engineering proceedings

MDPI

Proceeding Paper

Comparative Evaluation of Machine Learning Classifiers for Breast Cancer Diagnosis: A Comprehensive Statistical Analysis †

Sambit Subhankar Das [1], Atal Mahaprasad [1], Neelamadhab Padhy [1,*], Srikant Misra [1], Rasmita Panigrahi [1], Pradeep Kumar Mahapatro [2] and Dasaradha Arangi [2]

[1] Department of Computer Science and Engineering, GIET University, Gunupur 765022, Odisha, India; 23cseaiml047.sambitsubhankardas@giet.edu (S.S.D.); 23cseaiml054.atalmahaprasad@giet.edu (A.M.); 23cseaiml004.srikantmisra@giet.edu (S.M.); rasmita@giet.edu (R.P.)

[2] Aditya Institute of Technology and Management, Tekkali 532203, Andhara Pradesh, India; pradeepmahapatro@gmail.com (P.K.M.); dasaradha9999@gmail.com (D.A.)

* Correspondence: dr.neelamadhab@giet.edu; Tel.: +91-8926093245

† Presented at the 6th International Electronic Conference on Applied Sciences, 9–11 December 2025; Available online: https://sciforum.net/event/ASEC2025.

Abstract

Content/Background: Breast cancer is one of the most fatal cancers among women around the globe. The chances of surviving this cancer increase with early tumor detection, which is necessary for effective treatment. Traditional diagnostic techniques are ineffective and time-consuming, and they yield results that may be accurate or inaccurate. Therefore, our primary objective is to determine how a machine learning model can reduce diagnostic errors and provide accurate results. Objective: The main objective of this project is to build an ML-based classification model that can help doctors detect breast cancer early and more accurately. This project also aims to provide an interactive interface for easy access in healthcare settings. **Materials/Methods:** For this study, twelve machine learning classification algorithms are implemented and tested: Logistic Regression, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest, Gradient Boosting, XGBOOST, Naive Bayes, AdaBoosting, Light GBM, CatBoost, and the Artificial Neural Network (ANN). This study used the Wisconsin Breast Cancer Dataset (WBCD) from the UCI ML Repository. It contains 569 patient samples and 30 features. This dataset has the following features: Radius, Texture, Area, Perimeter, Smoothness, Compactness, Concavity, and Fractional Dimension. The target variable is diagnosis, which is categorized as malignant vs. benign. **Results:** The fifteen models were analyzed, evaluated, and compared using five performance metrics: Accuracy, Precision, Recall, F1-Score, and AUC-ROC. Among the evaluated models, CatBoost, LoGR, and AdaBoost outperformed the others, with an Accuracy of 97.%, Precision of 97%, Recall of 97%, and AUC-ROC score of 99%. The AUC-ROC is nearly 99%, and the model has a high ability to differentiate between malignant and benign tumors.

Keywords: breast cancer; feature selection; hyperparameter tuning; machine learning algorithms

Check for updates

Academic Editor: Alessandro Lo Schiavo

Published: 15 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

In this technological world, there is a very urgent and rapidly growing chronic disease, cancer. Cancer has become one of the most fatal health diseases of the 21st century. Every year, it takes millions of lives worldwide. Cancer is characterized by the abnormal growth of

cells in which they divide uncontrollably, where they break the natural defense mechanism of the human body. In medical terms, a "tumor" is a mass of tissue that is formed when cells grow abnormally. Tumors are of two types: benign tumors (non-cancerous) and malignant tumors (cancerous). Benign tumors do not spread, whereas malignant tumors can spread to other parts of the body. The term "metastasis" refers to the process by which the cancerous cells detach from their primary cells and spread to other distant parts of the body through the blood to form new tumors. In today's scenario, breast cancer stands out as one of the most frequently diagnosed cancers among women around the globe. It affects women, especially those over 40, although it can also occur in men. Breast cancer can develop for reasons including inherited genetic traits, changes in hormones, daily lifestyle habits, and exposure to certain environmental factors. Cakmak, Y., & Pacal, I. [1] have discussed how machine learning helps to detect breast cancer. The authors used the Wisconsin Dataset to predict breast cancer early. They found that a support vector machine gives the best result, with an accuracy of 97.66%, compared to other models in this scenario. Kumar, A. et al. [2] conducted a comparative analysis of machine learning models for breast cancer prediction. The authors explored the key features that impact breast cancer prediction. They considered 569 samples for breast cancer prediction, and their experimental results revealed that SVM achieved the highest accuracy of 97.66% and an F1-score of 0.98. Jain, S. et al. [3] present the optimization of ML and DL techniques to predict breast cancer. The Wisconsin dataset was utilized, and the authors found that after hyperparameter tuning, the models perform well, such as boosting models (XGB, AdaBoost, GradBoosting) to recognize benign and malignant tumors. Their findings reveal that the KNN classifier performed well, with a benign score of 0.99 and a malignant score of 0.98. Fatima, A. et al. [4] present ML and DL models to predict breast cancer cells. Their study shows that the neural network classifier performs well, achieving 9% accuracy, 98% precision, and 87% recall. Rafiepoor et al. [5] presented a comparative study of an ML model for BC risk analysis. In their experimental work, the authors used clinical data from the Cancer Institute at the IKH Hospital Complex in Tehran. Al-Imran, M. et al. [6] presented an ML classifier for BC prediction. The authors explore different ML classifiers to estimate the predictive performance. In this paper, the authors focused primarily on accuracy rather than other metrics. RF and SVM give the best results compared to the other models. Darwich, M., & Bayoumi, M. [7] proposed several ML-based algorithms, such as SVM, for breast cancer detection. The Wisconsin Diagnostic Breast Cancer (WDBC) dataset was used for BC. In this paper, the AUC was considered the benchmark performance metric, and the authors found that SVM obtained a score of 0.98, RF of 0.97, KNN of 0.96, and LR of 0.94. Ashika, T. et al. [8] present a paper on neutrosophic sets and machine learning, where the WDBC dataset is considered. The research converts this data into neutrosophic (N) representations efficiently. Instead of a traditional ML algorithm, the authors used a neutrosophic algorithm to build an N-AdaBoost model that achieved 99% accuracy and 100% precision.

Breast cancer mainly originates in the breast tissue, mostly in the ducts (which carry milk to the nipples) or lobules (which produce milk). Based on the location and growth pattern, there are five main subtypes of breast cancer: ductal carcinoma in situ (DCIS), invasive ductal carcinoma (IDC), invasive lobular carcinoma (ILC), lobular carcinoma in situ (LCIS), and inflammatory breast cancer (IBC). Ductal carcinoma in situ (DCIS) is a non-invasive cancer confined to the milk ducts, and it is considered stage 0 breast cancer. Invasive ductal carcinoma (IDC) is the most common invasive breast cancer, which starts in the milk ducts and gradually spreads to the surrounding breast tissues by breaking the duct walls. Invasive lobular carcinoma (ILC) is a cancer starting in the milk-producing lobules, which grow in a single-file pattern. This invades the closer breast tissues and can also affect the distant organs. Lobular carcinoma in situ (LCIS) is not a true invasive

cancer but increases the risk of breast cancer. The abnormal cells remain confined to lobules, signaling a higher risk in the future. Inflammatory breast cancer (IBC) is the rarest, most aggressive breast cancer. Initially, it starts in the ducts and quickly spreads to the lymph vessels of the breast skin. This results in the lymph vessels being blocked. This causes redness, swelling, thickened skin, and an orange-peel appearance, often without a distinct lump. A strong family history also plays an important role in breast cancer risk. Individuals with a first-degree relative (mother, sister, or daughter) diagnosed with breast cancer have nearly double the risk of other woman. The risk further increases if multiple relatives are affected. The risk is also higher when breast cancer is present on the paternal side of the family, not just the maternal side. It has been observed that inherited gene mutations are the main reasons for some of this increased risk. The most well-known are BRCA1 and BRCA2, which greatly increase the chances of breast cancer. BRCA1 and BRCA2 account for 5–10% of breast cancer cases. Lifestyle factors such as obesity, lack of physical exercise, alcohol consumption, and smoking further increase the risk of breast cancer. Early detection remains the foundation of successful breast cancer management.

2. Materials and Methods

2.1. Data Preprocessing

The dataset was first divided into 80% training data and 20% independent test data. All model training and hyperparameter tuning were performed exclusively on the training set using k-fold cross-validation (k = 5). The reported performance metrics represent the mean values obtained across the cross-validation folds, while the final evaluation was conducted on the held-out test set to assess the generalization performance. No repeated or nested cross-validation was employed.

