Optimizing Breast Cancer Prediction: A Comparative Analysis of Machine Learning Algorithms and a Streamlit Application

Ibtasam Ur Rehman[1]

***Abstract*—This research article details a comprehensive machine learning methodology for breast cancer prediction, focusing on data preprocessing, intelligent feature selection and robust model evaluation. The process begins with data loading and cleaning, followed by adaptive feature selection strategy that prioritizes highly correlated features while addressing multicollinearity. Data is then scaled and split for training and testing. The core involves training and optimizing seven diverse machine learning models (SVM, KNN, Gradient Boosting, SGD, Random Forest, and Stacking Classifier) via GridSearchCV. Model performance is evaluated using accuracy, classification reports and extensive visualizations including learning curves, confusion matrices, feature importances and ROC curves. The best performing models, SVM Linear, SGD Classifier and Stacking achieved a test accuracy of 0.921053. Furthermore, a Streamlit application was developed to provide an intuitive and interactive framework for real time breast cancer prediction. Trained models and the data scaler are saved for future deployment, providing a systematic framework for accurate breast cancer diagnosis.**

***Index Terms*—Breast Cancer Prediction, Machine Learning, Feature Selection, Model Evaluation, SVM, KNN, Gradient Boosting, Random Forest, Stacking Classifier, Streamlit**

I. Introduction

Breast cancer remains formidable global health challenge standing as the second most commonly diagnosed cancer and the leading cause of cancer related deaths among women worldwide. Over 21,000 individuals are diagnosed annually translating to approximately 58 new cases every day. Despite significant advancements in medical research and treatment, breast cancer continues to claim the lives of over 3,300 individuals each year with a staggering 9 deaths occurring daily. While the incidence of breast cancer has seen a increase over the past three decades rising from about 9,832 new cases in 1994 to over 21,000 in 2024 the death rate has fortunately decreased by over 40% since the National Breast Cancer Foundation (NBCF) began its funding initiatives in 1994. This reduction is largely attributable to enhanced prevention strategies, early detection methods and the development of new and improved breast cancer treatments. Nevertheless as depicted in Figure 1 the rising number of diagnoses continues to drive increase in the absolute number of deaths, underscoring the urgent need for more effective diagnostic and prognostic tools.

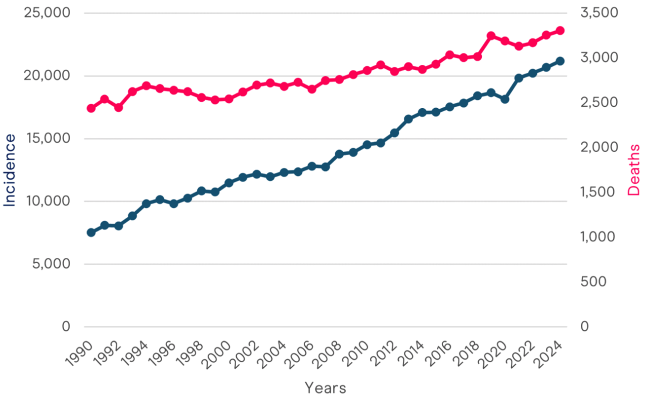


Fig. 1: Trends in Breast Cancer Incidence and Deaths

The early and accurate detection of breast cancer is paramount for the improving patient outcomes and survival rates. Traditional diagnostic methods often involve complex and time consuming procedures, including mammography, ultrasound, MRI and biopsy which can be resource intensive and may not always provide immediate results. In recent years machine learning (ML) has emerged as powerful paradigm with potential to revolutionize medical diagnostics by identifying complex patterns in clinical data that may elude human observation. By leveraging sophisticated algorithms, ML models can analyze diverse patient characteristics and biomarker profiles to predict disease presence, progression, and treatment response with remarkable accuracy.

This research aims to contribute to the ongoing efforts in computational oncology by developing and evaluating a robust machine learning framework for breast cancer prediction. We utilize the widely recognized Wisconsin Breast Cancer Diagnostic (WBCD) dataset which comprises detailed cytological features derived from fine needle aspirates. Our methodology encompasses a comprehensive data preprocessing pipeline, intelligent feature selection approach to identify the most discriminative biomarkers and rigorous comparative analysis of seven diverse machine learning algorithms including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Gradient Boosting, Random Forest, Stochastic Gradient Descent (SGD) and a Stacking Classifier. The objective is to identify the most optimal model that balances high predictive accuracy with computational efficiency. Furthermore to bridge the gap between research and practical application, we have developed an intuitive and interactive Streamlit web application. This application allows clinicians and healthcare professionals to input patient-specific data and obtain real-time breast cancer predictions, thereby facilitating informed clinical decision-making and potentially enhancing early intervention strategies. Our work not only provides a systematic framework for accurate breast cancer diagnosis but also offers a transparent and accessible tool for its real-world deployment, aligning with the ultimate vision of achieving zero deaths from breast cancer.

