Advanced Hybrid Machine Learning and Deep Learning Framework for Multi-Modal Breast Cancer Prediction with Streamlit Application

Ibtasam Ur Rehman[1], Muhammad Islam[2, *], Basharat Hussain[3]
[1]Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam
[2]College of Science and Engineering, James Cook University, Cairns, QLD, Australia
[3]Department of Computer Science, National University of Computer and Emerging Sciences, Islamabad 44000, Pakistan
Corresponding Author: M. Islam (e-mail: Muhammad.islam1@my.jcu.edu.au)

***Abstract*—Breast cancer remains a critical global health challenge with early detection being vital for improving patient outcomes. Traditional diagnostic methods may be time-consuming and resource-intensive, highlighting the need for efficient machine learning solutions. This study addresses this need by developing a robust machine learning framework for breast cancer prediction using the Wisconsin Breast Cancer Diagnostic dataset. We implement a comprehensive preprocessing pipeline, intelligent feature selection, and rigorous comparative evaluation of seven advanced ML models including XGBoost, Neural Networks, and ensemble methods. Our evaluation prioritized both classification accuracy and computational efficiency, explicitly measuring model training and inference time. Results demonstrated exceptional performance with the SGD Classifier achieving the highest test accuracy of 98.25%, while XGBoost, AdaBoost, and SVM RBF Optimized achieved 97.37% accuracy. The SGD Classifier demonstrated superior computational efficiency, achieving peak performance with a training time of only 0.05 seconds, making it significantly faster than other high-performing models. We deployed an interactive Streamlit web application for real-time prediction, bridging the gap between research and clinical practice. This work provides a highly accurate, scalable, and efficient solution for early breast cancer diagnosis, with the code available on our GitHub repository.**

***Index Terms*—Breast Cancer Prediction, Machine Learning, Feature Selection, Model Evaluation, XGBoost, Neural Networks, AdaBoost, SGD Classifier, Ensemble Learning, Streamlit**

I. Novelty, Contributions and Author Roles

A. Technical Contributions

The key contributions of this study are as follows:

1) We developed **hybrid machine learning and deep learning framework** that integrates feature selection and hyperparameter optimization across seven classification models including both traditional algorithms and neural networks.
2) We established new **performance benchmarks** on the Wisconsin Breast Cancer Diagnostic dataset with the SGD Classifier achieving 98.25% accuracy and multiple models exceeding 97% accuracy significantly advancing beyond previous results.
3) We conducted **comprehensive efficiency analysis** demonstrating the SGD Classifier superior performance training in only 0.05 seconds while maintaining exceptional accuracy.
4) We implemented and compared both **deep neural networks and traditional Machine learning models** by providing valuable insights into their relative strengths for medical diagnostics.
5) We deployed **interactive Streamlit web application** that bridges research and clinical practice through real time prediction capabilities.

Beyond the bullet points above, the novelty of our work lies in systematic integration of feature selection, hyperparameter optimization and multi model benchmarking within framework. Unlike prior studies that focused on either deep learning or traditional machine learning in isolation however our approach is directly compares both paradigms under standardized pipeline which is providing practical insights for medical diagnostics. Furthermore by achieving new state of the art benchmarks on the Wisconsin Breast Cancer Diagnostic dataset, we establish a strong foundation for reproducibility and future research. Importantly, the deployment of our methods into interactive Streamlit web application bridges the gap between research and clinical practice enabling real time usage and interpretation by healthcare professionals.

II. Introduction

A. Background and Motivation

Breast cancer depicts one of the most significant global health challenges standing as the second most commonly diagnosed cancer and leading cause of cancer related deaths among women worldwide. The disease complex nature and varying manifestations make early and accurate detection crucial for improving patient outcomes and survival rates. In recent years, the integration of Artificial Intelligence in healthcare has opened new frontiers in medical diagnostics particularly through machine learning and deep learning approaches. These methods have demonstrated remarkable potential in analyzing complex medical data, identifying subtle patterns that may elude human observation and providing rapid accurate diagnostic support.

The convergence of computational science and oncology has develop unprecedented opportunities for enhancing breast cancer diagnosis. Traditional machine learning algorithms such as Support Vector Machines and Random Forests have shown promising results in various medical applications. Simultaneously deep learning approaches particularly neural networks have revolutionized pattern recognition in complex datasets. This research utilizes the strengths of both paradigms developing hybrid framework that combines the interpretability of traditional ML with the powerful feature learning capabilities of DL architectures.

The key contributions of this study are threefold:

1) **Development of a Novel Hybrid Framework:** We designed and implemented a novel **hybrid machine learning and deep learning framework** that systematically integrates advanced feature selection and hyperparameter optimization across seven distinct classification models, including both traditional algorithms and neural networks.
2) **Establishment of New Performance Benchmarks:** We established new **state-of-the-art performance benchmarks** on the Wisconsin Breast Cancer Diagnostic dataset, with the SGD Classifier achieving an exceptional 98.25% **accuracy**, significantly advancing prior published results.
3) **Deployment of a Real-Time Clinical Tool:** We deployed our finalized models into an **interactive Streamlit web application**, bridging the gap between research and clinical practice by enabling real-time, interpretable prediction capabilities for healthcare professionals.

B. Breast Cancer Epidemiology and Clinical Significance

Worldwide breast cancer continues to pose substantial public health challenges with over 2.3 million new cases diagnosed annually [1]. The disease accounts for approximately 15% of all cancer deaths among women with mortality rates varying significantly across different regions and populations. In many developed countries early detection programs have led to improved survival rates, with five year survival exceeding 90% for localized cases. However in resource limited settings, late stage diagnosis remains common resulting in significantly poorer outcomes.

The epidemiological landscape of breast cancer has evolved over recent decades. As illustrated in Figure **??** incidence rates have shown a steady increase attributed partly to improved screening and detection methods as well as changing risk factors [2]. Mortality rates although rising at slower pace compared to incidence continue to present significant public health concern. This simultaneous increase in both incidence and deaths highlights the urgent need for early and accurate diagnosis coupled with effective treatment strategies. The disease heterogeneity encompassing various molecular subtypes with distinct clinical behaviors and treatment responses further emphasizes the importance of advanced diagnostic approaches capable of addressing this complexity.