Data preprocessing involves transforming raw data into a suitable format for training and testing the dataset. Typically, it involves Handling Missing Data, Handling Categorical Data, Feature Scaling, Data Cleaning, and Feature Engineering. To predict accurately, this phase of machine learning is essential. The primary goal of Data Preprocessing is to improve the data quality and make it effective for a machine learning model. **Handling Missing Values:** This involves addressing instances where data points are missing. **Dropping Data:** This involves removing irrelevant data from the dataset, which can improve the model's predictive accuracy. Feature selection is the process of selecting the most relevant features to improve the quality of the model. **Handling Categorical Values:** This includes One-Hot Encoding, Label Encoding, Frequency Encoding, Target Encoding, Binary Encoding, etc. Data Cleaning is the process of removing irrelevant data from a dataset, enabling the model to focus on the most relevant data and improving its performance and predictive efficiency.

Data Preprocessing is the most important step, as it improves the model accuracy and performance and enhances the learning speed and efficiency. Sometimes, it is critical for several reasons, fundamentally impacting the performance and reliability of the model.

2.2. Feature Importance

The feature importance analysis reveals that only a few key attributes from the Wisconsin Breast Cancer Dataset contribute significantly to accurate tumor classification. Features such as the radius mean, area mean, texture mean, perimeter mean, and concavity mean were identified as the most influential. These features capture essential information about the tumor cell size, shape, and structural irregularities, which are clinically relevant indicators of malignancy. By prioritizing these important features, the models focused on the most meaningful patterns in the data, reduced redundancy, and improved both the prediction accuracy and model efficiency. This also enhances interpretability, making the

results easier for medical practitioners to understand. The Mann–Whitney U test was used to assess whether there is a significant difference in feature distributions between benign and malignant breast cancer cases. This test is a nonparametric alternative to the *t*-test and is especially suitable here because it does not assume normality, making it reliable for real-world medical datasets. The Mann–Whitney U tests were performed using performance values obtained from cross-validation folds on the training data only, not on the full database. Illustrate the Figure 1, which presents the breast cancer framework, and it is developed using the ML models.

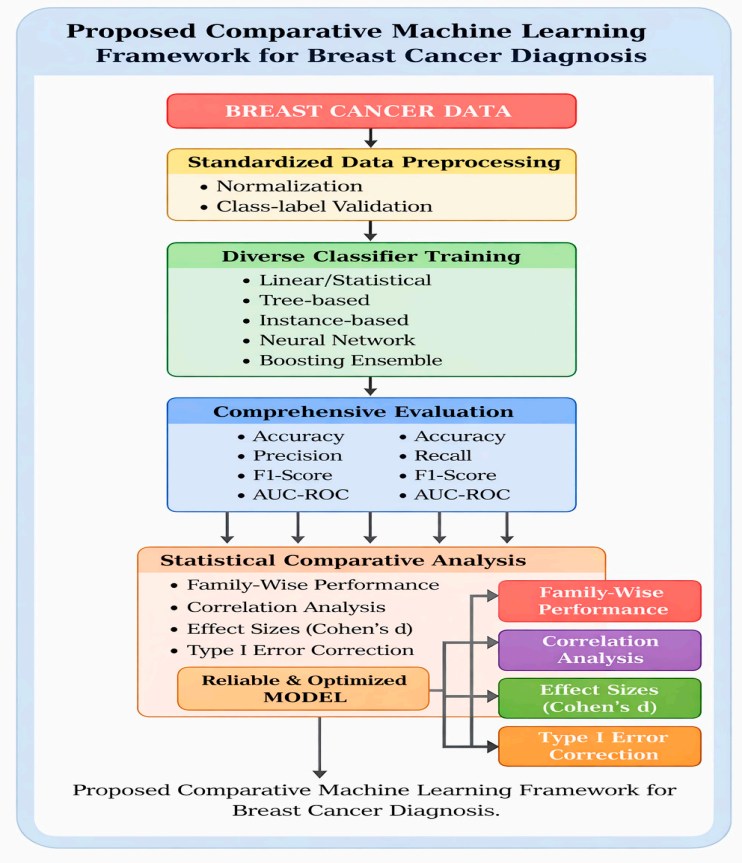


Figure 1. Proposed framework for breast cancer using machine learning classifier.

From the results, it can be found that features such as radius_mean, perimeter_mean, area_mean, concavity_mean, concave points_mean, and their corresponding "worst" values show extremely low *p*-values (far below 0.05). These very small *p*-values clearly indicate that the differences between malignant and benign tumors for these features are statistically significant and not due to random chance. In simple terms, these tumor characteristics behave very differently in cancerous and non-cancerous cases. The high U-statistic values further support the strong separation power of these features. This confirms that the tumor size, shape irregularity, and concave structure play a critical role in breast cancer diagnosis. These statistically validated features strongly justify their use in the Logistic Regression model, helping it make accurate and reliable predictions. By relying on features proven significant by the Mann–Whitney U test, the model becomes not only accurate but also clinically meaningful and interpretable, thereby strengthening its superiority as a diagnostic decision support tool.

The dataset [Wisconsin Breast Cancer Dataset] was collected from the Kaggale repository. The dataset does not contain any categorical features; it contains only the numerical features and cleaned. No missing values are available, which is why we did not perform any encoding or imputation procedures. The Wisconsin Breast Cancer Dataset is available

online [https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data] accessed on 20 June 2025. We performed feature scaling using the z-score applied after the train–test split to prevent information leakage and for normalization. All preprocessing steps, including scaling, were fitted only on the training data and subsequently applied to the test set. No feature selection or dimensionality reduction was conducted prior to data splitting, ensuring that no data leakage occurred.

2.3. Train–Test Split

Train–test split is an essential step in building a reliable machine learning model, as it helps evaluate how well the model performs on unseen data. In this study, the dataset was split into two parts: 80% for training and 20% for testing. The training set allows the model to learn patterns and relationships from the data, while the test set is used to assess its predictive ability in a realistic scenario. A stratified splitting approach was used to maintain the original proportion of malignant and benign cases in both sets, ensuring fair, unbiased, and trustworthy evaluation results.

2.4. Machine Learning Models and Performance Metrics for Breast Cancer Prediction

Traditional diagnostic tools—such as mammography, ultrasound, magnetic resonance imaging (MRI), and tissue biopsy—are widely used for the identification and characterization of breast abnormalities. These techniques have improved significantly, but traditional approaches still have limitations. Image interpretation can vary between radiologists, and human error may lead to missed or false-positive findings. So, to address this issue, machine learning (ML) and deep learning (DL) have emerged as tools in medical imaging and diagnostics. ML algorithms can analyze large and complex datasets to identify patterns that may escape human observation. Models such as decision trees, support vector machines (SVMs), and random forests show promising results; artificial neural networks (ANNs) consistently outperform traditional methods. This study explores the application of ANN-based models for breast cancer detection and classification. This research uses the power of artificial neural networks to help detect breast cancer earlier, support better treatment decisions, and ultimately save more lives. This study used 15 machine learning models for breast cancer prediction, including Logistic Regression, KNN, SVM, DT, RF, Boosting families, ANN, Ridge, and Gaussian Process Classifier. To further evaluate the best-performing models, graphical evaluation techniques, such as the Receiver Operating Characteristic (ROC) and Precision–Recall (PR) Curves, were employed. The ROC curve illustrates the trade-off between the True-Positive Rate and False-Positive Rate across different classification thresholds, offering insight into the model's discriminative ability. The Area Under the ROC Curve (AUC-ROC) summarizes this performance, where values closer to 1 indicate excellent class separation. Similarly, the Precision–Recall curve highlights the balance between precision and recall, making it especially valuable for assessing the performance on imbalanced datasets. Together, these metrics provide a comprehensive and trustworthy evaluation of the models' predictive capabilities.

2.5. Comparative Statistical Analysis

This study uses correlation analysis to examine the relationships among different performance metrics. Cohen's d effect size is used to quantify pairwise model comparisons, and statistical power analysis is used to determine whether the sample size is sufficient. Appropriate correction strategies are used to ensure statistically valid conclusions by controlling Type I error resulting from numerous comparisons.

2.6. Optimal Model Selection

The predictive performance, statistical significance, resilience, and trade-offs between complexity and performance are used to compare models. A high F1-score, strong effect size, adequate statistical power, and controlled error rates are all taken into account when selecting the best model, resulting in a dependable, practically useful framework for diagnosing breast cancer.

