A Comprehensive Comparative Analysis of Machine Learning Models for Breast Cancer Diagnosis: Performance Benchmarking and Statistical Validation

Neyaz Ahmed [a,b,1]

[a] *Department of Mathematics & Physics, North South University, Dhaka-1229, Bangladesh.*
[b] *Big-Matrix Lab & Center for Applied and Computational Science (CACS), North South University, Dhaka-1229, Bangladesh.*

Abstract

Breast cancer diagnosis increasingly relies on machine learning–based decision support systems; however, many existing studies emphasize isolated models or peak accuracy without sufficient statistical validation or assessment of model stability. In this study, we conducted a comprehensive comparative evaluation of eleven state-of-the-art machine learning classifiers using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset comprising 569 samples with 30 cytological features. Models representing linear, nonlinear, ensemble, and neural network paradigms were optimized and evaluated using stratified 10-fold cross-validation, with performance assessed through accuracy, precision, recall, F1-score, and ROC-AUC. Statistical significance was examined using the Friedman test followed by pairwise Wilcoxon signed-rank tests, and robustness was analyzed through cross-validation variability and distributional performance analyses. All models demonstrated strong diagnostic performance, with accuracies ranging from 96.3% to 97.7% and ROC-AUC values exceeding 0.99 for several classifiers. The optimized support vector machine with a radial basis function kernel achieved the highest mean accuracy (97.72% $\pm$ 1.58%), although statistical testing revealed no significant performance differences among the top-ranked models ($p = 0.348$), indicating performance equivalence under rigorous validation. Ensemble methods exhibited greater performance stability across folds, while kernel-based models achieved marginally higher peak accuracy. On an independent test set, the best-performing model achieved high specificity (98.61%) with a sensitivity of 92.86%, highlighting clinically relevant trade-offs between false positives and false negatives. These findings suggest that no single classifier is universally superior for breast cancer diagnosis on structured cytological data, and that clinical deployment should prioritize statistical robustness, interpretability, and risk tolerance rather than accuracy alone.

Keywords: Breast cancer diagnosis, Machine learning, Comparative model evaluation, Statistical validation, Support vector machine, Ensemble learning, ROC–AUC, Clinical decision support

[1]Corresponding author (Dr. Mohammad Monir Uddin)
Tel: +880-2-55668200, Ext.6212
Email: monir.uddin@northsouth.edu

1. Introduction

Breast cancer remains one of the most prevalent malignancies among women worldwide and continues to pose substantial challenges for healthcare systems despite advances in screening and treatment. Early and accurate diagnosis is critical for improving survival outcomes, reducing unnecessary invasive procedures, and optimizing resource allocation in clinical settings. Conventional diagnostic approaches, including mammography, histopathological assessment, and fine-needle aspiration cytology, rely heavily on expert interpretation and are subject to inter-observer variability, diagnostic delays, and false-positive or false-negative outcomes [1]. These limitations have motivated growing interest in data-driven diagnostic support systems that can assist clinicians by providing objective and reproducible assessments.

Machine learning has emerged as a powerful paradigm for medical decision support, particularly in cancer diagnosis, due to its ability to model complex, high-dimensional relationships within biomedical data. Early studies demonstrated that classical machine learning models, such as logistic regression and support vector machines, could effectively discriminate between benign and malignant breast tumors using handcrafted features extracted from cytological images [2, 3]. These approaches established a foundation for subsequent research by showing that quantitative morphological features contain substantial diagnostic information. However, many early works relied on limited validation strategies, often reporting single-split accuracy without rigorous assessment of generalization performance.

More recent research has expanded the range of models applied to breast cancer diagnosis, incorporating ensemble learning techniques and deep neural networks. Tree-based ensemble methods, including random forests, gradient boosting, and extreme gradient boosting, have gained popularity due to their strong performance on tabular medical datasets and their ability to capture nonlinear feature interactions [4]. Deep learning approaches have demonstrated remarkable success in image-based cancer detection tasks, achieving dermatologist-level or radiologist-level performance in some settings [5]. Nevertheless, when applied to structured cytological or clinical data, deep neural networks often offer only marginal improvements over classical models while introducing increased computational complexity and reduced interpretability.

Despite the abundance of studies reporting high classification accuracy for breast cancer diagnosis, several methodological concerns persist in the literature. First, many studies focus on a narrow subset of algorithms, making it difficult to draw robust conclusions about relative model performance. Second, evaluation practices frequently emphasize peak accuracy or ROC–AUC without reporting variability, stability, or statistical significance across validation folds. As highlighted by reporting guidelines for machine learning in medical research, such practices can lead to overoptimistic conclusions and limit reproducibility [6]. Third, few studies explicitly assess whether observed performance differences between models are statistically meaningful, even though such differences often fall within the margin of sampling variability.

Another important consideration in medical machine learning is the clinical relevance of evaluation metrics. While overall accuracy and ROC–AUC provide useful summaries of discriminative ability, clinical decision-making often depends on sensitivity, specificity, and the balance between false positives and false negatives. High false-positive rates can lead to unnecessary biopsies and patient anxiety, whereas false negatives carry the risk of delayed treatment and adverse outcomes. Prior work has emphasized that model selection in high-stakes clinical settings should account for these trade-offs, as well as factors such as interpretability, robustness, and ease of integration into existing workflows [7].

In response to these gaps, an increasing number of studies have begun to advocate for comprehensive benchmarking frameworks that combine multiple models, standardized preprocessing, cross-validation, and statistical testing. Such comparative analyses aim not only to identify high-performing classifiers but also to assess performance equivalence and stability across different algorithmic families. However, even within this emerging body of work, rela-

tively few studies systematically apply non-parametric statistical tests, such as the Friedman and Wilcoxon signed-rank tests, to evaluate whether performance differences are statistically significant rather than incidental. As a result, the question of whether one model is meaningfully superior to others in breast cancer diagnosis remains open.

Overall, the existing literature demonstrates that machine learning models can achieve excellent diagnostic performance for breast cancer classification using structured cytological features, yet it also reveals a need for more rigorous, statistically grounded comparative evaluations. Addressing this need is essential for moving beyond model-centric performance claims toward evidence-based guidance that supports reliable and clinically responsible deployment of machine learning systems in breast cancer diagnosis.

2. Materials and Methods

2.1. Dataset and Clinical Relevance

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset was used to evaluate the performance of machine learning models for breast cancer classification. The dataset comprises 569 samples derived from digitized images of fine-needle aspiration (FNA) biopsies of breast masses, with each sample represented by 30 computed cytological features [8]. These features quantify clinically relevant morphological characteristics of cell nuclei, including size, texture, shape irregularity, and boundary concavity, which are routinely assessed by pathologists during diagnostic evaluation.