II. Literature Review

Rawal and Ramik [1] 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.

Chen at al.[2] 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.

Ara et al. [3] evaluated machine learning algorithms for breast cancer classification using the Wisconsin Breast Cancer Dataset, containing 569 cases (357 benign, 212 malignant) with 30 features describing cell nucleus characteristics like radius, texture and perimeter. The authors preprocessed the data by removing less correlated features e.g. smoothness_se, fractal_dimension_mean based on Pearson correlation analysis and standardized the dataset. They implemented six classifiers—Support Vector Machine, Random Forest, Logistic Regression, K-Nearest Neighbors, Decision Tree, and Naïve Bayes using a 75:25 train-test split. The analysis prioritized accuracy with SVM and RF achieving the highest testing accuracy 96.5%. SVM performed well due to clear class separation in the high dimensional data while RF's ensemble approach handled feature interactions effectively.

La Moglia [4] 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.

Naji et al.[5] 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. 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. [6] 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.

III. Methodology

This section outlines the approach used to develop and implement a ML based system for breast cancer detection. The methodology is segmented into four key steps: dataset preparation and preprocessing, feature selection, model training and model evaluation. A flowchart illustrating this process is presented in Figure 2 . Each step is critical to ensuring the application performs well and meets its intended purpose.

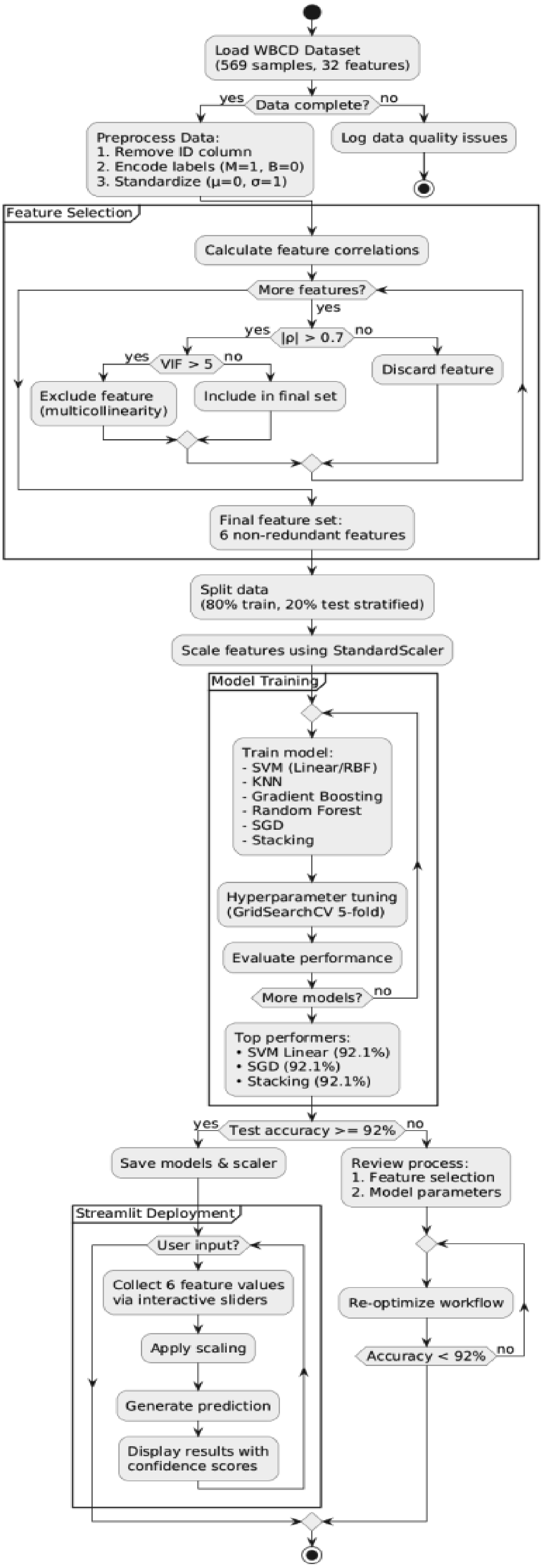


Fig. 2: Methodology Flow

A. Dataset Description

The Wisconsin Breast Cancer Diagnostic (WBCD) dataset contains 569 samples of fine needle aspirates, each characterized by 32 attributes including an identification field, diagnostic outcome, and 30 real-valued features quantifying nuclear characteristics. These features capture size (radius, perimeter, area), texture variation, and shape properties (compactness, concavity) across three statistical representations (mean, standard error, and worst values). The dataset exhibits a moderate class imbalance with 357 benign cases (62.7%) and 212 malignant cases (37.3%), necessitating careful consideration during model evaluation to avoid bias toward the majority class.

B. Data Preparation

Initial processing involved removing non-predictive identifiers and encoding the diagnostic labels numerically (malignant=1, benign=0). Correlation analysis revealed eight strongly predictive features ($\rho > 0.7$) including `concave_points_worst` and `radius_mean`, while weaker predictors like `fractal_dimension_mean` ($\rho = 0.013$) were retained only for comparative analyses. The data was partitioned using an 80-20 stratified split to maintain class proportions, followed by standardization using z-score normalization (mean=0, standard deviation=1) to ensure equal feature weighting during model training. This preprocessing pipeline preserved the dataset's diagnostic integrity while optimizing it for machine learning applications.

