Technology in Cancer Research & Treatment
Volume 23: 1-28

DOI: 10.1177/15330338241234791
journals.sagepub.com/home/tct

Breast Cancer Prediction Based on Multiple Machine Learning Algorithms

Sheng Zhou[1], Chujiao Hu[2], Shanshan Wei[1], and Xiaofan Yan[2]

Abstract

Introduction: The incidence of breast cancer has steadily risen over the years owing to changes in lifestyle and environment. Presently, breast cancer is one of the primary causes of cancer-related deaths among women, making it a crucial global public health concern. Thus, the creation of an automated diagnostic system for breast cancer bears great importance in the medical community. **Objectives:** This study analyses the Wisconsin breast cancer dataset and develops a machine learning algorithm for accurately classifying breast cancer as benign or malignant. **Methods:** Our research is a retrospective study, and the main purpose is to develop a high-precision classification algorithm for benign and malignant breast cancer. To achieve this, we first preprocessed the dataset using standard techniques such as feature scaling and handling missing values. We assessed the normality of the data distribution initially, after which we opted for Spearman correlation analysis to examine the relationship between the feature subset data and the labeled data, considering the normality test results. We subsequently employed the Wilcoxon rank sum test to investigate the dissimilarities in distribution among various breast cancer feature data. We constructed the feature subset based on statistical results and trained 7 machine learning algorithms, specifically the decision tree, stochastic gradient descent algorithm, random forest algorithm, support vector machine algorithm, logistics algorithm, and AdaBoost algorithm. **Results:** The results of the evaluation indicated that the AdaBoost-Logistic algorithm achieved an accuracy of 99.12%, outperforming the other 6 algorithms and previous techniques. **Conclusion:** The constructed AdaBoost-Logistic algorithm exhibits significant precision with the Wisconsin breast cancer dataset, achieving commendable classification performance for both benign and malignant breast cancer cases.

Keywords

breast cancer, machine learning, high correlation filtering method, confusion matrix, spearman correlation coefficient, Wilcoxon rank sum test

Abbreviations

AUC, area under the curve; K-NN, K-nearest neighbor; MG, mammography; ROC, receiver operating characteristic curve; SVM, support vector machines; SGD, stochastic gradient descent; US, ultrasonography.

Received: September 7, 2023; Revised: November 13, 2023; Accepted: January 22, 2024.

Introduction

Background of the Study

On January 12, 2023, the renowned international clinical journal Cancer Journal for Clinicians published its annual report online, titled "Cancer Statistics 2023". According to this report, the 10 most frequently occurring tumors among women are breast, lung, colorectal, cervical, cutaneous melanoma, non-Hodgkin's lymphoma, thyroid tumors, pancreatic, kidney, and leukemia. Among these, breast cancer stands out as the most prevalent, making up approximately 31%.[1]

Breast cancer is the leading cause of cancer-related death among women and ranks fifth in terms of overall cancer

[1] Department of Preventive Medicine, Guizhou Medical University, Guiyang, China
[2] Department of Medicine and Health Management, Guizhou Medical University, Guiyang, China

Corresponding Author:
Xiaofan Yan, Room904, Building 5, Faculty Apartment, Guizhou Medical University, Huaxi District, Guiyang City, Guizhou Province, China.
Email: 137859829@qq.com

mortality. According to the 2020 global cancer statistics released by the International Agency for Research on Cancer of the World Health Organization, Female breast cancer has now surpassed lung cancer as the most commonly diagnosed cancer worldwide; The estimated 2.3 million new cases indicate that 1 in every 8 cancers diagnosed in 2020 is breast cancer; The disease is the fifth leading cause of cancer mortality worldwide, with 685 000 deaths in 2020. In women, breast cancer accounts for 1 in 4 cancer cases and 1 in 6 cancer deaths, and the disease ranks first in terms of incidence and mortality in most countries around the world (in 159 and 110 countries, respectively). Each year, more than 20 000 new cases of breast cancer are reported, resulting in 266.85 million deaths globally.[2] In 2016, China reported approximately 306 000 new cases of breast cancer, making it the most prevalent cancer among women in the country. As of January 2022, approximately 4.1 million women in the United States have a history of breast cancer. Among these cases, invasive cancer has been diagnosed in about 13% of women, with the highest risk of diagnosis occurring in women aged 70–79 years.[3] The peak incidence of new cases was observed in the age range of 60–79 years, with higher urban incidence rates compared to rural areas (where breast cancer ranked as the second most common cancer). Furthermore, breast cancer was responsible for 72 000 deaths.[4]