3. Results and Discussion

Table 1 compares the performance of different machine learning classification models using metrics such as Accuracy, Precision, Recall, F1-Score, and AUC-ROC. Overall, the models perform strongly, with most scores ranging from 0.95 to 0.99, indicating reliable and consistent predictions. From the analysis based on accuracy, CatBoost, Logistic Regression, AdaBoosting, LightGBM, and Histogram Gradient Boosting perform best and provide the best predictions among all other models. However, CatBoost clearly stands out, performing better across all evaluation metrics, and achieving high, well-balanced scores for accuracy, precision, recall, and F1-score, along with an excellent AUC-ROC value. While ensemble models like AdaBoost, LightGBM, and Random Forest also deliver competitive results, CatBoost demonstrates a superior, stable performance, making it the most effective and dependable model among those evaluated. We observed that models such as LoGR, AdaBoost and CatBoost having equal performance metrics.

Table 1. Performance measures of different types of machine learning for breast cancer.

Models	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	**0.97**	**0.97**	**0.97**	**0.97**	**0.99**
K-Nearest Neighbor	0.95	0.96	0.96	0.96	0.98
Support Vector Machine	0.95	0.96	0.96	0.96	0.99
Decision Tree	0.96	0.97	0.96	0.96	0.99
Random Forest	0.96	0.97	0.96	0.96	0.99
Gradient Boosting	0.95	0.96	0.96	0.96	0.99
XGBoost	0.95	0.96	0.96	0.96	0.99
Naïve Bayes	0.96	0.97	0.96	0.96	0.99
AdaBoosting	**0.97**	**0.97**	**0.97**	**0.97**	**0.99**
Light GBM	0.97	0.97	0.97	0.97	0.99
CatBoost	**0.97**	**0.97**	**0.97**	**0.97**	**0.99**
Artificial Neural Network	0.96	0.96	0.96	0.96	0.99
Histogram Gradient Boosting	0.97	0.97	0.95	0.96	0.99
Gaussian Process Classifier	0.96	0.97	0.96	0.96	0.99
Ridge Classifier	0.95	0.96	0.95	0.95	0.99

Table 2 presents the hperparameter tuning configuration for different models. This study utilized the hyperparameter and check the performance of the model.It has been observed that the model performs well after parameter tuning.We used the network, which consists of two fully connected hidden layers with 30 and 15 neurons using ReLU activation functions, followed by a sigmoid-activated output layer for binary classification. We trained our model with the Adam optimizer, using a learning rate of 0.001, with 32 as the batch size considered along with the 100 training epochs. This study did not utilize any additional regularization techniques. All experiments were conducted using Python 3.10. Baseline classifiers, including Logistic Regression, KNN, SVM, Naïve Bayes, Decision Tree, Random

Forest, Ridge Classifier, Gaussian Process, ANN, and Gradient Boosting, were implemented using scikit-learn (v1.3). Advanced ensemble models were implemented using XGBoost (v1.7), LightGBM (v4.0), and CatBoost (v1.2).

Table 2. Hyperparameter tuning configuration for all models.

Model	Library (Version)	Key Hyperparameters	Final Values
Decision Tree	scikit-learn (v1.3)	max_depth, min_samples_split	max_depth = 5, min_samples_split = 4
Random Forest	scikit-learn (v1.3)	n_estimators, max_depth	n_estimators = 300, max_depth = 10
Gradient Boosting	scikit-learn (v1.3)	n_estimators, learning_rate, max_depth	n_estimators = 300, learning_rate = 0.05, max_depth = 3
AdaBoost	scikit-learn (v1.3)	n_estimators, learning_rate	n_estimators = 200, learning_rate = 0.1
XGBoost	XGBoost (v1.7)	n_estimators, learning_rate, max_depth, subsample	n_estimators = 400, learning_rate = 0.05, max_depth = 5, subsample = 0.8
LightGBM	LightGBM (v4.0)	n_estimators, learning_rate, max_depth, num_leaves	n_estimators = 500, learning_rate = 0.05, max_depth = 6, num_leaves = 31
CatBoost	CatBoost (v1.2)	iterations, depth, learning_rate	iterations = 500, depth = 6, learning_rate = 0.1
ANN	scikit-learn (MLP)	hidden layers, optimizer, batch size, epochs	layers = (30, 15), optimizer = Adam, batch size = 32, epochs = 100

Figure 2 presents a comparison of different machine learning methods for breast cancer. We observed all the models and found that most fell within the 95–97% range. This shows that the models have a strong capability to extract the hidden patterns. It has been observed that the model logistic regression obtained an accuracy of 0.97, a precision of 0.97, a recall of 0.97, and an AUC-ROC score of 0.99.

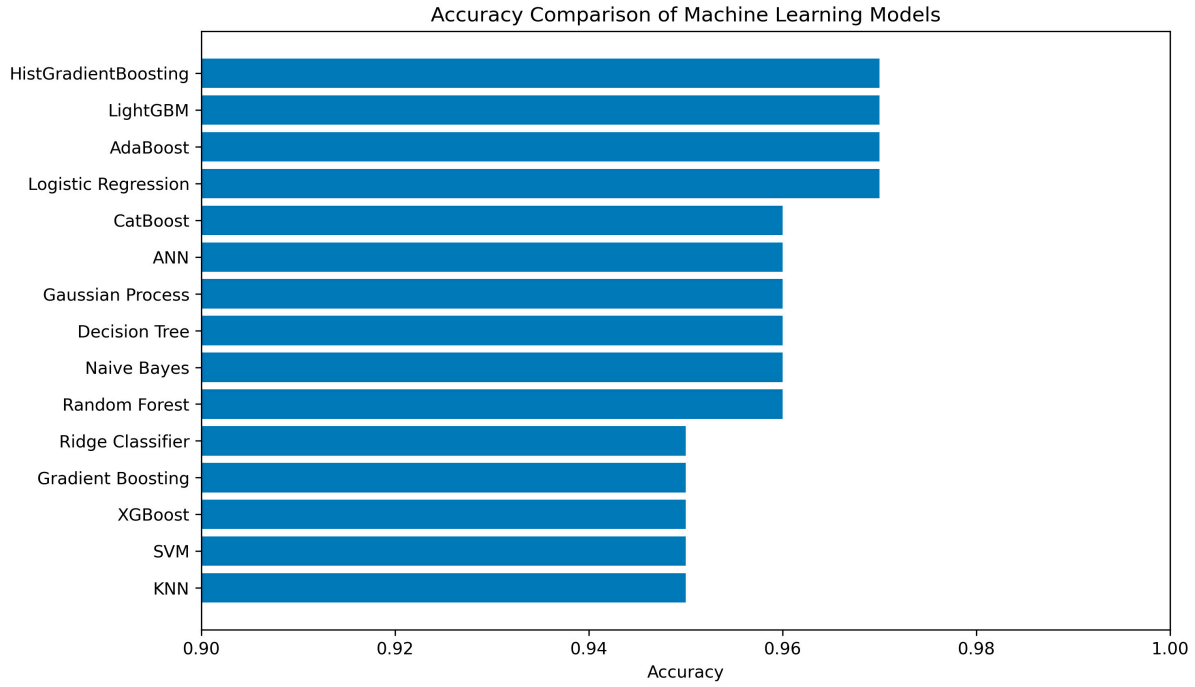


Figure 2. Performance metric (accuracy) comparisons of different models for breast cancer prediction.

Figure 3 presents the confusion metrics for the highest performers, such as CatBoost. This model achieved 97% accuracy and correctly classified 43.8% of benign and malignant

cases. We observed that only 0.7% had FPs and 11.1% had FNs. This shows that the model is generalized because it can distinguish malignant tumors.