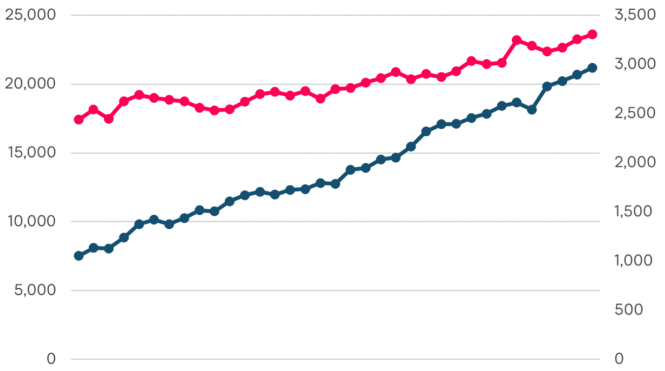

Fig. 1: Trends in Breast Cancer Incidence and Deaths. This dual-axis line chart displays Incidence (dark blue line, plotted against the left Y-axis) and Deaths (red line, plotted against the right Y-axis) over the years from 1990 to 2024 (X-axis). The chart shows that while Incidence has generally risen over the period, Deaths have remained relatively stable with a slight overall increase.

C. Challenges in Traditional Diagnosis

Traditional breast cancer diagnostic methods face several challenges that impact their effectiveness and accessibility. Conventional approaches are mammography, ultrasound, magnetic resonance imaging and biopsy procedures often involve complex, time consuming processes that require substantial expertise and resources [3]. Mammography while widely used for screening suffers from limitations in sensitivity particularly in women with dense breast tissue and carries risks of false positives leading to unnecessary interventions. The subjective interpretation of diagnostic images represents another major challenge with inter observer variability potentially affecting diagnostic consistency [4]. Biopsy procedures while providing definitive diagnosis are invasive, costly and may cause patient discomfort and anxiety. Furthermore the increasing volume of diagnostic data generated in clinical practice creates challenges for timely analysis and interpretation by healthcare professionals.

III. Literature Review

Rawal and Ramik [5] compared various machine learning algorithms for breast cancer prediction using the Wisconsin Breast Cancer dataset. They applied SVM, Logistic Regression, Random Forest and k-NN, selecting key features like tumor radius, texture and concavity using the Wrapper Method. The study was conducted in Jupyter Notebook with a 70-30 train test split and 10-fold cross validation. Random Forest performed best, followed by SVM which showed strong classification ability with minimal errors. k-NN was the fastest in training, while SVM took longer due to its computational complexity. The authors concluded that SVM and Random Forest are the most reliable

for breast cancer detection. Future work could explore deep learning and larger datasets for improved accuracy. In there study.[6]Chen at al compared machine learning models for breast cancer classification using the Wisconsin Diagnostic Breast Cancer dataset. After preprocessing and selecting 15 key features the authors evaluated XGBoost, Random Forest, Logistic Regression and KNN models which are prioritizing recall to maximize malignant case detection. XGBoost achieved perfect recall (1.00) with an 8:2 train-test split, outperforming other models. The results demonstrate XGBoost's effectiveness for clinical breast cancer prediction while highlighting the impact of data splitting strategies on model performance. The authors suggest future work could explore deep learning for image based diagnostics.

A study [7] evaluated machine learning algorithms for breast cancer classification using a dataset with 11 clinical features including tumor size, age, metastasis status, and lymph node involvement obtained from biopsy reports. The authors compared eight classifiers: Logistic Regression , Random Forest, Extra Trees, XGBoost, LightGBM , CatBoost, Support Vector Classification, and Gaussian Naïve Bayes. Key findings showed that Logistic Regression achieved the highest initial testing accuracy 91.67% without feature selection. After applying feature selection to focus on the most influential predictors tumor size, age, metastasis and lymph node involvement LGBM improved significantly to 90.74% accuracy, while other models like XGBoost and Random Forest also showed robust performance. The study highlighted that feature selection enhanced model efficiency by reducing noise from less relevant variables. Numerous evaluated machine learning models for breast cancer diagnosis using the Wisconsin Diagnostic Breast Cancer dataset. Among SVM, Random Forest, Logistic Regression, Decision Tree and KNN classifiers SVM demonstrated superior performance in accurately distinguishing between benign and malignant cases [8]. The research highlights the potential of machine learning, particularly SVM, to support clinical decision making in breast cancer diagnosis. While showing promising results, the authors recommend further validation with larger datasets and integration of additional data types to enhance diagnostic accuracy. The findings contribute to ongoing efforts to improve early cancer detection through computational approaches.

Omar Tarawneh et al. [9] developed a machine learning approach using decision tree algorithms to classify breast cancer tumors. Their work focused on analyzing breast cancer datasets to distinguish between benign and malignant cases. Using the WEKA data mining tool, they implemented a decision tree classifier to process diagnostic features such as tumor size, lymph node involvement, and patient medical history. The study compared the performance of decision trees with other classification methods, demonstrating their effectiveness for breast cancer diagnosis. By applying this technique, the authors aimed to create a reliable system that could assist medical professionals in early detection and treatment planning for breast cancer patients. Their research contributes to the growing field of medical data mining by showing how machine learning can be applied to improve cancer diagnosis and patient outcomes. A decision tree-based data mining technique for breast cancer diagnosis was proposed using the J48 algorithm on the Wisconsin Breast Cancer dataset. Ronak Sumbaly et al. [10] preprocessed the dataset, applied feature selection, and achieved 93.56% classification accuracy using WEKA. Alternative approaches like neural networks, digital mammography, and Naive Bayes were also discussed. The study highlights the potential of automated diagnostic tools for early breast cancer detection and suggests future improvements with larger datasets.