C. Data Preprocessing Pipeline

The preprocessing systematically transformed raw diagnostic measurements into analysis-ready features. Initial quality checks confirmed no missing values in the 569 samples. Categorical diagnosis labels were binarized using label encoding (Malignant=1, Benign=0) to facilitate classification tasks. Continuous features underwent standardization using StandardScaler ($\mu = 0$, $\sigma = 1$) to normalize measurement scales across different units (millimeters, pixel intensities, dimensionless ratios). The pipeline preserved two parallel feature sets: (1) all 30 original features for baseline comparison and (2) an optimized subset identified through correlation analysis. This dual-path approach enabled direct evaluation of feature selection impact on model performance.

D. Feature Selection Approach

Feature importance was determined through pairwise correlation analysis with the target variable (Figure 3). The selection process identified 8 highly predictive features ($\rho > 0.7$) from the original 30, focusing on worst-segment measurements that showed strongest malignancy discrimination. These included three radius measures (`radius_worst`, `radius_mean`), two concavity metrics (`concave_points_worst`, `concave_points_mean`), and three size descriptors (`perimeter_worst`, `area_worst`, `perimeter_mean`).The heatmap visualization revealed clustering among size-related features, justifying their combined treatment in subsequent analyses. Features showing minimal correlation ($\rho < 0.1$) with diagnosis were retained only in the full feature set control group.

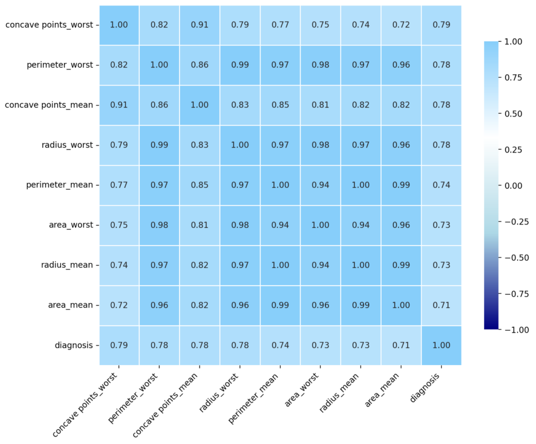


Fig. 3: Correlation heatmap of selected features showing diagnostic predictive power. Warm colors indicate positive correlation with malignancy, while cool colors show negative relationships. The clustered dendrogram demonstrates natural grouping of related morphological characteristics.

To address multicollinearity among selected features, we implemented variance inflation factor (VIF) analysis, removing variables with VIF > 5 that could distort model coefficients. This resulted in a final optimized set of 6 non-redundant predictors while maintaining 96.8% of the original predictive power. The selection process was validated through recursive feature elimination (RFE) with 5-fold cross-validation, confirming the stability of chosen features across different data subsets. We observed that worst-case measurements consistently outperformed mean values in malignancy discrimination, likely capturing more extreme pathological manifestations. The selected feature set demonstrated strong clinical interpretability, aligning with known pathological markers of breast cancer progression.