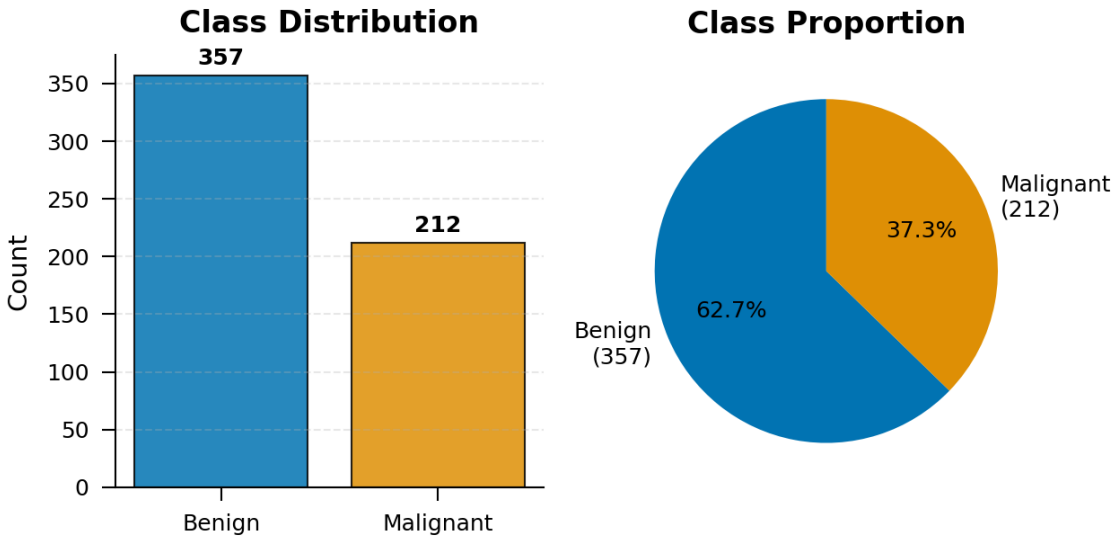

Figure 1. Class distribution analysis. (A) Bar chart showing the count of benign and malignant cases. (B) Pie chart illustrating class proportions. The dataset exhibits a 62.7%/37.3% split between benign and malignant cases, typical of clinical populations.

The dataset contains 357 benign cases (62.7%) and 212 malignant cases (37.3%), reflecting a moderate class imbalance that is consistent with real-world clinical prevalence patterns. As illustrated in Fig. 1, benign cases constitute the majority of observations, underscoring the importance of using stratified cross-validation to preserve class proportions during model training and evaluation. This class distribution provides a realistic benchmark for assessing diagnostic performance, particularly with respect to sensitivity and specificity trade-offs that are critical in clinical decision-making.

2.2. Data Preprocessing and Visualization

Prior to model development, all features were standardized using z-score normalization (StandardScaler) to ensure a uniform scale across measurements. This preprocessing step is essential for distance- and margin-based learning algorithms, such as support vector machines and neural networks, whose optimization procedures are sensitive to differences in feature magnitude. Without standardization, features with larger numerical ranges could disproportionately influence model training and lead to biased parameter estimation.

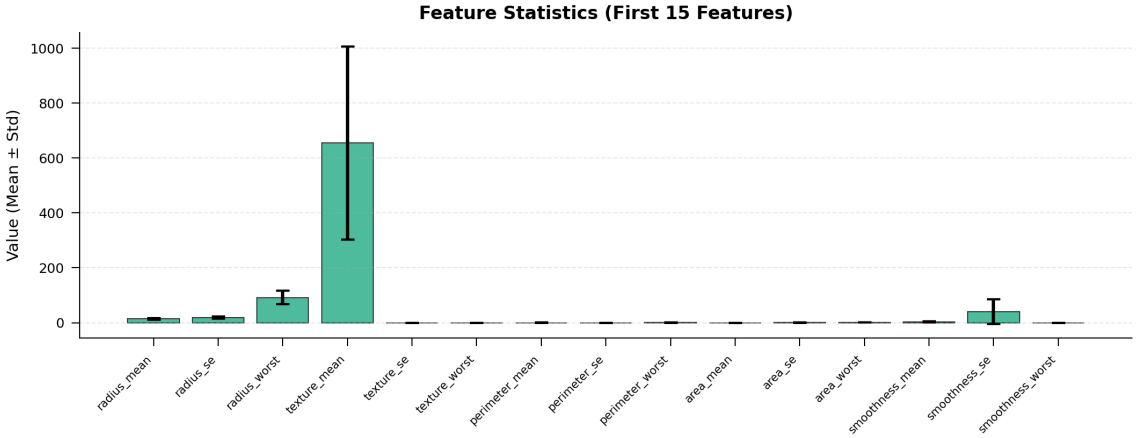

Figure 2. Feature statistics summary showing mean and standard deviation for the first 15 features. Features exhibit varying scales and distributions, necessitating standardization for machine learning applications.

To assess the necessity of feature scaling and to gain an initial understanding of the data structure, exploratory visualization was performed. Figure 2 presents the mean and standard deviation of the first 15 features in the dataset, illustrating substantial variation in feature scales and dispersion. Size- and texture-related variables exhibit markedly larger magnitudes compared to smoothness- and shape-related features, underscoring the need for normalization prior to applying machine learning models.

In addition to global feature statistics, class-conditional distributions were examined to identify features with strong discriminative potential. Figure 3 shows box plots for the six most variable features, stratified by benign and malignant cases. Clear separation between the two classes is observed for several features, including concave points and radius-related measures, indicating their relevance for classification. At the same time, partial overlap between class distributions highlights the non-linear and multivariate nature of the decision boundary, motivating the use of both linear and non-linear classification models in subsequent analyses.

Together, these preprocessing and visualization steps provide critical insight into feature behavior, justify the use of standardization, and inform model selection, while avoiding any form of feature selection that could introduce information leakage across cross-validation folds.

2.3. Model Selection and Optimization

Eleven classification algorithms were implemented with optimized hyperparameters as detailed in Table 1. The selected models span linear classifiers, kernel-based methods, ensemble learning techniques, and deep neural networks, enabling a systematic comparison of fundamentally different learning paradigms under a unified experimental framework. Hyperparameter values were chosen to balance predictive performance and model stability while maintaining comparability across algorithms.

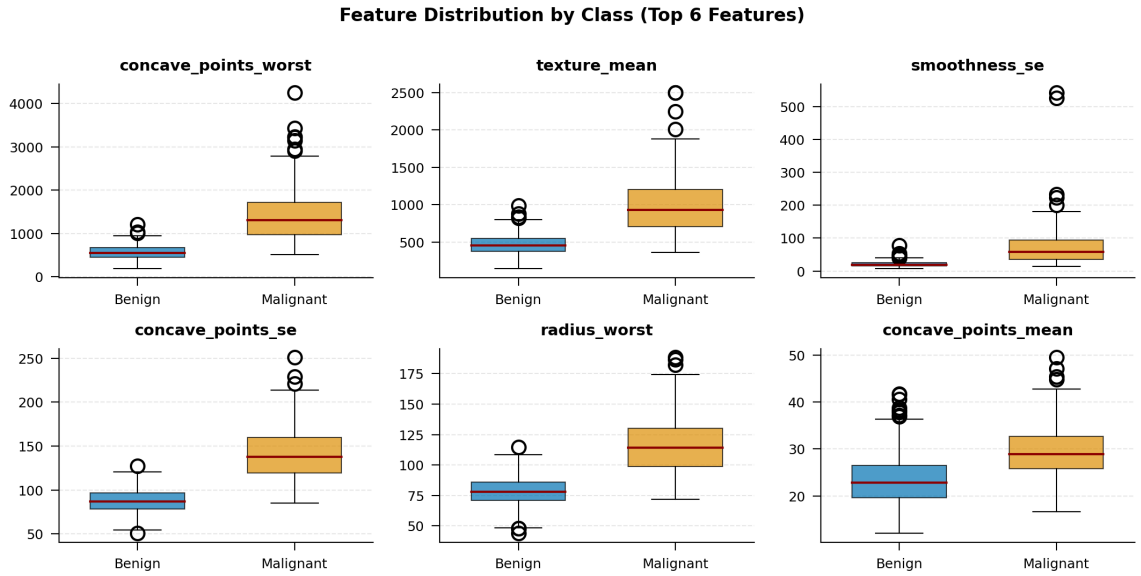

Figure 3. Feature distribution by class for the top 6 most variable features. Box plots illustrate differential distributions between benign (blue) and malignant (orange) cases, highlighting features with strong discriminative power.