Despite the increasing incidence and mortality rates of breast cancer, its causes and pathogenesis remain poorly understood. Current evidence suggests that breast cancer may be affected by the intrauterine environment, that exposures during adolescence are particularly important, and that pregnancy has a dual effect on breast cancer risk: An early increase followed by long-term protection. Great variation exists in the structural development of the breast ductal system already in the newborn —and by inference in utero—and a pregnancy induces permanent structural changes in the mammary gland.[5] Furthermore, delayed childbearing and a decline in the number of births in countries undergoing social and economic transition may also contribute to the increased incidence of breast cancer. With the persistence of these cumulative risk factors, the likelihood of developing breast cancer rises. As a result, regular screening, accurate diagnosis, and timely and effective treatment play a crucial role in breast cancer prevention and serve as an important means of preserving women's lives and health.

Commonly used clinical screening and diagnostic methods for breast cancer include mammography (MG), ultrasonography, magnetic resonance imaging, nucleic acid hybridization system, real-time quantitative fluorescence polymerase chain reaction system, protein hybridization system, and flow cytometry.[6] Given that breast cancer must be differentiated from benign conditions such as breast adenoma, cystic hyperplasia, and plasma cell mastitis, the diagnosis of breast cancer remains susceptible to misdiagnosis and underdiagnosis due to complex clinical factors and the quality of acquired images and equipment. Hence, the application of automated techniques in breast cancer surveillance is crucial.

Research status

Currently, the primary machine learning algorithms employed in breast cancer classification prediction research and risk assessment include logistic regression, random forest, multilayer perceptron, BP neural network, XGBoost, K-nearest neighbor (K-NN), support vector machine (SVM), and others.[7] Yu-Dong Zhang et al[8] processed digital mammograms to develop a computer-aided diagnostic system. They applied weighted fractional-order Fourier transform, principal component analysis, SVM, and K-NN algorithm. Among them, the weighted fractional-order Fourier transform combined with principal component analysis and SVM resulted in a sensitivity of 92%. 22%±4.16%, 92.10% ± 2.75% specificity, and 92.16% ± 3.60% accuracy were reported in the study, demonstrating the proficiency of SVMs in diagnosing breast cancer. Wu et al[9] conducted an evaluation of SVM, K-NN, simple Bayesian, and decision tree models based on features selected from The Cancer Genome Atlas datasets. Their findings demonstrated that the SVM model outperformed the other 3 models in classifying The Cancer Genome Atlas datasets. Zhang et al[10] proposed a method based on contrast-constrained adaptive histogram equalization and chaotic adaptive real number coded biogeography optimization for breast cancer diagnosis in a study, and their experimental results showed that their proposed CAR-BBO method achieves superior performance to genetic algorithms, particle swarm algorithms and AR-BBO algorithms in terms of accuracy, specificity, precision, and sensitivity. Monirujjaman Khan et al[11] utilized the Wisconsin Breast Cancer Diagnostic dataset to train and assess random forest, logistic regression, decision tree, and K-NN algorithms. Their results indicated that the Logistic Regression model achieved 98% accuracy, surpassing the performance of random forest, decision tree, and K-NN algorithms. Kumar et al[12] developed an optimized stacked integrated learning (OSEL) model for early breast cancer prediction utilizing a dataset from the University of California Irvine repository. Their approach demonstrated higher accuracy compared to individual machine learning prediction models such as AdaBoostM1, Gradient Boosting, Stochastic Gradient Boosting, CatBoost, and XGBoost. This underscores the efficacy of an integrated learning approach for breast cancer prediction. Aamir et al[13] devised a machine learning framework employing random forest, gradient boosting, SVMs, artificial neural networks, and multilayer perceptron algorithms. They incorporated a linkage-based feature selection technique to enhance the performance of the framework. Their study emphasized the importance of effective data preprocessing and feature selection in improving model accuracy. Wang et al[14] conducted a research study on digital MG diagnosis, where they designed and trained a feedforward neural network classifier using the Jaya algorithm as a parameter-free method. The findings revealed that the Jaya algorithm was more efficient than BP, MBP, GASA, and PSO when training feedforward neural networks.