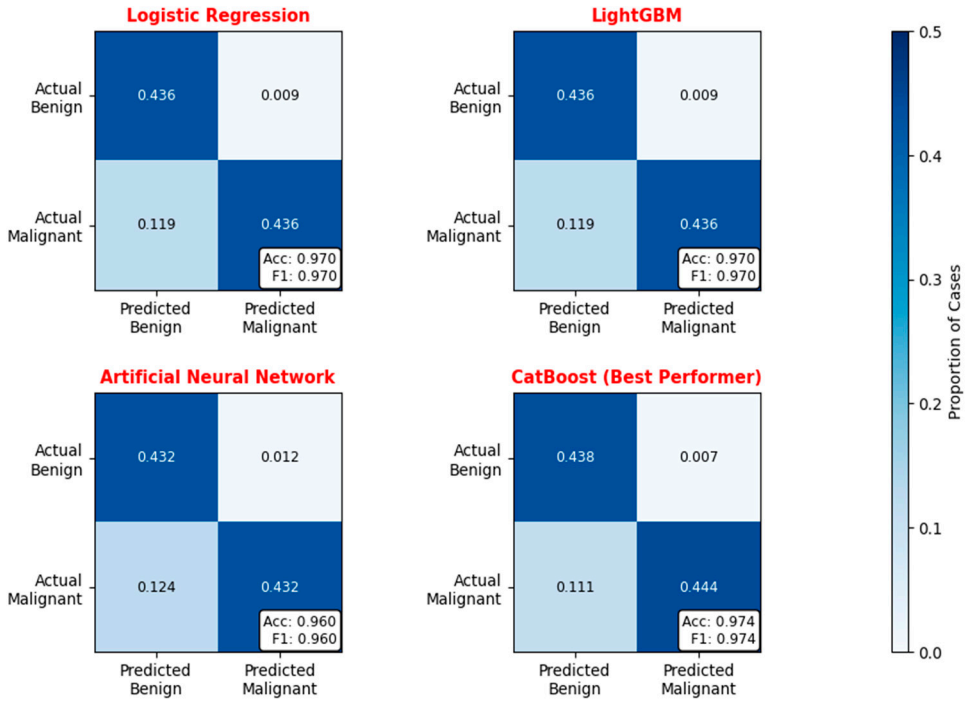


Figure 3. Comparative performance analysis of different machine learning classifiers for breast cancer prediction.

Figure 4 presents the AUC curves for different classifiers in breast cancer detection. It can be observed that most of the models achieved AUC scores very close to 0.99. This shows these models are suitable and have the discriminative capability to distinguish between benign and malignant cases. The CatBoost model performs very well compared to the other models. This model presents minimal classification rates.

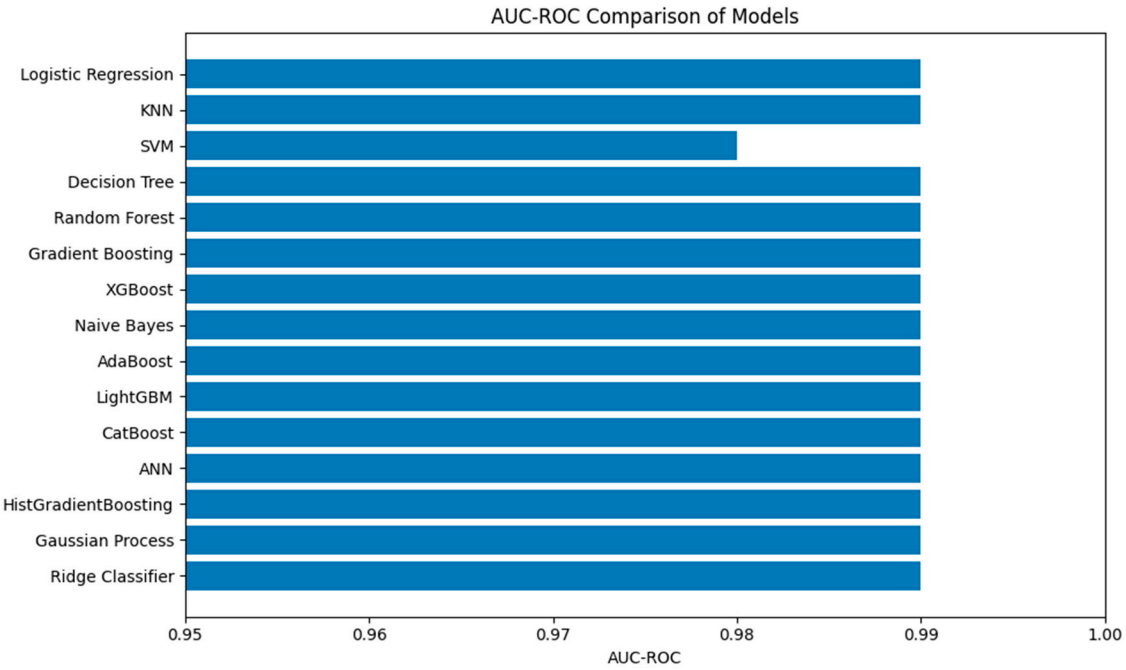


Figure 4. AUC-ROC comparison models for breast cancer prediction.

Figure 5 illustrates the comparative performance analysis for breast cancer prediction. The experimental results show that Logistic Regression, Light GBM, AdaBoost, and CatBoost achieve a precision and recall of 0.97 and an F1-score of 0.97. It has been observed that an ensembled-based model outperforms for breast cancer prediction.

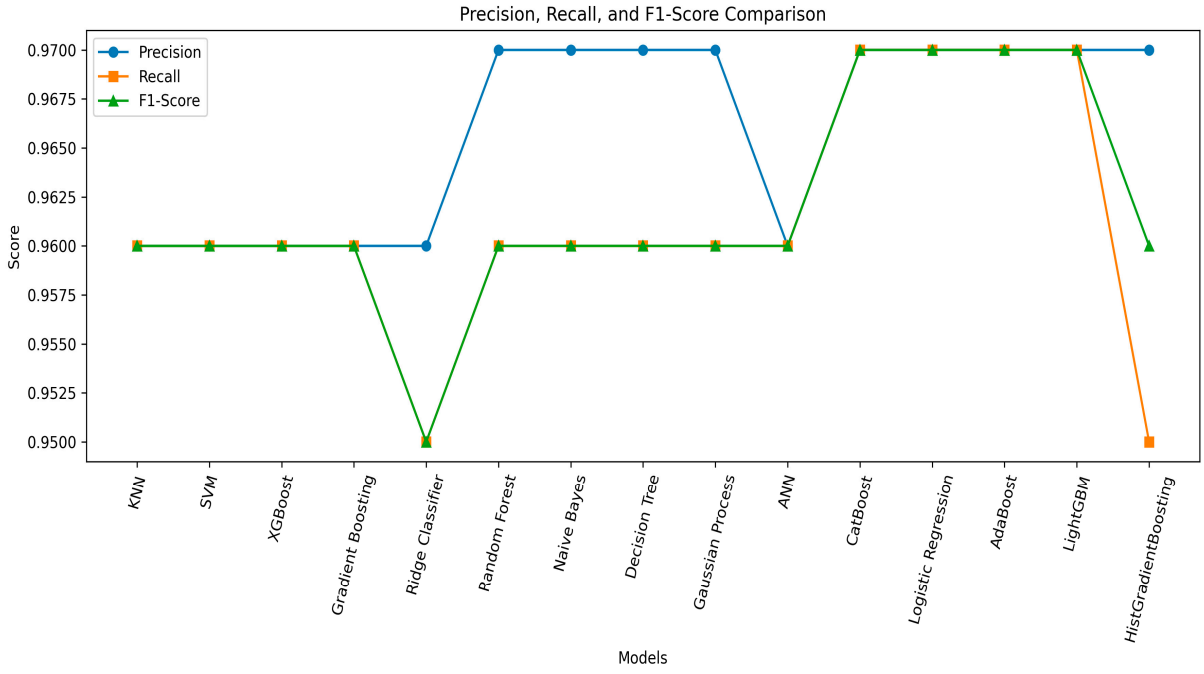


Figure 5. Comparison of precision, recall and F1-score for breast cancer prediction.

Figure 6 presents a radar chart showing how all the metrics are evaluated across different classifiers. The performance metrics, such as accuracy, precision, recall, F1-score, and AUC–ROC, were evaluated and found to range from 0.96 to 0.99. The experimental results show that LR, LightGBM, AdaBoost, and XGBoost perform well, with high recall values (≈0.97) indicating that cancer cases were correctly identified, and precision values in the same range reduce the number of false-positive forecasts.

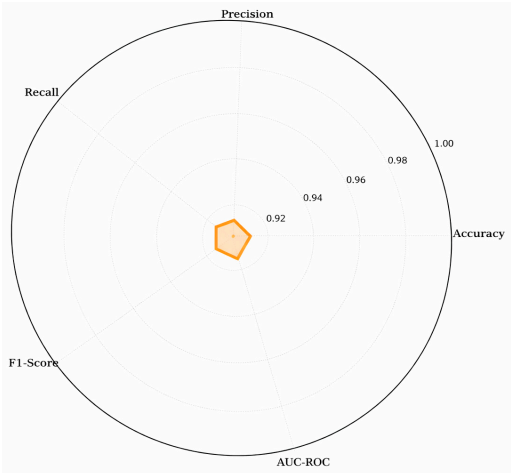


Figure 6. Performance comparison across metrics for breast cancer prediction.