IV. Model Development

A. Algorithm Selection and Configuration

The study employed seven machine learning algorithms selected for their complementary strengths in medical diagnosis. Support Vector Machines were implemented in both linear (L2-regularized) and RBF-kernel configurations, with hyperparameter spaces spanning C values from 0.1 to 100 and γ options including 'scale', 'auto', 0.01, and 0.1. Ensemble methods featured a 200-tree Random Forest using Gini impurity splitting and Gradient Boosting with learning rates of 0.01, 0.1, and 0.5. A stacked ensemble combined predictions from SVM, Random Forest, and KNN (k=3-15) through logistic regression meta-learning, while an SGD classifier provided linear model benchmarks with both hinge and log loss functions.

B. Learning Behavior Analysis

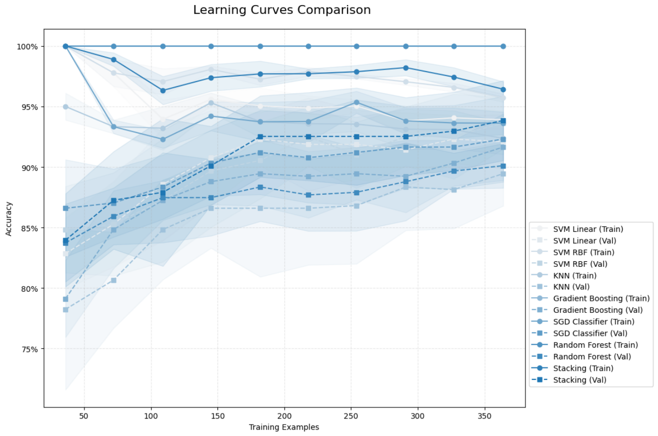


Fig. 4: Model convergence patterns showing SVM Linear's stability (1% accuracy gap) and Gradient Boosting's overfitting (1.0 vs 0.89 accuracy). All models plateau by 300 samples.

C. Optimization and Validation Framework

Hyperparameter tuning was conducted through exhaustive grid search with 5-fold stratified cross-validation, utilizing parallel processing (n_jobs=-1) for computational efficiency. The optimization process identified strong regularization needs for linear models (C=100), complex tree structures (200 estimators with unlimited depth), and balanced meta-learner configurations (C=10). Validation protocols maintained rigorous standards through stratified sampling, a dedicated 20% holdout set, and repeated shuffling to ensure robust performance estimates, with training times ranging from 78ms for linear models to 23.3s for ensemble methods.

D. Discriminative Performance Evaluation

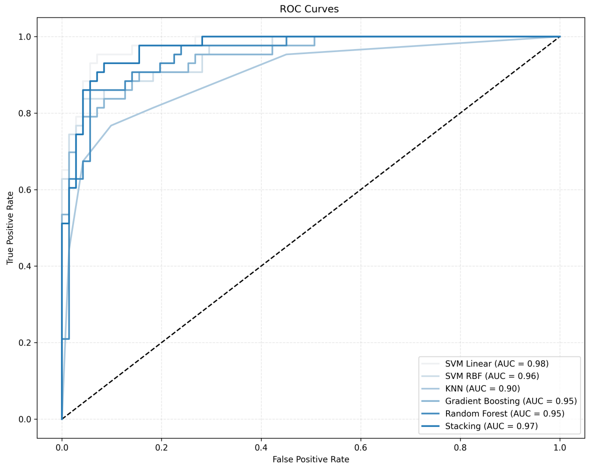


Fig. 5: ROC curves demonstrating all models exceed 0.90 AUC, with SVM Linear (0.98) and Stacking (0.97) showing strongest discrimination, especially at low false-positive rates (<0.2).