In the current body of literature, it has been well-established that a viable approach to developing breast cancer prediction models involves utilizing breast cancer data, employing appropriate feature selection techniques, and leveraging both machine learning and deep learning algorithms. Additionally, the utilization of an integrated learning approach has been discovered to improve model accuracy when compared to the use of a single machine learning algorithm.

Significance of the Study

Theoretical Significance. This paper integrates health big data, machine learning, and medicine to establish a machine-learning model for breast cancer classification. It demonstrates the feasibility and effectiveness of utilizing the Pearson correlation coefficient,[15] the grid search algorithm,[16] the AdaBoost algorithm,[17] and machine learning models in disease diagnosis within the medical field. In addition, this article proposes to use machine learning methods for predicting and diagnosing cancer to assist traditional medical and pathological diagnosis. This innovative approach not only provides a new idea and framework for the development and establishment of diagnostic technologies for other diseases but also serves as a case study in combining artificial intelligence (AI) with medicine. This integration promotes the merging and application of medicine with engineering and biology, harnessing the complementary advantages of different disciplines.

Practical Significance. The machine learning model developed in this study aims to address various challenges in traditional breast cancer diagnosis, including long patient waiting times, missed tests, and misdiagnosis. Additionally, it has the potential to decrease the workload of clinical laboratory personnel and enhance the efficiency and accuracy of breast cancer diagnosis in healthcare organizations. Moreover, the breast cancer prediction results obtained from machine learning models can aid clinicians in formulating treatment plans, thereby minimizing the time and economic burdens on patients. Finally, by analyzing the performance of the feature subset extracted through the feature selection technique in the machine learning algorithm, the statistical relationship between the feature subset and the benignity or malignancy of breast cancer is unveiled. This discovery provides valuable insights into the role of feature subsets in breast cancer pathogenesis, disease progression, and diagnostic processes, thereby guiding future research within the medical community.

Research Objective. The objectives of this study are threefold. Firstly, we aim to analyze a subset of features from the Wisconsin Breast Cancer Dataset that closely correlate with the benign and malignant characteristics of breast cancer. These feature subsets are instrumental in guiding the medical field's exploration of the causative agents, pathogenesis, disease evolution, and prognosis of breast cancer. Secondly, we seek to train a machine learning model based on these feature subsets and evaluate its classification accuracy and generalization ability to optimize accurate diagnosis for both benign and malignant breast cancer. Lastly, we assess the strengths and weaknesses of the top-performing machine learning algorithms trained on the Wisconsin Breast Cancer Dataset, providing valuable insights and directions for algorithm optimization, deployment, and as a reference solution for leveraging machine learning techniques to address medical problems. After preprocessing, analyzing, and visualizing the data, we develop and train 6 different machine learning algorithms (decision tree, stochastic gradient descent [SGD], random forest, K-NN, SVM, logistic regression, and AdaBoost logistic) in this study. The final evaluation results demonstrate that the AdaBoost-Logistic model achieves a remarkably high classification accuracy of 99.21%, effectively facilitating the classification and diagnosis of breast cancer.

Materials and Methods

In this research paper, we trained machine learning algorithms using the publicly available Wisconsin breast cancer dataset from Baidu Ai Studio.[18] The dataset underwent several data preprocessing steps, including imputation of missing values, removal of anomalous data, and transformation of data types. Subsequently, we utilized a high correlation filter based on the Pearson correlation coefficient and data visualization techniques to select a subset of 10 features that are closely associated with benign and malignant breast cancer. Normalization was applied to ensure scale consistency of the selected datasets. The processed data was then used to train various machine learning algorithms, such as decision trees, random forests, K-NNs, logistic regression, SVMs, and SGD algorithms. The classification performance of these algorithms was evaluated. The best model was extracted, and further optimization was performed using grid search and AdaBoost algorithms. As a result, we obtained the best algorithms with training and testing errors of less than 1%. The overall research framework is illustrated in Figure 1:

Data Acquisition

Data, as a fundamental component of AI, plays a crucial role in influencing subsequent feature selection and model performance. In this study, we utilized the Wisconsin breast cancer dataset, obtained from the Baidu Flying Paddle Little Love Studio platform, to construct machine learning models. The dataset consists of 569 rows of data distributed across 32 columns. Out of these, 30 columns represent physiological characteristics related to breast cancer, one column represents ID numbers, and the last column represents breast cancer categories. In our analysis, the categories are denoted by "B" for benign tumors and "M" for malignant tumors.