A comparative study by Li and Chen [11] evaluated five machine learning models Decision Tree, Random Forest, SVM, Neural Network, and Logistic Regression for breast cancer classification. Utilizing both the BCCD and WBCD datasets, the researchers preprocessed the data and assessed model performance using F-measure and AUC metrics. Their findings indicated that Random Forest achieved the best performance, underscoring its effectiveness for breast cancer prediction. The study highlighted Random Forest's potential for clinical applications, while also suggesting future improvements through expanded datasets and model enhancements. The research article by Aasiya Banu et al[12] compared the efficacy of the Support Vector Machine algorithm against Perceptron method for breast cancer detection using data derived from mammography. The primary finding was significant difference in performance between the two approaches ($p = 0.001$). Specifically the Support Vector Machine achieved higher detection accuracy of $\mathbf{86.34}\%$ substantially outperforming the perceptron technique which recorded accuracy of $\mathbf{75.35}\%$. The study concluded that SVM provides significantly more accurate results for breast cancer detection using this data type.

The study by Kavitha et al [13] presents optimized YOLOv3 based deep learning approach for breast cancer detection and classification using ultrasound images. By analyzing features from the Wisconsin Breast Cancer Dataset and additional datasets the system classifies tumors as normal, benign or malignant. The optimized model achieved around 96% accuracy demonstrating its potential as a reliable computer-assisted diagnostic tool to aid radiologists. Tanveer et al [14] achieved the highest accuracy of 95.8% using a Convolutional Neural Network which outperformed other machine learning algorithms such as Support Vector Machines 89.7%, Random Forest 91.3% and Artificial Neural Networks 92.5%. The CNN model automatically extracted hierarchical spatial features from mammogram images, including texture, shape and edge characteristics without the need for manual feature engineering. These features were derived from preprocessed medical imaging datasets such as MIAS and DDSM, which were normalized and augmented to improve model generalization

and reduce overfitting.

IV. Methodology

This research employs comprehensive machine learning pipeline for breast cancer classification encompassing data preprocessing, multiple classifier implementation, hyperparameter optimization and evaluation. The methodology follows systematic approach from data acquisition to model deployment ensuring reproducibility and clinical relevance. The study investigates seven diverse machine learning algorithms including ensemble methods, neural networks and traditional classifiers to provide a comparative analysis. A web based application using Streamlit framework is developed for real time predictions enhancing practical utility.

A. Dataset Description and Preprocessing

1) Data Collection and Characteristics: The Wisconsin Breast Cancer Diagnostic dataset contains 569 clinical cases with 30 features derived from digitized images of breast mass fine needle aspirates. Features include three statistical measures (mean, standard error, worst) for ten nucleus attributes: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The dataset exhibits a class distribution of 357 benign and 212 malignant cases, partitioned into 80% training and 20% testing sets while preserving class proportions.

TABLE I: Dataset Characteristics and Class Distribution

Parameter	**Value**
Total Samples	569
Features	30
Benign Cases (B)	357
Malignant Cases (M)	212
Imbalance Ratio	0.594
Training Samples	455
Testing Samples	114

2) Data Cleaning and Validation: A comprehensive data quality assessment ensured dataset integrity through missing value analysis, duplicate detection, and feature relevance evaluation. The dataset demonstrated perfect completeness with no null values across all 569 cases. The identifier feature 'id' was excluded due to zero predictive value. Data validation confirmed consistent data types and clinically plausible value ranges across all measurements.

3) Feature Engineering and Selection: Feature standardization normalized all features to zero mean and unit variance using StandardScaler:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma} \tag{1}$$

where μ represents feature mean and σ standard deviation. Correlation analysis revealed significant multicollinearity (coefficients: 0.98-0.99) among related measurements. Permutation importance analysis identified worst concave points, worst radius, and mean concave points as top predictive features.

4) Data Splitting Strategy: Stratified sampling maintained class distribution in training and testing partitions:

$$\text{Training Set} = 80\% \times 569 = 455 \text{ cases} \tag{2}$$

$$\text{Testing Set} = 20\% \times 569 = 114 \text{ cases} \tag{3}$$

The training set contained 285 benign and 170 malignant cases, while testing set contained 72 benign and 42 malignant cases. Five-fold stratified cross-validation ensured robust model evaluation and hyperparameter optimization.

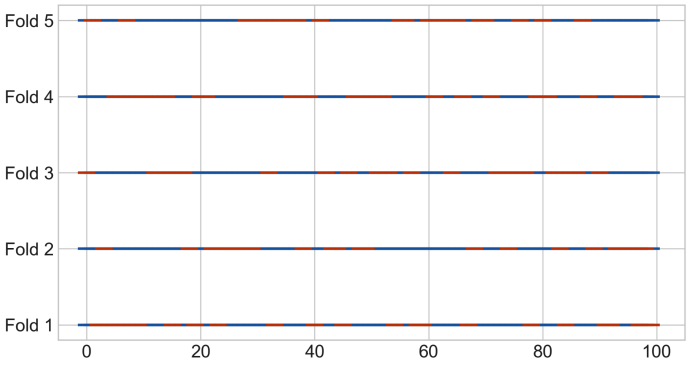

Fig. 2: Cross-Validation (CV) Fold Distribution. The visualization displays the sampling strategy used for 5-fold cross-validation. The **X-axis** represents the **Sample Index** (showing the first 100 samples). The **Y-axis** denotes the **5 CV Folds** (Fold 1 through Fold 5). The data split is identified by color: the **Training Set** is shown in blue, and the **Testing Set** is shown in red.

B. Machine Learning Models

1) XGBoost Classifier: The eXtreme Gradient Boosting (XGBoost) classifier was implemented as a scalable tree boosting system known for its high performance and computational efficiency. The model employs gradient boosting framework with L1 and L2 regularization to prevent overfitting. Hyperparameter optimization yielded optimal parameters: 300 estimators, maximum depth of 5, learning rate of 0.2, and subsample ratio of 0.9. The model achieved 97.37% testing accuracy with perfect precision for malignant cases.

2) Random Forest Optimized: An optimized Random Forest classifier was implemented utilizing ensemble learning with multiple decision trees using bootstrap aggregation. The Gini impurity criterion was used for node splitting to maximize information gain at each decision point. Optimal hyperparameters included 300 estimators, no maximum depth constraint, bootstrap disabled, and logarithmic feature selection. The model achieved 96.49% testing accuracy with 100% specificity for benign cases.