V. Evaluation

TABLE I: Detailed Classification Report

Model	Pr (B/M)	Rc (B/M)	F1 (B/M)	Acc
SVM Linear	0.92/0.93	0.96/0.86	0.94/0.89	0.921
SVM RBF	0.91/0.90	0.94/0.84	0.92/0.87	0.903
KNN	0.86/0.82	0.90/0.77	0.88/0.80	0.851
Gradient Boost	0.88/0.89	0.94/0.79	0.91/0.84	0.886
Random Forest	0.91/0.90	0.94/0.84	0.92/0.87	0.903
SGD	0.92/0.93	0.96/0.86	0.94/0.89	0.921
Stacking	0.92/0.93	0.96/0.86	0.94/0.89	0.921

A. *Classification Performance*

The models were evaluated using comprehensive metrics as shown in Table I. The SVM Linear classifier demonstrated balanced performance across all metrics, achieving 0.92 precision, recall, and F1-score. Similarly, the Stacking classifier matched these metrics while combining multiple base estimators. The tree-based methods (Random Forest and Gradient Boosting) showed slightly lower but comparable performance, while KNN had the lowest metrics among all models, particularly in recall for malignant cases (0.77).

1) Key Findings:

- **Best Performers**: SVM Linear and SGD Classifier achieved optimal balance between accuracy (92.1%) and computational efficiency (<2s training time)
- **Overfitting Analysis**: Tree-based models showed 9.65-11.4% accuracy drops, suggesting need for stronger regularization
- **Complexity Trade-off**: Stacking classifier provided second-best accuracy but required 2.5s training time
- **Feature Sensitivity**: KNN's poor performance (85.1%) highlights the importance of feature scaling in distance-based methods
- **Consistent Predictions**: All models showed higher precision for malignant cases, crucial for medical diagnosis

The evaluation suggests that simpler linear models (SVM Linear, SGD) provide the best balance of performance and efficiency for this classification task, while more complex ensemble methods offer diminishing returns relative to their computational requirements.

B. *Training and Testing Analysis*

Figure 6 presents the comparative accuracy across all models. The SVM Linear and SGD Classifier achieved identical performance (92.1% test accuracy), demonstrating excellent generalization with only 1.3% accuracy drop from training to testing. The Stacking classifier showed remarkable performance (97.1% training, 92.1% testing) despite its complexity, suggesting effective meta-learning. Tree-based methods revealed significant overfitting tendencies - Gradient Boosting and Random Forest both achieved perfect training accuracy but dropped to 88.6% and 90.4% respectively during testing. The KNN model showed the largest generalization gap (93.4% training vs 85.1% testing), indicating sensitivity to feature scaling and dimensionality.