Data Preprocessing

After loading the data, missing values were identified and handled by filling in the gaps. The "diagnosis" attribute, which represents the breast cancer category, was encoded as "0" for benign tumors and "1" for malignant tumors for visualization purposes. Upon analyzing the data, 357 instances of

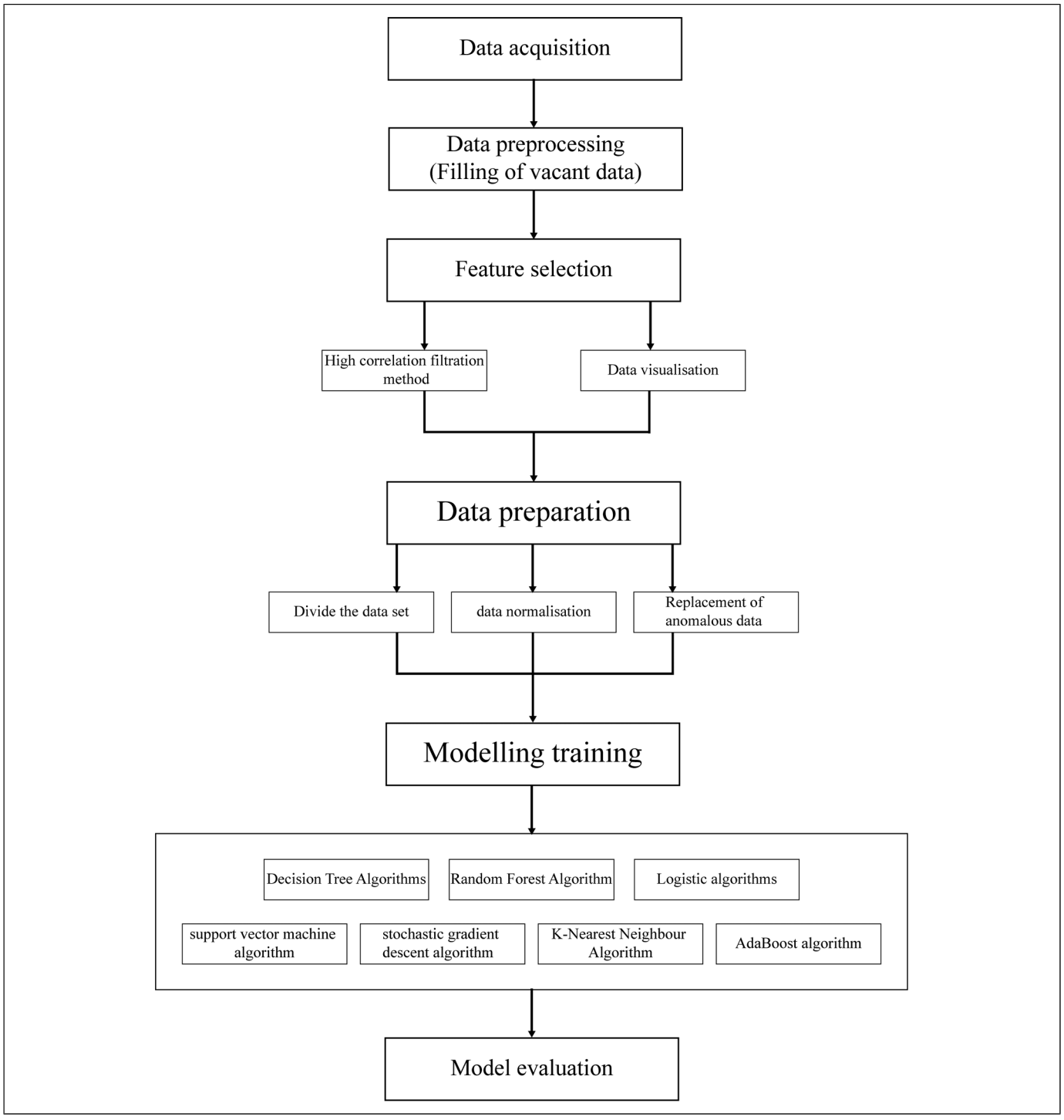

Figure 1. Research process.

benign tumors and 212 instances of malignant tumors were found within the "diagnosis" attribute. The distribution of benign and malignant breast cancers is presented in Figure 1 through a histogram and Gaussian kernel density estimate plot. Figure 2 displays the number of data points for category "M" and category "B", while Figure 3 illustrates the distributions of category "M" and category "B" for breast cancer using kernel density estimation.