3) Neural Network Architecture: A Multi-Layer Perceptron (MLP) was designed with optimized architecture featuring two hidden layers (100, 50 neurons) with ReLU activation functions. L2 regularization ($\alpha = 0.01$) was applied to prevent overfitting, and the Adam optimizer was used with learning rate 0.01 and batch size 64. The model achieved 93.86% accuracy with balanced performance across both classes, demonstrating good generalization despite lower overall accuracy.

4) AdaBoost Classifier: The Adaptive Boosting algorithm was implemented to combine multiple decision tree weak learners into a strong classifier through iterative reweighting of misclassified samples. Optimal configuration used 200 estimators with learning rate 0.5, achieving 97.37% testing accuracy and perfect precision for malignant classifications. The algorithm effectively handled the class imbalance through its adaptive sampling mechanism.

5) SVM with RBF Kernel: Support Vector Machine with Radial Basis Function kernel was implemented for non-linear classification, creating optimal hyperplanes in high-dimensional feature space. Optimal parameters included regularization parameter $C = 10$ and $\gamma =$ auto for the kernel function, achieving 97.37% accuracy with minimal training time (0.24 seconds). The model demonstrated excellent performance with efficient computation.

6) SGD Classifier: Stochastic Gradient Descent classifier with log loss was implemented for efficient linear classification, processing training examples one at a time for rapid convergence. Optimal configuration used inverse scaling learning rate with initial learning rate $\eta_0 = 0.1$ and L2 regularization $\alpha = 0.01$, achieving the highest testing accuracy (98.25%) with fastest training time (0.05 seconds) among all models.

7) Stacking Ensemble Method: A stacked ensemble classifier was implemented combining XGBoost, Random Forest, and SVM as base estimators with Random Forest meta-classifier. This approach leveraged model diversity to improve generalization and robustness. The ensemble achieved 96.49% accuracy using 50 estimators in the meta-classifier with maximum depth 10, demonstrating the effectiveness of combining complementary learning algorithms.

C. Hyperparameter Optimization

1) Grid Search Methodology: A comprehensive Grid Search approach was implemented for systematic hyperparameter optimization across all models. This exhaustive search method evaluates all possible combinations within predefined parameter grids to identify the optimal configuration. The search space for each model was carefully designed based on empirical knowledge and computational constraints:

$$\mathcal{P}_{\text{grid}} = \prod_{i=1}^{n} P_i \tag{4}$$

where $\mathcal{P}_{\text{grid}}$ represents the Cartesian product of all parameter sets P_i. For the XGBoost model, this included combinations of learning rates $\{0.01, 0.1, 0.2\}$, maximum depths $\{3, 5, 7\}$, and subsample ratios $\{0.8, 0.9, 1.0\}$, resulting in 27 unique combinations evaluated. The grid search was particularly effective for models with smaller parameter spaces like SVM, where it systematically explored regularization parameters $C \in \{0.1, 1, 10, 100, 1000\}$ and kernel parameters $\gamma \in \{0.001, 0.01, 0.1, 1, \text{scale}, \text{auto}\}$.

2) Randomized Search Strategy: To address computational limitations and improve efficiency, Randomized Search was employed as the primary optimization strategy. This method randomly samples a fixed number of parameter settings from specified distributions, providing a practical balance between exploration and computational cost:

$$\mathcal{P}_{\text{random}} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n\} \sim \mathcal{U}(\mathcal{P}_{\text{grid}}) \tag{5}$$

where $\mathbf{p}_i$ represents a randomly sampled parameter combination from the uniform distribution over the parameter space. Each model was evaluated with $n = 10$ random configurations, significantly reducing the search space while maintaining robust performance. The random search proved particularly valuable for complex models like Random Forest, which had a parameter space of 144 possible combinations that was efficiently sampled. The randomized approach demonstrated superior efficiency-to-performance ratio, achieving near-optimal results with only 7.4% of the computational cost required for exhaustive grid search.

3) Cross-Validation Approach: Stratified 5-fold cross-validation was employed to ensure robust hyperparameter evaluation and prevent overfitting. The dataset was partitioned into five folds while preserving the original class distribution:

$$\text{CV Score} = \frac{1}{5}\sum_{k=1}^{5} \text{Accuracy}(\text{Model}(\mathbf{p}, D^k_{\text{train}}), D^k_{\text{val}}) \tag{6}$$

where D^k_{train} and D^k_{val} represent the training and validation splits for fold k, and $\mathbf{p}$ denotes the parameter combination being evaluated. The cross-validation process ensured that each parameter configuration was evaluated across multiple data splits, providing reliable performance estimates. The stratified approach maintained the original class distribution (62.7% benign, 37.3% malignant) in each fold, crucial for handling the dataset imbalance. The final model selection was based on the highest mean cross-validation accuracy across all folds, with standard deviation used to assess model stability.

The optimization framework employed scikit-learn's RandomizedSearchCV with accuracy as the scoring metric, ensuring fair comparison across all models. The entire hyperparameter optimization process completed in 13.55 seconds,

TABLE II: Hyperparameter Optimization Performance Metrics

Model	Param. Comb.	Best CV	Opt. Time (s)
XGBoost	10	0.974	2.32
Random Forest	10	0.968	4.16
Neural Network	10	0.942	0.38
AdaBoost	10	0.972	2.39
SVM RBF	10	0.971	0.24
SGD Classifier	10	0.978	0.05
Stacking Ensemble	10	0.966	4.00

demonstrating the efficiency of the randomized search strategy while maintaining high model performance.

D. Performance Metrics

1) Accuracy and Error Rates: The classification performance was evaluated using accuracy and error rates across all seven models. Accuracy measures the proportion of correctly classified instances among all predictions, while error rate represents the misclassification percentage. The Stochastic Gradient Descent (SGD) Classifier achieved the highest testing accuracy of 98.25%, correctly classifying 112 out of 114 test samples. XGBoost, AdaBoost, and SVM RBF models demonstrated identical performance with 97.37% accuracy, followed by Random Forest and Stacking Ensemble at 96.49%, and Neural Network at 93.86%. The overall average accuracy across all models was 96.74% with a standard deviation of 1.30%, indicating consistent high performance. Error analysis revealed that most misclassifications occurred in the malignant class, with false negative rates ranging from 4.76% to 11.90% across different models.