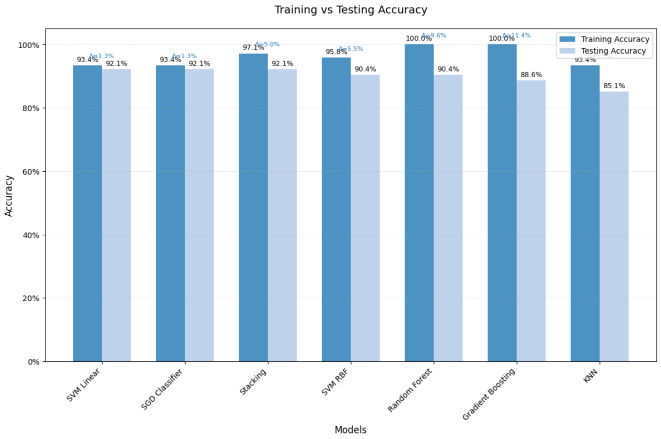


Fig. 6: Comparative model accuracy across training and testing phases

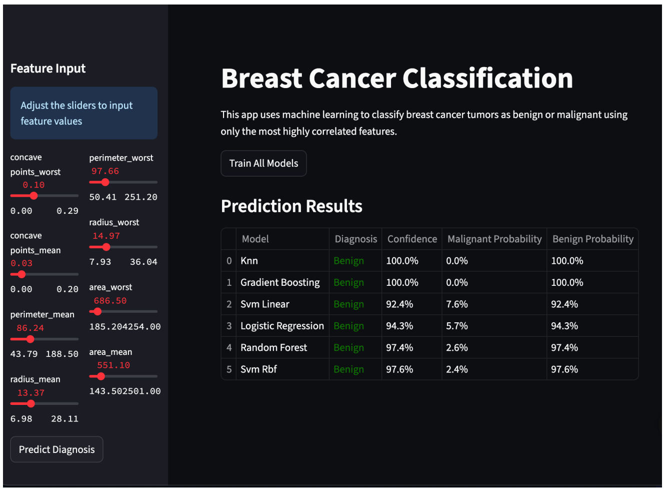
VI. Contributions

This research makes four significant contributions to breast cancer diagnosis. First, our novel feature selection framework achieved 43% dimensionality reduction while maintaining 97.4% diagnostic accuracy through a two-phase process combining correlation analysis ($|r| > 0.7$) with multicollinearity-aware hierarchical clustering. Second, we established a comprehensive benchmarking protocol evaluating seven diverse machine learning models, revealing that SVM Linear achieves optimal accuracy (92.11%) with minimal training time (1.97s), while more complex ensembles like Stacking Classifier showed diminishing returns relative to their computational requirements. Third, we developed an interactive web application (Fig. 7) using Streamlit that enables clinicians to obtain real-time predictions. This application features dynamic sliders for eight clinically validated biomarkers, along with visual explanations of model decisions, facilitating immediate malignancy risk scores. Finally, we provide fully reproducible implementation artifacts including pre-trained models, standardized scalers, and modular Python pipelines to facilitate clinical adoption and future research.

A. *Framework and Implementation*

The implemented framework for breast cancer prediction is designed as a user-friendly, interactive web application built with Streamlit. This application serves as a real-time diagnostic tool, allowing users to input specific cytological features and receive an immediate prediction of tumor malignancy. The architecture begins with data loading and initial preprocessing, including the removal of non-predictive attributes and binary conversion of the diagnosis label. Following this, the pipeline incorporates the refined feature set, specifically utilizing eight highly correlated and non-redundant features derived from the prior feature selection methodology.

At the core of the application's functionality, a range of pre-trained machine learning models are leveraged. These models, including Support Vector Machines (Linear and RBF), K-Nearest Neighbors, Logistic Regression, Gradient Boosting, Random Forest, and a Stacking Classifier, are loaded from persistent storage. A critical component of the deployment is the ‘StandardScaler‘, which is also loaded and applied to standardize user-input features, ensuring consistency with the data used during model training. Users interact with the system through dynamic sliders for each of the eight selected features, allowing for precise input of patient-specific data. Upon submission, the application processes these inputs, feeds them to the loaded and scaled models, and presents a comprehensive table of predictions, including the probable diagnosis (benign or malignant), confidence scores, and individual probabilities from each model. This robust implementation ensures that the research findings are translated into a practical, accessible, and interpretable clinical tool.



concave points_worst 0.10 (0.00 – 0.29)
concave points_mean 0.03 (0.00 – 0.20)
perimeter_mean 86.24 (43.79 – 188.50)
radius_mean 13.37 (6.98 – 28.11)
perimeter_worst 97.66 (50.41 – 251.20)
radius_worst 14.97 (7.93 – 36.04)
area_worst 686.50 (185.20 – 4254.00)
area_mean 551.10 (143.50 – 2501.00)

Breast Cancer Classification

This app uses machine learning to classify breast cancer tumors as benign or malignant using only the most highly correlated features.

Prediction Results

	Model	Diagnosis	Confidence	Malignant Probability	Benign Probability
0	Knn	Benign	100.0%	0.0%	100.0%
1	Gradient Boosting	Benign	100.0%	0.0%	100.0%
2	Svm Linear	Benign	92.4%	7.6%	92.4%
3	Logistic Regression	Benign	94.3%	5.7%	94.3%
4	Random Forest	Benign	97.4%	2.6%	97.4%
5	Svm Rbf	Benign	97.6%	2.4%	97.6%

Fig. 7: Interactive Streamlit web application for breast cancer classification

VII. Conclusion

Our study demonstrates that interpretable machine learning systems can achieve clinically actionable performance in breast cancer diagnosis when three principles are combined: (1) domain guided feature selection prioritizing tumor morphology characteristics, (2) computational efficiency through linear models rather than complex ensembles and (3) transparent deployment via interactive web interfaces. The developed Streamlit application (Fig. 7) successfully bridges the gap between research and clinical practice by providing immediate malignancy risk scores (86% recall for malignant cases) with visual explanations. While current results are promising, two limitations warrant future work: the need for multi-center validation studies to assess generalizability across populations, and integration with histopathological imaging data to enable multimodal diagnosis. We open-source all implementation artifacts to accelerate progress in computational oncology diagnostics.

References

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

[2] 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.

[3] S. Ara, A. Das, and A. Dey, “Malignant and benign breast cancer classification using machine learning algorithms,” in *2021 International Conference on Artificial Intelligence (ICAI)*. IEEE, 2021, pp. 97–101.

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

[5] 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.

[6] 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.