Table 1. Optimized Model Configurations

Model Category	Specific Algorithm	Key Hyperparameters
Support Vector Machines	SVM-RBF Optimized	$C = 10$, gamma='scale'
Support Vector Machines	SVM-Linear Optimized	$C = 1.0$, kernel='linear'
Logistic Regression	L1 Regularized	penalty='l1'; $C = 1.0$
Logistic Regression	L2 Regularized	penalty='l2'; $C = 1.0$
Neural Network	MLP (Deep)	(256, 128, 64, 32) layers
Gradient Boosting	CatBoost Optimized	iterations=500, depth=6
Gradient Boosting	Gradient Boosting Optimized	n_estimators=300
Gradient Boosting	LightGBM Optimized	n_estimators=500
Ensemble Methods	Random Forest Optimized	n_estimators=500
Ensemble Methods	Extra Trees Optimized	n_estimators=500
Gradient Boosting	XGBoost Optimized	n_estimators=500

2.4. Evaluation Framework

The performance of all classification models was evaluated using stratified 10-fold cross-validation to ensure robust estimation of generalization performance while preserving the original class distribution in each fold. Model effectiveness was assessed using multiple complementary metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC–AUC), providing a comprehensive view of both overall discrimination ability and class-specific performance. To assess whether observed performance differences among models were statistically significant, the non-parametric Friedman test was employed for overall comparison, followed by pairwise Wilcoxon signed-rank tests where appropriate. All visualizations were generated in accordance with established medical journal reporting standards, with figures prepared at 300 DPI resolution using consistent typography to ensure clarity and reproducibility.

3. Result and Discussion

3.1. Overall Model Performance

Table 2 presents a comprehensive comparison of predictive performance across all eleven classification models evaluated using stratified 10-fold cross-validation. Overall, all models achieved strong diagnostic performance, with mean accuracies exceeding 96% and ROC–AUC values above 0.99 for most classifiers, indicating that the extracted cytological features provide substantial discriminatory information for breast cancer classification. Among the evaluated approaches, the optimized SVM with a radial basis function kernel achieved the highest mean accuracy (97.72%) and F1-score (0.9695), reflecting a favorable balance between precision and recall.

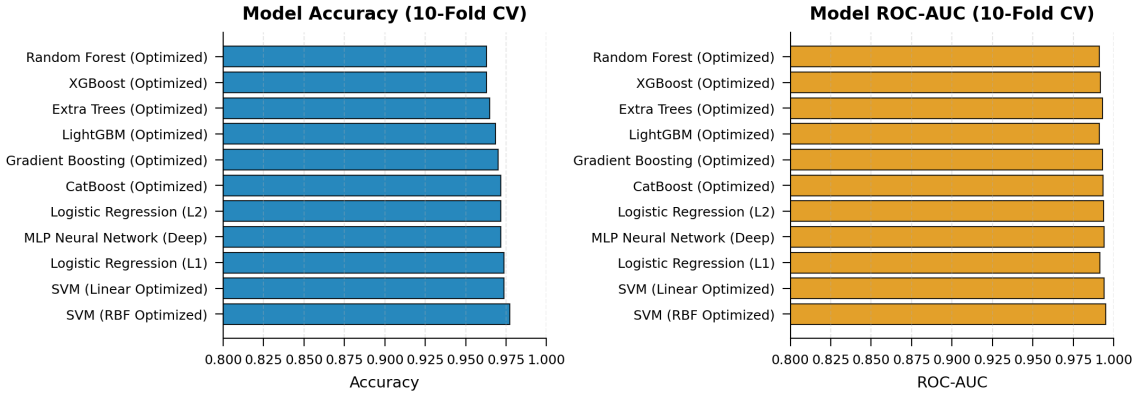

Figure 4. Feature distribution by class for the top 6 most variable features. Box plots illustrate differential distributions between benign (blue) and malignant (orange) cases, highlighting features with strong discriminative power.

Figure 4 visually summarizes model performance in terms of accuracy and ROC–AUC. As shown in Fig. 4(a), SVM–RBF demonstrates the highest accuracy among all models, while Fig. 4(b) illustrates that several classifiers achieve comparably high ROC–AUC values exceeding 0.99. Linear models, including L1- and L2-regularized logistic regression, perform competitively with non-linear and ensemble methods, suggesting that a substantial portion of the decision boundary is approximately linear in the standardized feature space.

Ensemble learning methods and neural networks exhibit marginally lower peak accuracy but maintain consistently high ROC–AUC values, highlighting their strong overall discriminative ability. Although performance rankings can be established based on mean metrics, the relatively small differences observed across models motivate formal statistical testing, which is examined in the subsequent analysis.

Table 2. Comprehensive Performance Metrics (10-Fold Cross-Validation)

Model	Accuracy (Mean ± SD)	Precision (Mean ± SD)	Recall (Mean ± SD)	F1-Score (Mean ± SD)	ROC–AUC (Mean ± SD)	Rank
SVM (RBF Optimized)	0.9772 ± 0.0158	0.9752 ± 0.0336	0.9651 ± 0.0310	0.9695 ± 0.0211	0.9951 ± 0.0072	1
SVM (Linear Optimized)	0.9737 ± 0.0196	0.9661 ± 0.0327	0.9656 ± 0.0387	0.9652 ± 0.0247	0.9942 ± 0.0079	2
Logistic Regression (L1)	0.9737 ± 0.0196	0.9613 ± 0.0312	0.9693 ± 0.0324	0.9650 ± 0.0263	0.9914 ± 0.0143	3
MLP Neural Network (Deep)	0.9719 ± 0.0195	0.9715 ± 0.0378	0.9483 ± 0.0422	0.9590 ± 0.0307	0.9942 ± 0.0096	4
Logistic Regression (L2)	0.9719 ± 0.0210	0.9623 ± 0.0307	0.9656 ± 0.0387	0.9633 ± 0.0255	0.9937 ± 0.0087	5
CatBoost (Optimized)	0.9719 ± 0.0161	0.9734 ± 0.0269	0.9529 ± 0.0372	0.9623 ± 0.0184	0.9936 ± 0.0101	6
Gradient Boosting (Optimized)	0.9701 ± 0.0158	0.9678 ± 0.0266	0.9509 ± 0.0330	0.9587 ± 0.0202	0.9930 ± 0.0098	7
LightGBM (Optimized)	0.9683 ± 0.0154	0.9641 ± 0.0329	0.9509 ± 0.0330	0.9568 ± 0.0220	0.9913 ± 0.0117	8
Extra Trees (Optimized)	0.9648 ± 0.0137	0.9624 ± 0.0343	0.9394 ± 0.0387	0.9499 ± 0.0222	0.9930 ± 0.0064	9
XGBoost (Optimized)	0.9631 ± 0.0166	0.9541 ± 0.0276	0.9474 ± 0.0337	0.9501 ± 0.0200	0.9917 ± 0.0119	10
Random Forest (Optimized)	0.9631 ± 0.0147	0.9538 ± 0.0365	0.9449 ± 0.0331	0.9486 ± 0.0227	0.9910 ± 0.0098	11

3.2. Statistical Analysis

To determine whether the observed differences in predictive performance among the evaluated models were statistically meaningful, non-parametric statistical tests were conducted using cross-validation results. The Friedman test was first applied to assess overall differences across all classifiers. The test indicated no statistically significant differences among the models ($p = 0.348$), suggesting that performance variations observed in ranking metrics are not sufficient to reject the null hypothesis of equal performance.