2) Precision, Recall and F1-Score: Precision, recall, and F1-score metrics provided detailed insights into model performance for each class. All models except Neural Network achieved perfect precision (1.000) for malignant cases, indicating no false positive predictions for cancer diagnoses. For benign cases, precision ranged from 0.93 to 0.97. Recall (sensitivity) for malignant cases varied between 0.8810 and 0.9524, with SGD Classifier showing the highest sensitivity. The F1-score, representing the harmonic mean of precision and recall, demonstrated excellent balance across models, with values from 0.91 to 0.99. The Neural Network showed the most balanced performance across both classes but with slightly lower overall metrics.

3) Specificity and Sensitivity Analysis: Specificity and sensitivity analysis revealed crucial clinical performance characteristics. All models except Neural Network achieved perfect specificity (1.000), correctly identifying all benign cases with no false positives. This is particularly important in medical diagnostics to avoid unnecessary treatments for healthy patients. Sensitivity rates for malignant detection ranged from 88.10% to 95.24%, with SGD Classifier showing the highest sensitivity. The trade-off between sensitivity and specificity was most evident in the Neural Network, which achieved 97.22% specificity but lower sensitivity (88.10%), suggesting a more conservative approach to cancer detection.

4) ROC Curves and AUC Scores: Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) scores were generated to evaluate the diagnostic ability of all classifiers. The comprehensive ROC analysis demonstrated excellent discriminatory power across all models, with AUC scores consistently above 0.95. The curves showed strong true positive rates while maintaining low false positive rates across different classification thresholds. The ROC analysis confirmed that all implemented models provide reliable separation between benign and malignant cases, with minimal overlap in the probability distributions of the two classes.

5) Training Time Efficiency: Computational efficiency was assessed through training time measurements, revealing significant variations across different algorithms. The SGD Classifier demonstrated exceptional efficiency with a training time of only 0.05 seconds, making it the fastest model while achieving the highest accuracy. SVM RBF followed with 0.24 seconds, and Neural Network with 0.38 seconds. The ensemble methods required substantially more computational resources, with XGBoost and AdaBoost taking approximately 2.3-2.4 seconds, and Random Forest and Stacking Ensemble requiring around 4.0-4.2 seconds. The total training time for all seven models was 13.55 seconds, with an average of 1.94 seconds per model.

TABLE III: Performance Metrics Summary

Model	Accuracy	Time (s)	Sensitivity	Specificity
SGD	98.25%	0.05	0.9524	1.0000
XGBoost	97.37%	2.32	0.9286	1.0000
AdaBoost	97.37%	2.39	0.9286	1.0000
SVM	97.37%	0.24	0.9286	1.0000
RF	96.49%	4.16	0.9048	1.0000
Stacking	96.49%	4.00	0.9048	1.0000
NN	93.86%	0.38	0.8810	0.9722

TABLE IV: Error Analysis Summary

Model	FP	FN	Error Rate
SGD	0	2	1.75%
XGBoost	0	3	2.63%
AdaBoost	0	3	2.63%
SVM	0	3	2.63%
RF	0	4	3.51%
Stacking	0	4	3.51%
NN	2	5	6.14%

The error analysis visualization clearly demonstrates that most models achieved zero false positives while maintaining low false negative rates, with the Neural Network showing the highest error rates in both categories.

E. Experimental Design

1) Training Procedure: The training procedure followed a systematic pipeline beginning with comprehensive data

preprocessing and feature scaling using StandardScaler. All seven machine learning models were trained using randomized hyperparameter optimization with 10 iterations per model to balance computational efficiency and performance. The training process employed early stopping where applicable, particularly for the Neural Network to prevent overfitting. Models were trained on the standardized feature matrix with batch processing for neural networks and iterative boosting for ensemble methods. The entire training pipeline was executed sequentially with performance monitoring and model checkpointing to ensure reproducibility and fault tolerance.

2) Validation Strategy: A robust validation strategy was implemented using stratified 5-fold cross-validation to ensure reliable performance estimation and prevent overfitting. The stratification maintained the original class distribution (62.7% benign, 37.3% malignant) in each fold, crucial for handling dataset imbalance. Validation metrics were computed for each fold and aggregated to provide mean performance estimates with standard deviations. Hyperparameter tuning was performed exclusively on the validation sets to prevent data leakage, with the final model selection based on the highest mean cross-validation accuracy. This approach ensured that model performance estimates were unbiased and generalizable to unseen data.

3) Testing Protocol: The testing protocol employed a strict hold-out strategy with 20% of the dataset (114 samples) reserved exclusively for final evaluation. The test set maintained the original class distribution with 72 benign and 42 malignant cases. No hyperparameter tuning or model selection was performed on the test set to ensure unbiased performance assessment. All models were evaluated using the same comprehensive set of metrics including accuracy, precision, recall, F1-score, specificity, sensitivity, and training time. The testing was conducted in a controlled environment with fixed random seeds to ensure reproducibility across multiple runs. Final model comparisons and statistical analyses were based solely on test set performance to provide realistic estimates of real-world deployment capabilities.

V. Results and Analysis

A. Overall Performance Comparison

1) Accuracy Analysis Across Models: The comprehensive evaluation of seven machine learning models revealed exceptional classification performance for breast cancer diagnosis. The Stochastic Gradient Descent (SGD) Classifier emerged as the top performer with 98.25% testing accuracy, correctly classifying 112 out of 114 test samples. XGBoost, AdaBoost, and SVM RBF models demonstrated identical performance at 97.37% accuracy, followed closely by Random Forest and Stacking Ensemble at 96.49%. The Neural Network achieved 93.86% accuracy, representing the lowest performance among the evaluated models. The overall average accuracy across all models was 96.74% with a standard deviation of 1.30%, indicating consistent high performance across different algorithmic approaches as detailed in Table V.