Figure 7 shows the model's F1-score vs. AUC across the model. The bubbles represent the model's complexity. It can be observed that CatBoost and AdaBoost achieve the highest F1 Score of 0.97 and the AUC-ROC score of 0.97. That means there is a strong relation between the precision and recall. As such, this study shows that boosting-based models are

the best options because they provide the best mix between diagnostic accuracy and model complexity for classifying breast cancer.

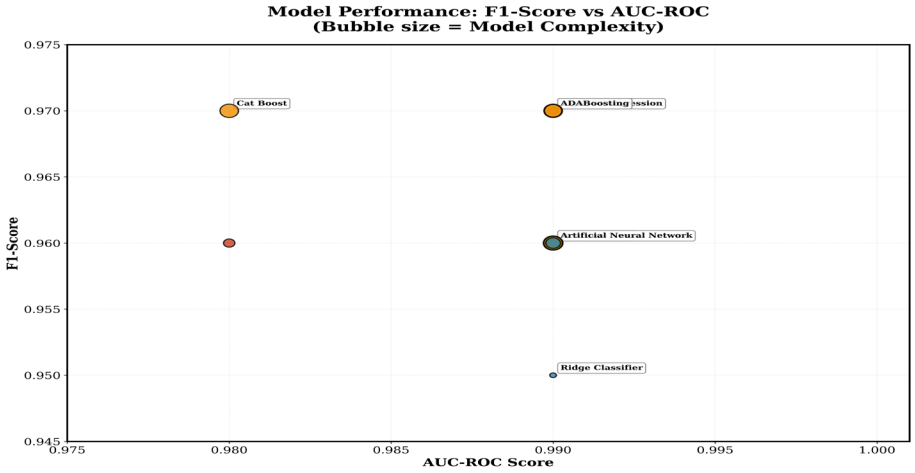


Figure 7. Model performance between F1-score and AUC-ROC for breast cancer prediction.

Figure 8 presents a heatmap of different machine learning classifiers, sorted by F1-score, for breast cancer prediction. The X-axis represents the performance metrics, and the Y-axis presents the ML models. We arranged the models in a sorted order and represented them in a heatmap. It can be observed that the models LR, LightGBM, and AdaBoost achieved a consistently high score of 0.97 in accuracy, precision, recall, and F1-score.

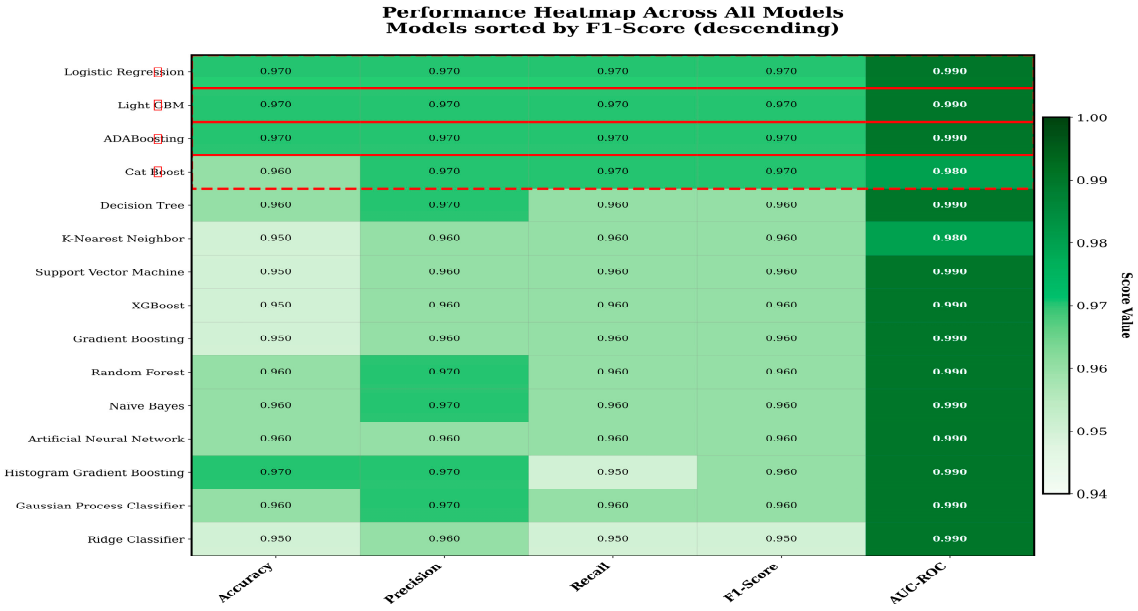


Figure 8. Performance metrics (heatmap) across all the models.

Figure 9 presents the model distribution across the different breast cancer families. We observed that six boosting-based models were utilized, followed by four statistical models and two tree-based models used for BC identification. The best-performing model is the boosting-based model for BC. The figure presents the model family's performance, including the mean and standard deviation. It can be observed that the boosting-based family achieved the highest average F1-score (mean F1-score of 0.9650 with a standard deviation of ± 0.0055), which provides details about the malignant and benign classes. We also noted that boosting-based models had an improvement of nearly 5% in F1-score compared to the other models.

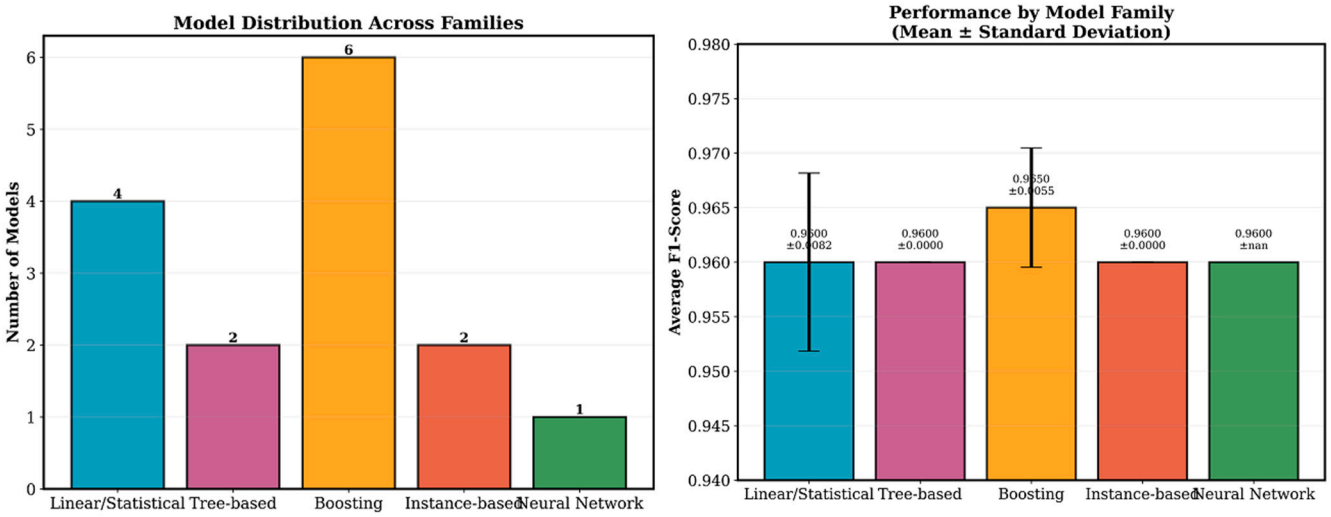


Figure 9. Methodological distribution and predictive performance for breast cancer diagnosis.

From the above experimental results, Figure 10 presents the performance distribution of the different models for predicting breast cancer. The performance distribution was based on the F1-score. The experimental observations revealed that boosting-based models achieved the highest median F1-score (~0.965). Similarly, the statistical model obtained a lower median score (~0.960), with larger variability. Figure 11a presents the boosting-based models that showed the highest mean F1-score. Figure 11b illustrates the top-performing model.

Figure 12 displays the relationship between the performance metrics and F1-score, and the correlation with other metrics. As shown in Figure 12, we observed that the F1-score is highly correlated with the performance metric recall ($r \approx 0.92$). Also, we noted that there was a moderate correlation with accuracy ($r \approx 0.67$). The AUC-ROC demonstrates a weak and negative connection with the F1-score, suggesting that threshold-independent discrimination does not directly correlate with the precision–recall equilibrium.