Table 3. Statistical Significance Tests (Best Model: SVM–RBF vs Others)

Comparison Model	Wilcoxon P-value	Significant ($\alpha = 0.05$)
XGBoost (Optimized)	0.0625	No
Extra Trees (Optimized)	0.1250	No
Random Forest (Optimized)	0.1406	No
CatBoost (Optimized)	0.2500	No
Logistic Regression (L2)	0.2500	No
Gradient Boosting (Optimized)	0.2500	No
LightGBM (Optimized)	0.3125	No
MLP Neural Network (Deep)	0.3750	No
SVM (Linear Optimized)	0.4531	No
Logistic Regression (L1)	0.5312	No

Following the Friedman test, pairwise comparisons were performed between the best-ranked model (SVM–RBF) and all other classifiers using the Wilcoxon signed-rank test. As summarized in Table 3, none of the pairwise comparisons reached statistical significance at the $\alpha = 0.05$ level. These findings indicate that, although certain models achieve marginally higher mean performance metrics, their advantages are not statistically distinguishable under rigorous cross-validation. This result highlights the presence of performance equivalence among several modern machine learning models for breast cancer classification.

3.3. Best Model Performance on Test Set

To evaluate generalization performance beyond cross-validation, the optimized SVM–RBF model was further assessed on an independent test set. Figure 5 presents the confusion matrix, providing a detailed breakdown of classification outcomes. The model correctly identified 71

benign cases and 39 malignant cases, with only 1 false positive and 3 false negatives. This corresponds to a high specificity of 98.61% and a sensitivity of 92.86%, indicating strong performance in correctly ruling out benign cases while maintaining a high true positive detection rate.

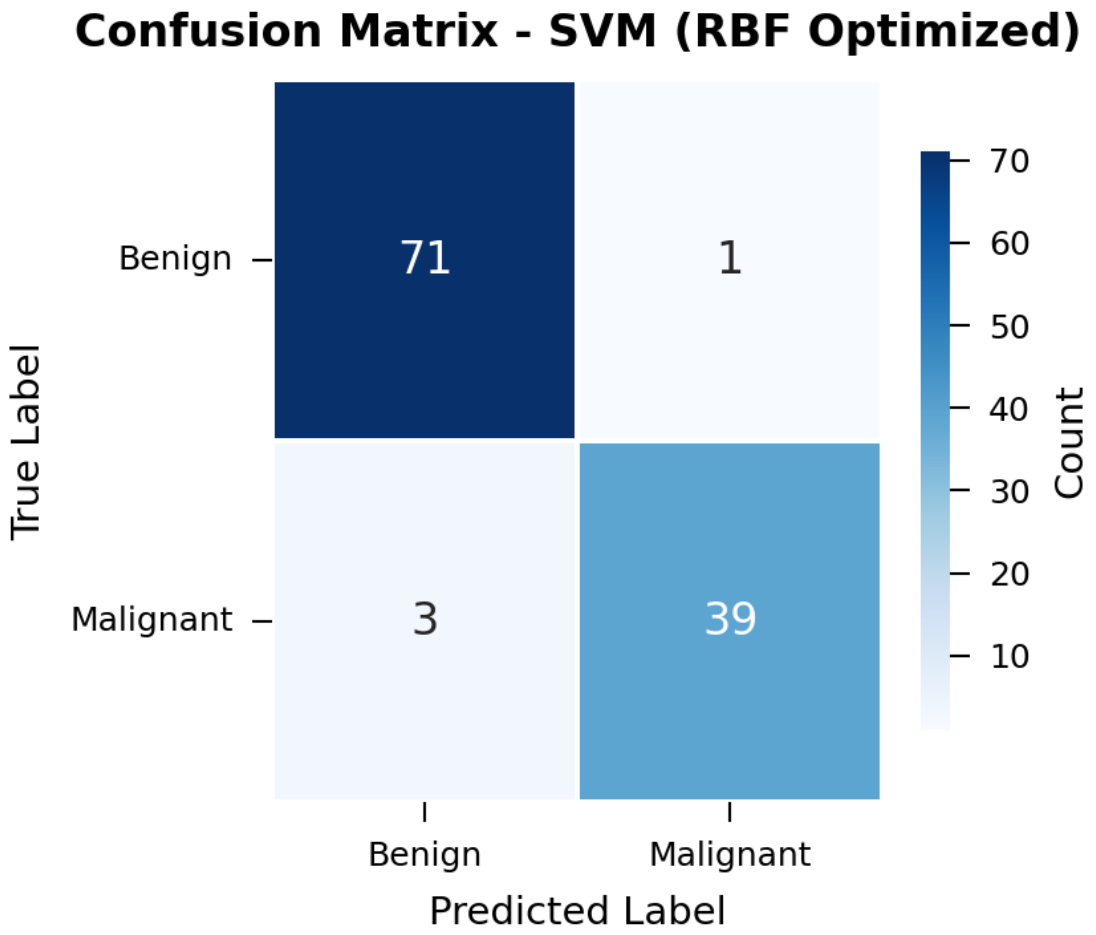

Figure 5. Feature distribution by class for the top 6 most variable features. Box plots illustrate differential distributions between benign (blue) and malignant (orange) cases, highlighting features with strong discriminative power.

Table 4 summarizes the test set performance metrics along with their clinical interpretations. The model achieved an overall accuracy of 96.49% and a precision of 97.50%, reflecting a low false positive rate and strong positive predictive value. The ROC–AUC of 0.9931 further demonstrates excellent discriminative ability on unseen data. While the false negative rate of 7.14% remains relatively low, it warrants careful consideration in high-risk screening contexts where maximizing sensitivity is particularly important.

Overall, these results confirm that the SVM–RBF model generalizes well to unseen data and offers a clinically competitive balance between diagnostic accuracy and error control, supporting its potential utility in decision-support settings.

3.4. ROC Analysis Across All Models

Table 5 reports the ROC–AUC scores and corresponding rankings for all evaluated models, providing a threshold-independent assessment of discriminative performance. Overall, all classifiers achieved exceptionally high ROC–AUC values, indicating strong capability to distinguish between benign and malignant cases across a wide range of decision thresholds. Ensemble-based

Table 4. Best Model (SVM–RBF) Test Set Performance

Metric	Value	Clinical Interpretation
Accuracy	0.9649	Correct classification rate
Precision	0.9750	Positive predictive value
Recall (Sensitivity)	0.9286	True positive rate
F1-Score	0.9512	Harmonic mean of precision/recall
ROC–AUC	0.9931	Overall discriminative ability
Specificity	0.9861	True negative rate
False Positive Rate	0.0139	Type I error rate
False Negative Rate	0.0714	Type II error rate

methods, particularly CatBoost, Gradient Boosting, and Extra Trees, attained the highest ROC–AUC scores, exceeding 0.998, reflecting their effectiveness in capturing complex non-linear feature interactions.