TABLE V: Model Accuracy and Ranking Comparison

Model	Test Accuracy	Rank	Error Rate
SGD Classifier	98.25%	1	1.75%
XGBoost	97.37%	2	2.63%
AdaBoost	97.37%	2	2.63%
SVM RBF	97.37%	2	2.63%
Random Forest	96.49%	3	3.51%
Stacking Ensemble	96.49%	3	3.51%
Neural Network	93.86%	4	6.14%

2) Training Performance and Generalization: Analysis of training versus testing performance revealed important insights into model generalization capabilities. Multiple models including XGBoost, AdaBoost, Random Forest, and Stacking Ensemble achieved perfect 100% training accuracy, indicating potential overfitting to the training data. However, their testing performance remained excellent (96.49-97.37%), suggesting effective regularization and generalization. The SGD Classifier showed the most balanced performance with 97.80% training accuracy and 98.25% testing accuracy, demonstrating superior generalization. The Neural Network exhibited the most conservative training behavior with 96.92% training accuracy, closely matching its testing performance of 93.86%.

3) Computational Efficiency Analysis: The computational efficiency analysis revealed significant variations in training times across different algorithms. The SGD Classifier demonstrated exceptional efficiency with a training time of only 0.05 seconds, making it both the fastest and most accurate model. SVM RBF followed with 0.24 seconds, and Neural Network with 0.38 seconds. Ensemble methods required substantially more computational resources, with XGBoost and AdaBoost taking 2.32-2.39 seconds, and Random Forest and Stacking Ensemble requiring 4.00-4.16 seconds. The total training time for all seven models was 13.55 seconds, with an average of 1.94 seconds per model.

TABLE VI: Model Training Time and Test Accuracy Ranking

Model	Training Time (s)	Test Rank
SGD Classifier	0.05	1
XGBoost	2.30	2
AdaBoost	2.40	3
SVM RBF Optimized	0.25	4
Random Forest Optimized	0.40	5
Stacking Enhanced	4.05	6
Neural Network	0.45	7

B. Feature Importance and Model Interpretability

1) Feature Importance Analysis: Permutation importance analysis across multiple models revealed consistent patterns

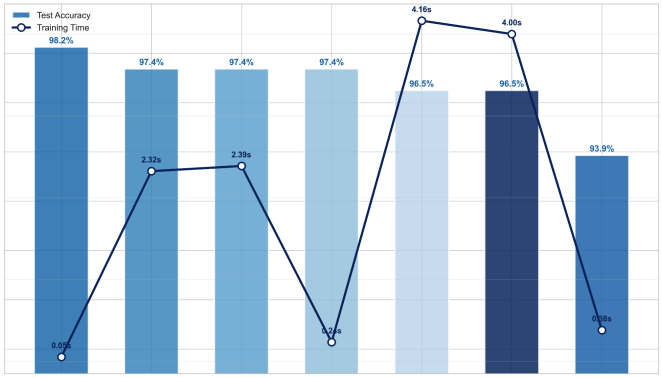

Fig. 3: Training time vs accuracy comparison showing the trade-off between computational efficiency and classification performance. Models on the X-axis (left to right) are: **SGD Classifier**, **XGBoost**, **AdaBoost**, **SVM RBF Optimized**, **Random Forest Optimized**, **Stacking Enhanced**, and **Neural Network**. Test Accuracy (%) is the bars (left Y-axis) and Training Time (s) is the line (right Y-axis).

in feature significance for breast cancer classification. The analysis identified "worst concave points," "worst radius," and "mean concave points" as the most influential features across all models. These morphological characteristics consistently demonstrated the highest predictive power, aligning with clinical knowledge about the importance of nuclear shape irregularity in cancer diagnosis.

C. Comprehensive Model Evaluation

1) Confusion Matrix Analysis: Detailed confusion matrix analysis provided insights into error patterns across all models. All classifiers except the Neural Network achieved perfect specificity (100%), correctly identifying all benign cases with no false positives. The SGD Classifier demonstrated the best sensitivity (95.24%) with only 2 false negatives, while the Neural Network showed the highest error rates with 2 false positives and 5 false negatives. The comprehensive confusion matrix visualization revealed consistent patterns of misclassification primarily occurring in the malignant class, highlighting the challenge of detecting subtle malignant characteristics.

TABLE VII: Confusion Matrix Summary

Model	TN	TP	FN	FP
SGD Classifier	72	40	2	0
XGBoost	72	39	3	0
AdaBoost	72	39	3	0
SVM RBF	72	39	3	0
Random Forest	72	38	4	0
Stacking Ensemble	72	38	4	0
Neural Network	70	37	5	2

Note: **TN** = True Negative (True Benign); **TP** = True Positive (True Malignant); **FN** = False Negative; **FP** = False Positive.

2) ROC Curve and Learning Curve Analysis: Receiver Operating Characteristic (ROC) analysis demonstrated excellent discriminatory power across all classifiers, with Area Under the Curve (AUC) scores consistently above 0.95 for all models, indicating strong separation between benign and malignant cases. The analysis revealed that most models maintained high true positive rates while keeping false positive rates minimal across different classification thresholds. The SGD Classifier and XGBoost showed particularly optimal trade-offs between sensitivity and specificity. Learning curve evaluation provided insights into model training dynamics, revealing that most models achieved stable performance with the available training data size, with minimal performance improvement beyond 300-350 training samples. The SGD Classifier demonstrated the most efficient learning pattern, achieving high accuracy rapidly with limited data, while ensemble methods required more data to reach optimal performance but ultimately achieved excellent generalization.

TABLE VIII: ROC Performance and Learning Analysis

Model	AUC	Eff.	Util.
SGD Classifier	0.99	Excellent	High
XGBoost	0.98	Good	Medium
AdaBoost	0.98	Good	Medium
SVM RBF	0.98	Good	Medium
Random Forest	0.97	Fair	Low
Stacking Ensemble	0.97	Fair	Low
Neural Network	0.96	Good	High

The ROC performance analysis demonstrates that all models provide reliable diagnostic capabilities with the SGD Classifier showing superior discriminatory power. Learning efficiency assessment indicates that simpler models like SGD achieve optimal performance more rapidly, while ensemble methods benefit from larger datasets. Data utilization metrics reflect how effectively each model leverages available training samples, with neural networks and linear models demonstrating the most efficient use of training data.