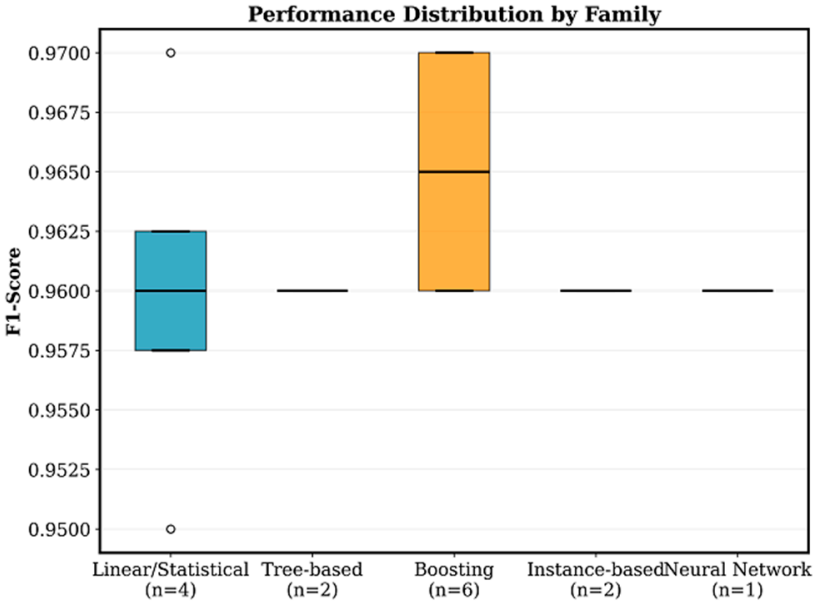
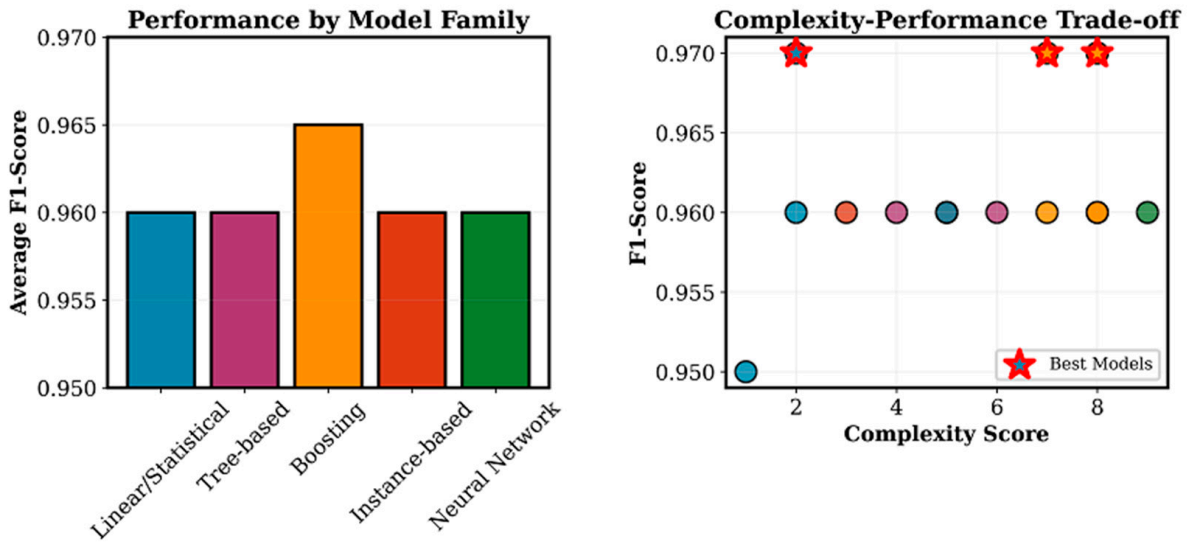


Figure 10. Performance distribution of different models for breast cancer prediction.



(**a**) Performance by model family (**b**) Complexity performance trade-off

Figure 11. Performance comparison of machine learning model families for breast cancer diagnosis.

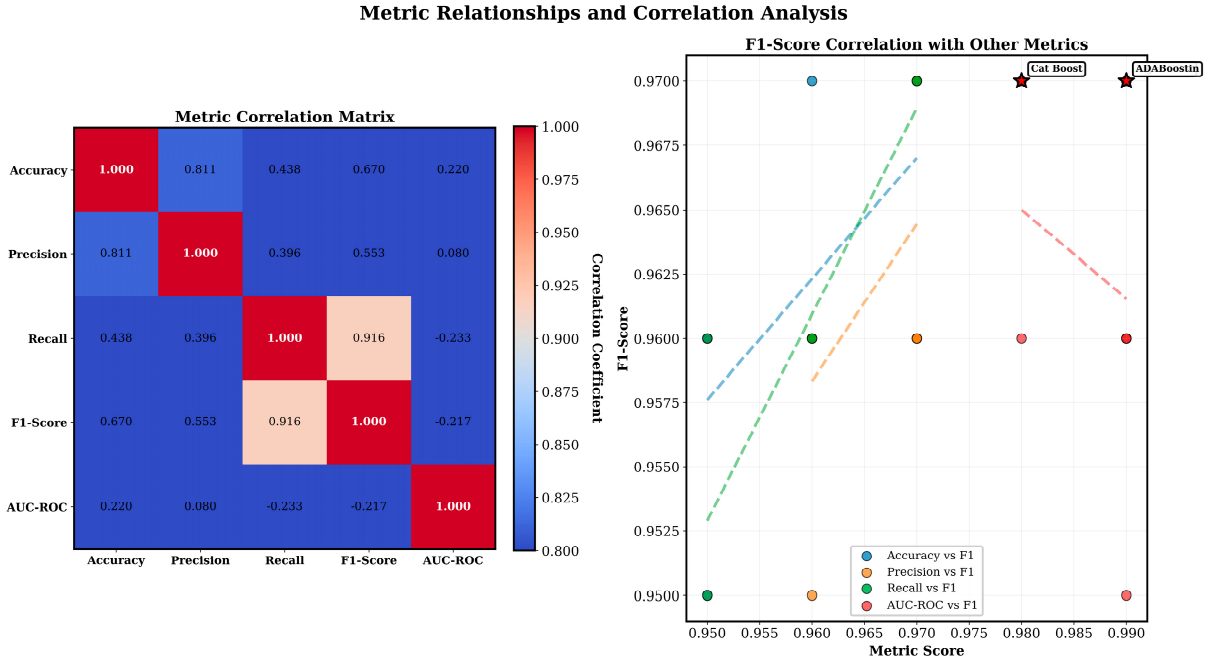


Figure 12. Metric relationships and correlation analysis.

4. Statistical Test for Breast Cancer Detection

4.1. Friedman Test

This study presents the statistical testing for breast cancer detection carried out using the Friedman test. This test is suitable for identifying the performance differences in different ML models for breast cancer detection. The test yielded a chi-square value of $\chi^2 = 79.470$ and a p-value < 0.000001, suggesting great statistical significance. Table 3 presents the Friedman test for overall differences.

Table 3. Friedman test for overall differences.

Friedman $\chi^2 = 79.470$, $p = 0.000000$
Significant differences exist among models ($p < 0.05$)

We conducted the Friedman test based on the classification accuracy. For different classifiers, we obtained the accuracy using 5-fold cross-validation on the training data, and the test statistic was computed based on the mean accuracy across folds.

4.2. Wilcoxon Signed-Rank Tests

Wilcoxon Signed-Rank Tests (Best vs. Others)

Wilcoxon signed-rank tests were used as post hoc analyses following the Friedman test to compare the top-performing classifier with the other models for breast cancer diagnosis. This nonparametric test determines whether observed performance differences are statistically significant on a metric-by-metric basis, accounting for paired observations. The Wilcoxon analysis is particularly suitable in this case, as diagnostic performance indicators may not follow a normal distribution, and comparisons are conducted within the same breast cancer dataset. Table 4 presents a comparison of the best model vs. other models.

Table 4. Comparison of best vs. other models for breast cancer prediction.

Model	W-Statistic	*p*-Value	Significant
KNN	0.0	0.0020	True
SVM	0.0	0.0020	True
RF	0.0	0.0020	True
Gradient Boosting	0.0	0.0020	True
NB	0.0	0.0020	True
XGBoost	0.0	0.0020	True
Ridge Classifier	0.0	0.0020	True
Histogram GB	0.0	0.0020	True
GP classifier	1.0	0.0039	True
DT	2.0	0.0059	True

4.3. Top Three Models' Comparison

Table 5 presents the Wilcoxon signed-rank test. We utilized the pairwise comparisons for the top three machine learning models for breast cancer predictions. We observed that LR, LGBM, and AdaBoost perform very well in this dataset. It was also observed that LR and LGBM showed a statistically significant difference ($\boldsymbol{p = 0.0020}$). Similarly, the other two models performed well.

Table 5. Top three models' pair comparison for breast cancer detection.