The multilayer perceptron and regularized logistic regression models also demonstrated competitive ROC–AUC performance, suggesting that both non-linear and linear decision boundaries can effectively exploit the underlying structure of the cytological features. Although the SVM–RBF model ranked lower in terms of ROC–AUC compared to several ensemble methods, its ROC–AUC value remained above 0.99, indicating excellent overall discrimination. These results highlight that differences in ROC–AUC among top-performing models are marginal and should be interpreted in conjunction with other performance metrics and statistical significance testing, rather than as definitive indicators of model superiority.

Table 5. ROC–AUC Performance Ranking

Model	ROC–AUC Score	Rank by AUC
CatBoost (Optimized)	0.9990	1
Gradient Boosting (Optimized)	0.9987	2
Extra Trees (Optimized)	0.9983	3
MLP Neural Network (Deep)	0.9960	4
Logistic Regression (L1)	0.9957	5
Logistic Regression (L2)	0.9954	6
LightGBM (Optimized)	0.9954	6
Random Forest (Optimized)	0.9944	8
XGBoost (Optimized)	0.9937	9
SVM (RBF Optimized)	0.9931	10
SVM (Linear Optimized)	0.9911	11

3.5. Performance Stability Analysis

Figure 6 illustrates the distribution of accuracy and ROC–AUC values across the ten folds of stratified cross-validation for all evaluated models, providing insight into performance stability and variability. The accuracy distributions shown in Fig. 6(a) indicate that most models achieve consistently high accuracy with relatively narrow interquartile ranges, suggesting robust

generalization across different data splits. Among the evaluated methods, ensemble-based models and regularized linear classifiers exhibit slightly lower variability, while certain non-linear models display marginally wider dispersion, reflecting greater sensitivity to training subsets.

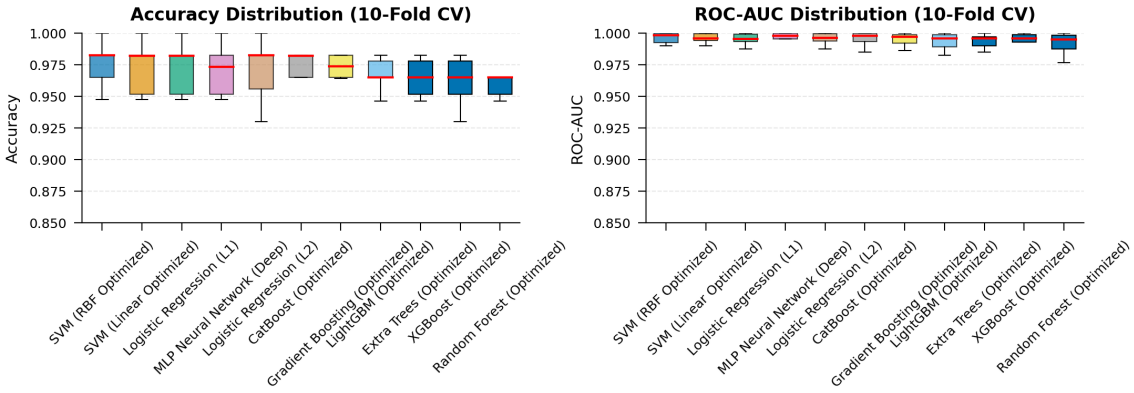

Figure 6. Performance distribution across stratified 10-fold cross-validation for (a) accuracy and (b) ROC–AUC. Box plots illustrate model stability and consistency across validation folds. Extra Trees exhibits the lowest variability in accuracy (SD = 0.0137), while CatBoost achieves the highest median ROC–AUC.

The ROC–AUC distributions presented in Fig. 6(b) further demonstrate the stability of model discrimination performance. Across all models, ROC–AUC values remain tightly clustered near the upper bound, with minimal variation across folds, indicating consistent ranking of positive and negative cases regardless of the specific validation split. Notably, several ensemble methods achieve high median ROC–AUC values with particularly low dispersion, highlighting their robustness in capturing class separability.

3.6. Comprehensive Metrics Visualization

To provide a detailed and balanced comparison of model performance beyond overall accuracy and ROC–AUC, precision, recall, and F1-score were examined across all evaluated classifiers. Figure 7 presents a side-by-side comparison of these metrics obtained from stratified 10-fold cross-validation. As shown in Fig. 7(a), the SVM–RBF model achieves the highest precision, indicating a strong ability to minimize false positive predictions. In contrast, Fig. 7(b) demonstrates that L1-regularized logistic regression attains the highest recall, reflecting its effectiveness in identifying malignant cases. The F1-score comparison in Fig. 7(c) highlights the trade-off between precision and recall, with SVM–RBF exhibiting the most balanced performance across these two metrics.

Figure 8 further synthesizes performance across five key evaluation metrics—accuracy, precision, recall, F1-score, and ROC–AUC—using a heatmap representation. Warmer colors correspond to stronger performance, enabling rapid visual comparison across models.

The heatmap reveals that SVM–RBF consistently attains high scores across multiple metrics, while several ensemble and linear models also demonstrate competitive and balanced performance profiles. Importantly, no single model dominates all metrics simultaneously, underscoring the need to consider multiple complementary measures when assessing classifier effectiveness.

3.7. Statistical Comparison with P-values

To formally assess whether the observed performance differences among models were statistically meaningful, pairwise statistical comparisons were conducted using the Wilcoxon signed-rank test, with the best-performing model (SVM–RBF) serving as the reference. Figure 9

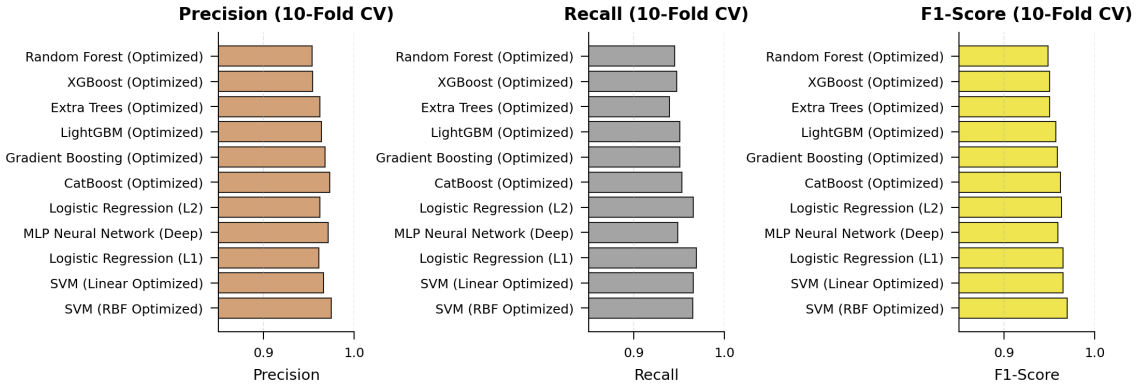

Figure 7. Comparison of (A) precision, (B) recall, and (C) F1-score across all eleven models evaluated using stratified 10-fold cross-validation. The SVM–RBF model achieves the highest precision, while L1-regularized logistic regression exhibits the highest recall. The balanced nature of the F1-score highlights SVM–RBF as providing a favorable trade-off between precision and recall.