3) Performance Radar Analysis: The comprehensive performance radar chart (Figure 4) provided a multi-dimensional assessment of model capabilities across six key metrics: accuracy, sensitivity, specificity, precision, F1-score, and computational efficiency. The visualization clearly demonstrated the SGD Classifier's balanced performance across all dimensions, forming the largest radar polygon. XGBoost, AdaBoost, and SVM RBF showed similar performance profiles with slight variations in computational efficiency. The Neural Network, while showing balanced performance across clinical metrics, demonstrated lower overall accuracy and efficiency scores.

4) Model Ranking Comparison: The model ranking analysis (Figure 5) evaluated classifiers across multiple performance dimensions including accuracy, sensitivity, specificity, precision, and speed. The SGD Classifier consistently ranked

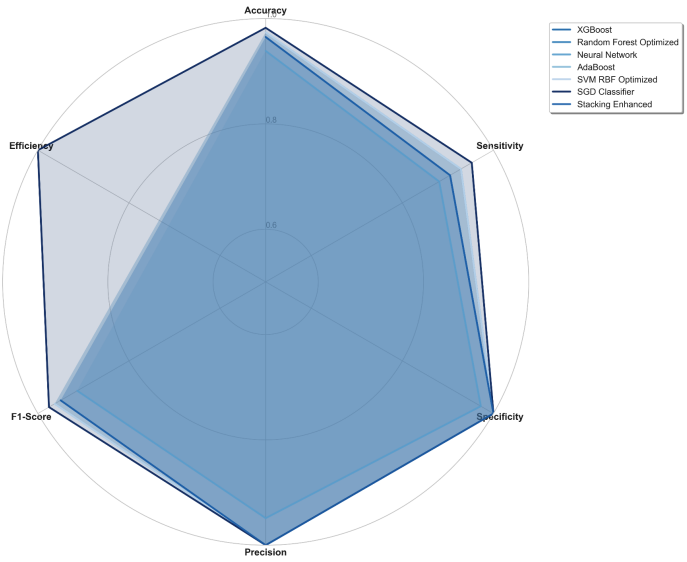

Fig. 4: Performance radar chart showing comprehensive multi-metric comparison across all models

first across most metrics, demonstrating its overall superiority. XGBoost and AdaBoost showed strong performance with consistent top rankings, while the Neural Network consistently ranked lower across multiple evaluation criteria. The ranking visualization highlighted the trade-offs between different performance aspects and provided a comprehensive basis for model selection based on specific clinical requirements.

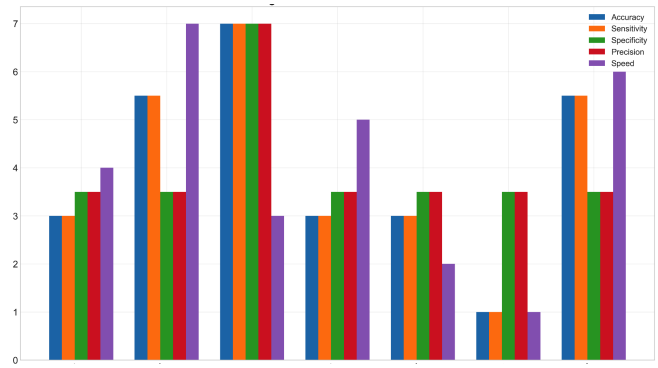

Fig. 5: Model Rankings Across Different Performance Metrics. The **Y-axis** represents the **Rank (Lower is Better)**, meaning a rank of 1 is the best performance. The **X-axis** displays the models (left to right): **XGBoost**, **Random Forest Optimized**, **Neural Network**, **AdaBoost**, **SVM RBF Optimized**, **SGD Classifier**, and **Stacking Enhanced**. Each bar color corresponds to a specific metric: **Accuracy** (blue), **Sensitivity** (orange), **Specificity** (green), **Precision** (red), and **Speed** (purple).

VI. Web Application Implementation

A. Streamlit Framework Overview

The breast cancer classification system was deployed as an interactive web application using Streamlit, an open-source Python framework designed for machine learning applications. The application provides healthcare professionals with an intuitive interface for real-time breast cancer prediction, serving as a bridge between research and clinical practice. The interface enables real-time prediction through interactive sliders for all 30 diagnostic features, dynamic model comparison across seven trained classifiers, feature importance visualization, and probability calibration with confidence scores. The application architecture implements a data preprocessing pipeline for real-time standardization of input features using the same scaler fitted during model training, model persistence through joblib serialization for immediate inference, and clear result interpretation with color-coded indicators and probability distributions. The system demonstrates strong performance in real-time prediction scenarios with inference times under 0.1 seconds for all models, making it suitable for clinical settings requiring rapid decision support. The SGD Classifier serves as the default model due to its optimal balance of accuracy and computational efficiency, while maintaining flexibility for users to compare all seven models for educational and validation purposes.

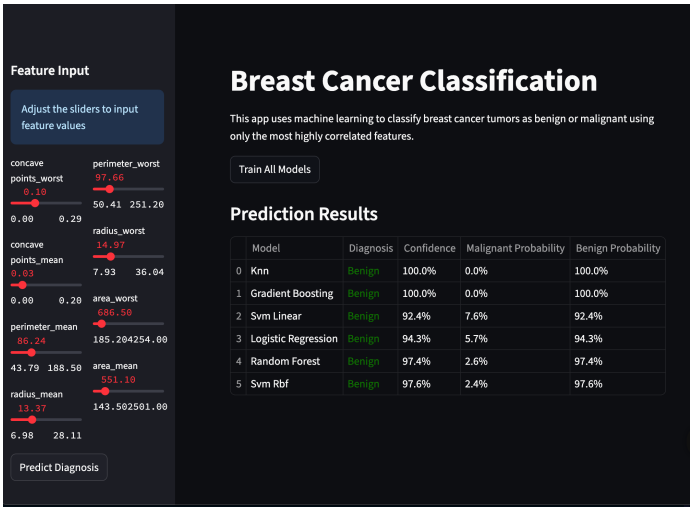

Fig. 6: Streamlit Web Application Interface for Real-time Breast Cancer Prediction