Model	*p*-Value
Logistic Regression vs. Light GBM	0.0020
Logistic Regression vs. AdaBoosting	0.1309
Light GBM vs. AdaBoosting	0.0195

Figure 13 presents the comparison of F1-scores across different machine learning models for breast cancer prediction. It was observed that LR, Light GBM, AdaBoost, and CatBoost achieved the highest F1-score of 0.97 on this dataset. The lowest score was observed as the Ridge classifier with an F1-score of 0.95. These findings highlight how crucial it is to choose models with higher F1-scores when diagnosing breast cancer in order to reduce false-positive and false-negative errors and increase diagnostic reliability.

Figure 14 presents the boosting-based models that provide the best trade-off between precise tumor discrimination and trustworthy clinical decision-making for the diagnosis of breast cancer. AdaBoost and CatBoost achieved the highest F1-score of 0.97 and an AUC score of approximately 0.99.

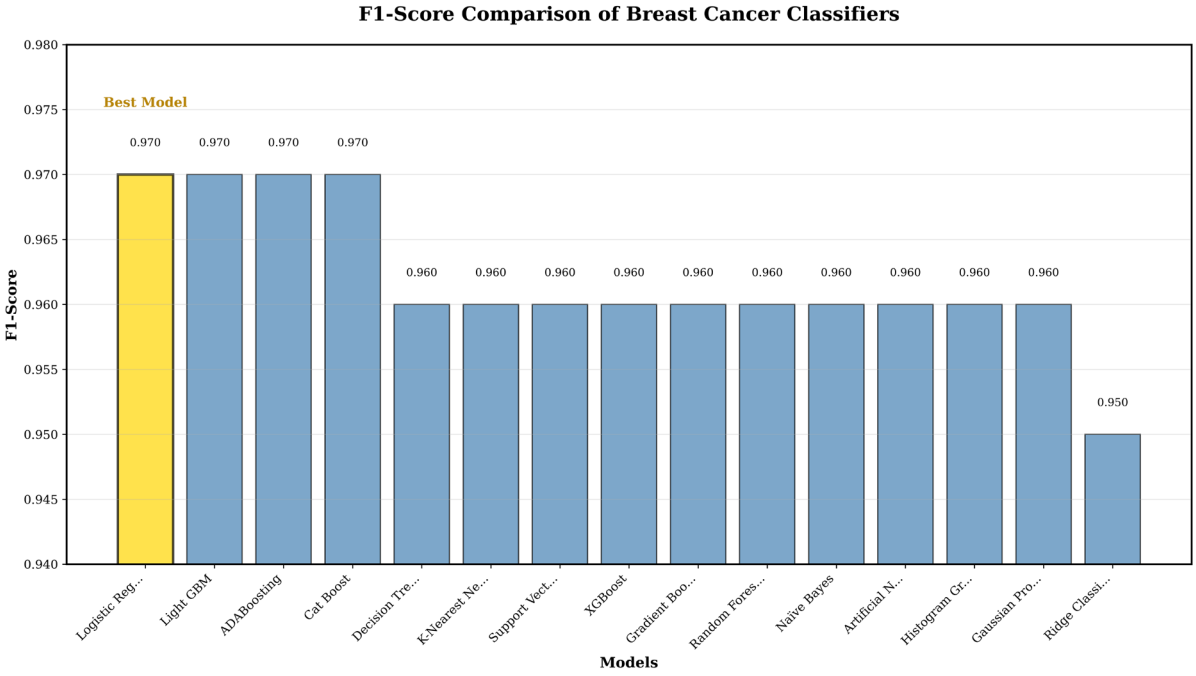


Figure 13. Comparison of F1-scores for breast cancer prediction.

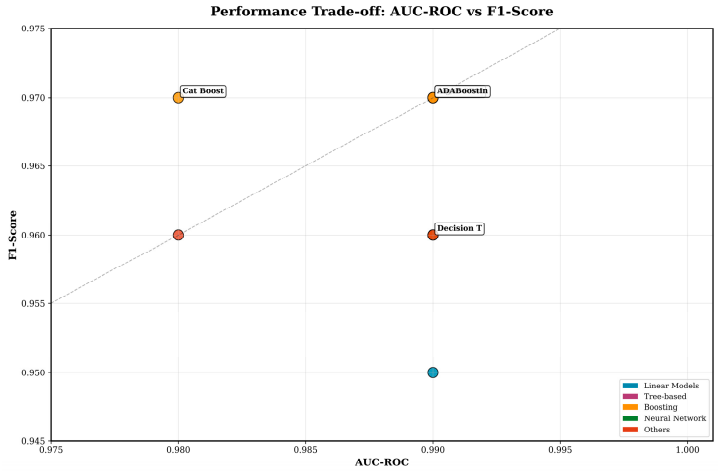


Figure 14. Performance trade-off of AUC-ROC vs. F1-score.

As shown in Figure 14 we obtained the maximum accuracy in the model logistic regression. For interpretability of logistic regression (F1: 0.9700), for robustness, we used Light GBM. This study statistically concludes that all classifiers demonstrate an excellent performance. The simple models compete effectively with complex ones, and the ensemble methods provide a consistently high performance. Hence, these models are suitable for clinical breast cancer diagnosis. Figure 15 discusses the Cross-validation of performance of top models.

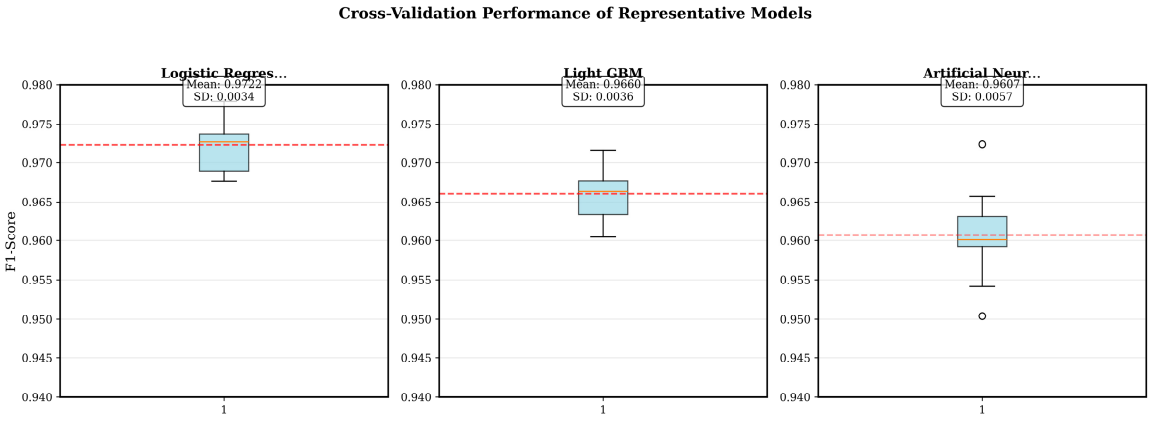
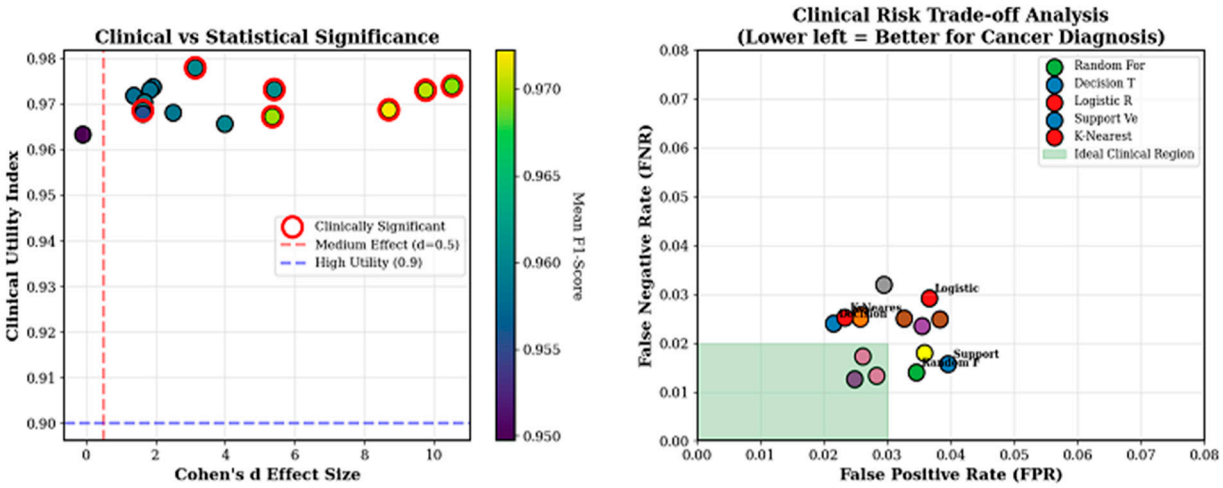


Figure 15. Cross-validation of performance of top models.