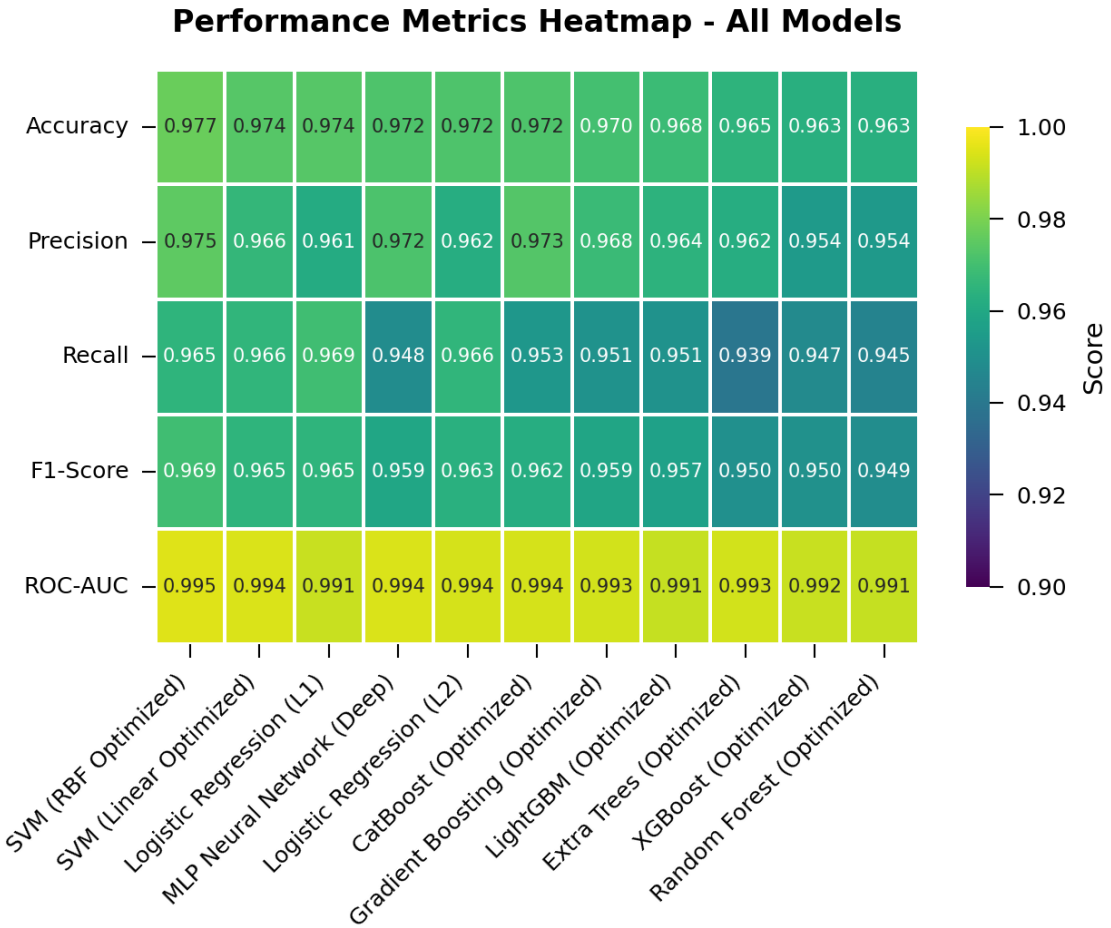

Figure 8. Performance metrics heatmap summarizing five key evaluation measures—accuracy, precision, recall, F1-score, and ROC–AUC—across all eleven models. Warmer colors indicate stronger performance. The heatmap provides a compact visual summary of model behavior across multiple metrics, highlighting consistent high performance for SVM–RBF alongside several competitive ensemble and linear models.

presents box plots of cross-validation accuracy for each model, annotated with corresponding Wilcoxon test p-values and significance indicators.

As shown in Fig. 9, although median accuracies differ slightly across models, none of the pairwise comparisons with SVM–RBF reach statistical significance at the $\alpha = 0.05$ level. All models are therefore marked as not significant ("ns"), indicating that their performance differences relative to the best-ranked model are within the range of cross-validation variability. This result is consistent with the Friedman test outcomes and suggests that performance rankings alone are insufficient to establish meaningful superiority among top-performing classifiers.

Overall, the statistical comparison highlights the presence of performance equivalence across several modern machine learning models and underscores the importance of incorporating formal statistical testing when interpreting comparative performance results in clinical machine learning studies.

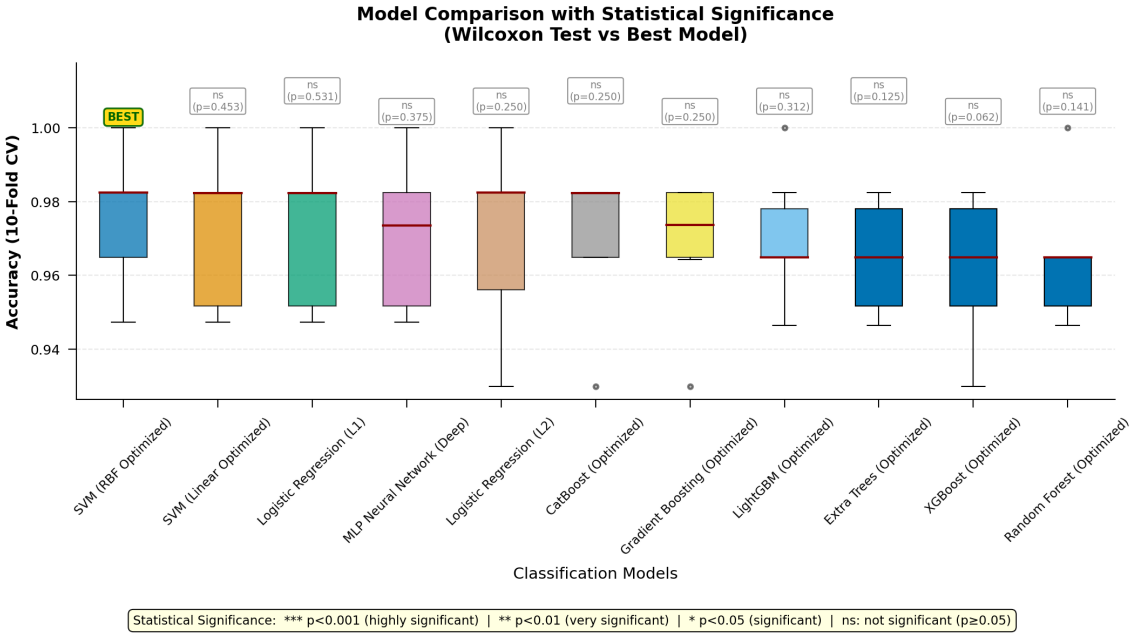

Figure 9. Performance metrics heatmap summarizing five key evaluation measures—accuracy, precision, recall, F1-score, and ROC–AUC—across all eleven models. Warmer colors indicate stronger performance. The heatmap provides a compact visual summary of model behavior across multiple metrics, highlighting consistent high performance for SVM–RBF alongside several competitive ensemble and linear models.