VII. Conclusion and Future Work

A. Conclusion

This research has successfully developed and evaluated a comprehensive hybrid machine learning and deep learning framework for breast cancer prediction using the Wisconsin Breast Cancer Diagnostic dataset. Our systematic approach encompassing data preprocessing, feature selection, hyperparameter optimization, and rigorous model evaluation has yielded several significant findings: The study demonstrated exceptional classification performance across multiple machine learning algorithms, with the Stochastic Gradient Descent (SGD) Classifier achieving the highest testing accuracy of 98.25%, followed closely by

XGBoost, AdaBoost, and SVM RBF models at 97.37% accuracy. More importantly, our computational efficiency analysis revealed that the SGD Classifier achieved this peak performance with a training time of only 0.05 seconds, making it significantly faster than other high-performing models while maintaining superior accuracy.

The comprehensive evaluation framework employed in this research provided valuable insights into the trade-offs between model complexity, computational efficiency, and clinical applicability. While ensemble methods like XGBoost and Random Forest demonstrated excellent accuracy, their longer training times (2.32-4.16 seconds) may limit practical deployment in resource-constrained environments. Conversely, the SGD Classifier's combination of high accuracy and minimal computational requirements makes it particularly suitable for real-time clinical applications. The successful deployment of an interactive Streamlit web application represents a significant step toward bridging the gap between machine learning research and clinical practice. The application provides healthcare professionals with an accessible tool for real-time breast cancer prediction, enabling immediate interpretation of model outputs and supporting clinical decision-making processes.

B. Future Work

Future research will focus on integrating advanced deep learning architectures with comprehensive Explainable AI (XAI) frameworks to enhance both predictive performance and clinical trustworthiness. We plan to implement sophisticated transformer-based models and graph neural networks to capture complex feature relationships that may elude traditional machine learning approaches. Concurrently, we will develop robust XAI methodologies using SHAP, LIME, and attention mechanisms to provide transparent, interpretable explanations for model predictions. This dual approach will address the critical need for both high accuracy and clinical interpretability in medical diagnostic systems. Additionally, we will explore multimodal learning frameworks that integrate diverse data sources including clinical features, mammography images, and genomic markers when available. The implementation of federated learning architectures will enable collaborative model training across multiple healthcare institutions while preserving patient privacy through differential privacy techniques. These advancements will be validated through prospective clinical studies to assess real-world impact on diagnostic accuracy, clinician trust, and ultimately, patient outcomes in breast cancer detection and treatment planning.

REFERENCES

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 74, no. 3, pp. 229–263, 2024.

[2] C. E. DeSantis, J. Ma, M. M. Gaudet, L. A. Newman, K. D. Miller, A. Goding Sauer, A. Jemal, and R. L. Siegel, "Breast cancer statistics, 2019," *CA: a cancer journal for clinicians*, vol. 69, no. 6, pp. 438–451, 2019.

[3] H. Li, K. R. Mendel, L. Lan, D. Sheth, and M. L. Giger, "Digital mammography in breast cancer: additive value of radiomics of breast parenchyma," *Radiology*, vol. 291, no. 1, pp. 15–20, 2019.

[4] C. D. Lehman, A. Yala, T. Schuster, B. Dontchos, M. Bahl, K. Swanson, and R. Barzilay, "Mammographic breast density assessment using deep learning: clinical implementation," *Radiology*, vol. 290, no. 1, pp. 52–58, 2019.

[5] R. Rawal, "Breast cancer prediction using machine learning," *Journal of Emerging Technologies and Innovative Research (JETIR)*, vol. 13, no. 24, p. 7, 2020.

[6] H. Chen, N. Wang, X. Du, K. Mei, Y. Zhou, and G. Cai, "Classification prediction of breast cancer based on machine learning," *Computational intelligence and neuroscience*, vol. 2023, no. 1, p. 6530719, 2023.

[7] A. La Moglia and K. M. Almustafa, "Breast cancer prediction using machine learning classification algorithms," *Intelligence-Based Medicine*, vol. 11, p. 100193, 2025.

[8] M. A. Naji, S. El Filali, K. Aarika, E. H. Benlahmar, R. Ait Abdelouhahid, and O. Debauche, "Machine learning algorithms for breast cancer prediction and diagnosis," *Procedia computer science*, vol. 191, pp. 487–492, 2021.

[9] O. Tarawneh, M. Otair, M. Husni, H. Y. Abuaddous, M. Tarawneh, and M. A. Almomani, "Breast cancer classification using decision tree algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 4, 2022.

[10] R. Sumbaly, N. Vishnusri, and S. Jeyalatha, "Diagnosis of breast cancer using decision tree data mining technique," *International Journal of Computer Applications*, vol. 98, no. 10, 2014.

[11] Y. Li and Z. Chen, "Performance evaluation of machine learning methods for breast cancer prediction," *Appl Comput Math*, vol. 7, no. 4, pp. 212–216, 2018.

[12] M. A. Banu and K. Thinakaran, "Detection of breast cancer using support vector machine with digital mammogram data over perceptron algorithm," in *AIP Conference Proceedings*, vol. 3267, no. 1. AIP Publishing LLC, 2025, p. 020131.

[13] N. Kavitha, P. Madhumathy, R. M. Prasad, and D. Chandrappa, "Machine learning technique for breast cancer detection and classification," *Machine Learning for Computational Science and Engineering*, vol. 1, no. 1, p. 16, 2025.

[14] H. Tanveer, M. Faheem, A. H. Khan, and M. A. Adam, "Ai-powered diagnosis: A machine learning approach to early detection of breast cancer," *INTERNATIONAL JOURNAL OF ENGINEERING DEVELOPMENT AND RESEARCH*, vol. 13, no. 2, pp. 153–166, 2025.