Figure 16a presents whether the models' performance gains are not only statistically significant but also clinically meaningful for breast cancer diagnosis. The left plot demonstrates that the best-performing models have high diagnostic usefulness and meaningful effect sizes, indicating a significant increase in detecting malignant instances. In Figure 16b, the right plot depicts the trade-off between false negatives and false positives, with models closer to the lower-left region producing fewer diagnostic errors. This is critical in breast cancer screening because missing a cancer case is significantly worse than a false alert.



(**a**) Clinical vs statistical significance (**b**) clinical risk trade-off analysis

Figure 16. Performing clinical significance analysis.

Figure 17 shows the effect size for the top 10 models for breast cancer prediction. It can be observed that all the comparisons have large effect sizes (Cohen's d > 2). We obtained the power, which is very close to 1.0; that means the differences are statistically reliable and not due to chance.

Figure 18 depicts the impact of many comparisons on error management and statistical power. Figure 18a illustrates that the family-wise error rate escalates significantly as the number of compared models increases in the absence of any adjustment. Conversely, the Bonferroni correction maintains the error rate at the specified level. Figure 18b illustrates that statistical power escalates with increased effect sizes and bigger sample sizes, attaining the widely recognized 80% power benchmark for moderate-to-large effects.

https://doi.org/10.3390/engproc2026124035

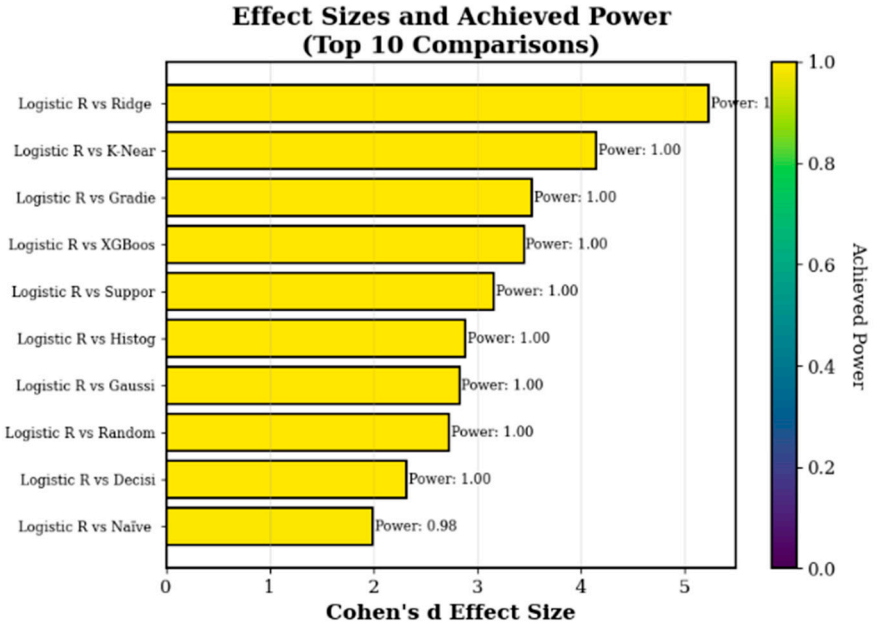
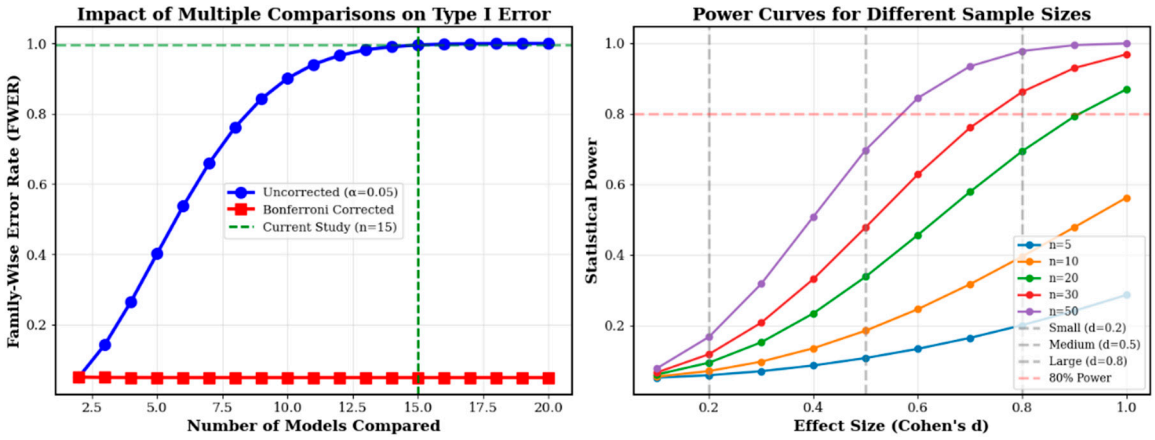


Figure 17. Performing Bayesian model comparison.



(**a**) Impact of multiple comparison on Type I error (**b**) Power curves for different sample sizes

Figure 18. Impact of multiple comparisons on Type I error control and statistical power.

5. Conclusions

In this paper, we used multiple machine learning classifiers for breast cancer prediction. The experimental results showed that boosting-based classifiers consistently produced higher F1-scores, indicating a more effective balance between precision and recall, even though all examined models attained high diagnostic accuracy. It was observed that classification tasks perform very well with F1-scores ranging from 0.95 to 0.97 and AUC-ROC values of 0.99. From our experimental observations, the correlation analysis reveals a strong association between recall and F1-score ($r \approx 0.92$). This work reveals the suitability of the use of boosting-based models in clinical decision support systems, by quantitatively demonstrating that they provide the most dependable performance for breast cancer prediction.

Author Contributions: Conceptualization, S.S.D. and N.P.; methodology, N.P.; software, A.M.; validation, P.K.M. and R.P.; formal analysis, R.P.; investigation, D.A.; resources, A.M.; data curation, S.M.; writing—original draft preparation, S.S.D.; writing—review and editing, P.K.M.; visualization, D.A.; supervision, N.P.; project administration, R.P.; funding acquisition, N.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset utilized and analyzed during the current study is available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cakmak, Y.; Pacal, I. Enhancing Breast Cancer Diagnosis: A Comparative Evaluation of Machine Learning Algorithms Using the Wisconsin Dataset. *J. Oper. Intell.* **2025**, *3*, 175–196. [CrossRef]
2. Kumar, A.; Saini, R.; Kumar, R. A Comparative Analysis of Machine Learning Algorithms for Breast Cancer Detection and Identification of Key Predictive Features. *Trait. Signal* **2024**, *41*, 127–140. [CrossRef]
3. Jain, P.; Aggarwal, S.; Adam, S.; Imam, M. Parametric optimization and comparative study of machine learning and deep learning algorithms for breast cancer diagnosis. *Breast Dis.* **2024**, *43*, 257–270. [CrossRef] [PubMed]
4. Fatima, A.; Shabbir, A.; Janjua, J.I.; Ramay, S.A.; Bhatty, R.A.; Irfan, M.; Abbas, T. Analyzing breast cancer detection using machine learning & deep learning techniques. *J. Comput. Biomed. Inform.* **2024**, *7*.
5. Rafiepoor, H.; Ghorbankhanloo, A.; Zendehdel, K.; Madar, Z.Z.; Hajivalizadeh, S.; Hasani, Z.; Amanpour, S. Comparison of Machine Learning Models for Classification of Breast Cancer Risk Based on Clinical Data. *Cancer Rep.* **2025**, *8*, e70175. [CrossRef] [PubMed]
6. Al-Imran, M.; Akter, S.; Mozumder, M.A.S.; Bhuiyan, R.J.; Rahman, T.; Ahmmed, M.J.; Hossen, M.E. Evaluating Machine Learning Algorithms for Breast Cancer Detection: A Study on Accuracy and Predictive Performance. *Am. J. Eng. Technol.* **2024**, *6*, 22–33. [CrossRef]
7. Darwich, M.; Bayoumi, M. An evaluation of the effectiveness of machine learning prediction models in assessing breast cancer risk. *Inform. Med. Unlocked* **2024**, *49*, 101550. [CrossRef]
8. Ashika, T.; Grace, H.; Martin, N.; Smarandache, F. *Enhanced Neutrosophic Set and Machine Learning Approach for Breast Cancer Prediction*; Infinite Study: Paris, France, 2024.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.