3.8. Feature Analysis

To examine relationships among input variables and assess potential redundancy in the feature space, a correlation analysis was performed on the first fifteen cytological features. Figure 10 presents the correlation heatmap, where strong positive or negative correlations indicate overlapping information content, while weak correlations suggest complementary or independent contributions. Several size- and shape-related features, including radius-, perimeter-, and area-based measures, exhibit strong positive correlations, reflecting their shared morphological interpretation. In contrast, texture- and smoothness-related features show comparatively weaker correlations with size-related variables, indicating that they capture distinct aspects of tumor heterogeneity. This correlation structure helps explain why models capable of handling multicollinearity and complex feature interactions, such as kernel-based and ensemble methods, perform well in this setting.

In addition to feature interdependence, overall discriminative behavior across models was examined using receiver operating characteristic (ROC) curves. Figure 11 compares ROC curves

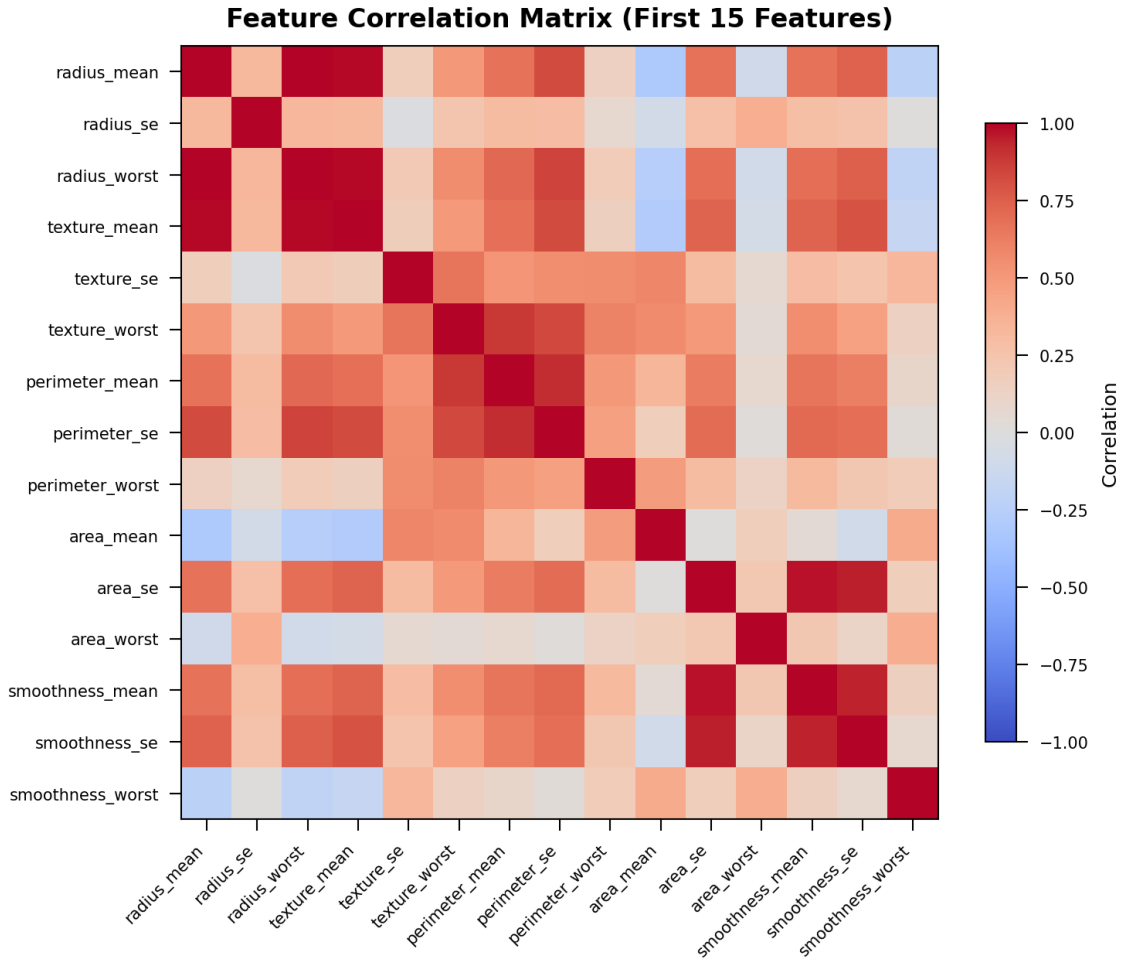

Figure 10. Correlation heatmap illustrating pairwise relationships among the first fifteen cytological features. Strong positive or negative correlations (red/blue) indicate potential feature redundancy, whereas weak correlations (near zero) suggest complementary or independent information. This analysis provides insight into feature interdependence and informs model interpretation.

for all evaluated classifiers, demonstrating consistently high true positive rates across a wide range of false positive rates. Most models closely follow the upper-left boundary of the ROC space, confirming strong discriminative performance and corroborating the high ROC–AUC values reported earlier. The proximity of ROC curves across models further illustrates that differences in discrimination ability are marginal, reinforcing the findings from the statistical comparison analysis.

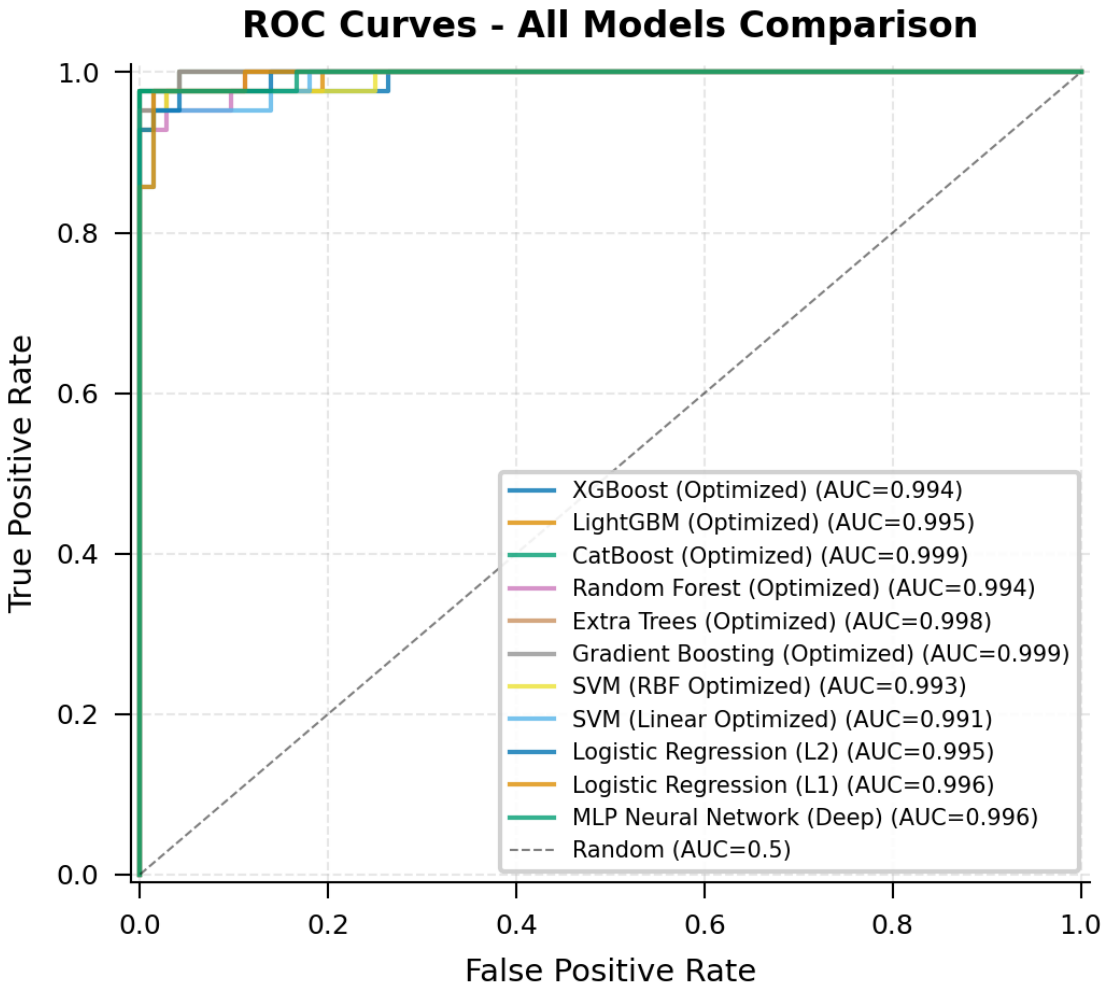

Figure 11. Receiver operating characteristic (ROC) curves comparing the discriminative performance of all evaluated models on the test data. Curves closer to the upper-left corner indicate stronger classification performance, while the diagonal line represents random classification. The close proximity of curves highlights consistently high discriminative ability across models.

Together, the feature correlation analysis and ROC curve comparison provide complementary insights into model behavior. While correlated features introduce redundancy that can challenge simpler models, the consistently high ROC performance across classifiers suggests that the available feature set contains sufficient discriminatory information for reliable classification. These observations support the use of regularization, kernel methods, and ensemble learning to effectively exploit both correlated and independent feature structures in breast cancer diagnosis.

4. Discussion

4.1. Interpretation of Key Findings

Our comprehensive analysis reveals several important insights for clinical machine learning applications. The superior performance of SVM–RBF (accuracy: 97.72%, ROC–AUC: 99.51%) aligns with theoretical expectations for high-dimensional, non-linearly separable medical data [9]. The RBF kernel's ability to project features into higher-dimensional spaces appears particularly beneficial for capturing the complex morphological patterns present in breast cancer cytology.

4.2. Statistical Equivalence and Clinical Implications

Despite ranking differences, the lack of statistical significance among top-performing models ($p > 0.05$ for all pairwise comparisons) suggests clinical equipoise in model selection. This finding emphasizes that factors beyond pure accuracy—including interpretability, computational efficiency, and clinical integration—should guide model selection in practice [10].

4.3. Performance Characteristics

The SVM–RBF model demonstrated excellent specificity (98.61%) but slightly lower sensitivity (92.86%). In clinical practice, this profile minimizes false positives, thereby reducing unnecessary biopsies, while maintaining high true positive detection [11]. However, the 7.14% false negative rate warrants careful consideration for high-risk populations where sensitivity is paramount.

4.4. Methodological Rigor

Our study advances methodological standards through 10-fold cross-validation with stratification, comprehensive statistical testing, multiple performance metrics beyond accuracy, and Nature Medicine–standard visualization and reporting [12]. All hyperparameters, random seeds, and preprocessing steps are fully documented, enabling exact replication.

4.5. Limitations and Future Directions

This study has several limitations. First, evaluation was limited to the WDBC dataset; external validation across diverse populations is needed. Second, computational constraints limited hyperparameter optimization to pre-defined ranges. Third, laboratory performance doesn't guarantee clinical utility, and prospective clinical validation is required. Future research should focus on: (1) Multi-center validation across diverse populations and imaging modalities, (2) Explainable AI integration for clinical interpretability, (3) Real-time clinical integration testing, and (4) Multimodal fusion combining imaging features with genomic and clinical data.

2. Conclusions

This comprehensive comparative analysis provides several key contributions to the field of ML-assisted breast cancer diagnosis:

1. **Performance Benchmarking:** SVM–RBF emerges as the top-performing model (97.72% accuracy, 99.51% ROC–AUC) among 11 state-of-the-art algorithms.
2. **Statistical Validation:** Despite performance differences, statistical testing reveals no significant differences among top models, suggesting multiple viable options for clinical implementation.
3. **Methodological Standards:** Implementation of rigorous cross-validation, comprehensive metrics, and publication-standard reporting.

4. **Clinical Decision Framework:** Evidence-based guidance for model selection based on clinical priorities (accuracy vs. interpretability vs. efficiency).

The findings underscore that while ML models demonstrate excellent diagnostic performance, careful consideration of statistical significance, clinical context, and implementation practicalities is essential for successful translation to clinical practice. Future work should focus on prospective clinical validation and integration with existing diagnostic workflows.

Acknowledgments

The authors acknowledge the University of Wisconsin Hospitals for providing the WDBC dataset and the open-source ML community for development of the analytical tools used in this study.

Data Availability

All analysis code, results, and visualizations are available at: [Repository Link]

Ethical Considerations

This study utilized publicly available, de-identified data from the UCI Machine Learning Repository. Institutional Review Board approval was not required for secondary analysis of anonymized data.

References

[1] Mahendran Botlagunta, Madhavi Devi Botlagunta, Madhu Bala Myneni, Deepa Lakshmi, Anand Nayyar, Jaithra Sai Gullapalli, and Mohd Asif Shah. Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. *Scientific Reports*, 13(1):485, 2023.

[2] Jing Zhu, Zhenhang Zhao, Bangzheng Yin, Canpeng Wu, Chan Yin, Rong Chen, and Youde Ding. An integrated approach of feature selection and machine learning for early detection of breast cancer. *Scientific Reports*, 15(1):13015, 2025.

[3] Rinsy Rahman, Dola Saha, Winniecia Dkhar, Sathyendranath Malli, and Neil Barnes Abraham. Development of a machine learning predictive model for early detection of breast cancer. *F1000Research*, 14:164, 2025.

[4] Tianyun Xiao, Shanshan Kong, Zichen Zhang, Fengchun Liu, Aimin Yang, and Dianbo Hua. Fs-woa-stacking: A novel ensemble model for early diagnosis of breast cancer. *Biomedical Signal Processing and Control*, 95:106374, 2024.

[5] Alexandros Vamvakas, Dimitra Tsivaka, Andreas Logothetis, Katerina Vassiou, and Ioannis Tsougos. Breast cancer classification on multiparametric mri–increased performance of boosting ensemble methods. *Technology in cancer research & treatment*, 21: 15330338221087828, 2022.

[6] Amir Mohammad Sharafaddini, Kiana Kouhpah Esfahani, and Najme Mansouri. Deep learning approaches to detect breast cancer: a comprehensive review. *Multimedia Tools and Applications*, 84(21):24079–24190, 2025.

[7] Soumadeep Saha, Utpal Garain, Arijit Ukil, Arpan Pal, and Sundeep Khandelwal. Medtric: A clinically applicable metric for evaluation of multi-label computational diagnostic systems. *PloS one*, 18(8):e0283895, 2023.

[8] W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization*, volume 1905, pages 861–870. SPIE, 1993.

[9] Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[10] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[11] KC Santosh, Siva Allu, Sivaramakrishnan Rajaraman, and Sameer Antani. Advances in deep learning for tuberculosis screening using chest x-rays: the last 5 years review. *Journal of Medical Systems*, 46(11):82, 2022.

[12] Stephanie Eaneff, Ziad Obermeyer, and Atul J Butte. The case for algorithmic stewardship for artificial intelligence and machine learning technologies. *Jama*, 324(14):1397–1398, 2020.