After completing vacancy data filling and label coding, we analyzed the data distribution. To determine the suitable statistical method, we used the normal test function from the scipy library in the Python environment to test whether the data follows a normal distribution or not. It is crucial to assess if the data meets the normality assumption before choosing the statistical method. The table below demonstrates the normality test outcomes for each feature.

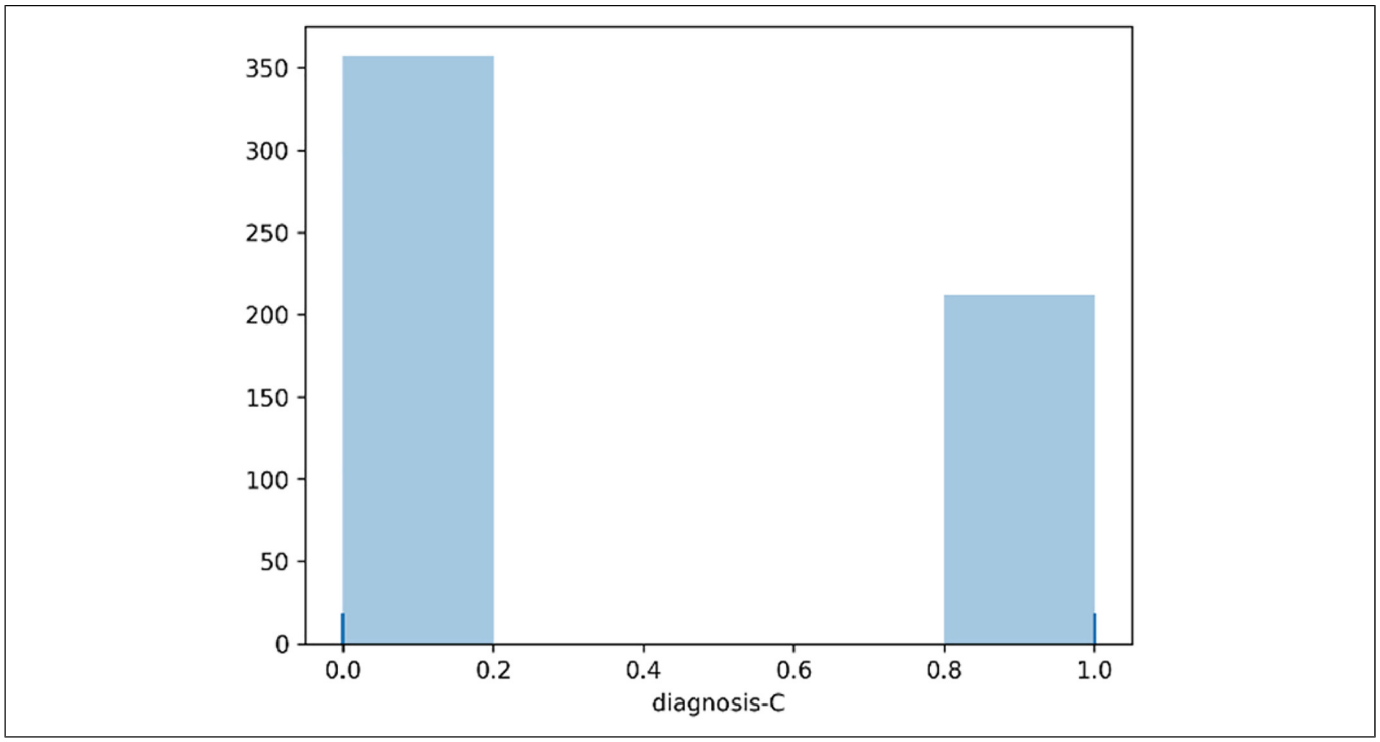

Figure 2. Histogram of diagnosisbiao categorie.

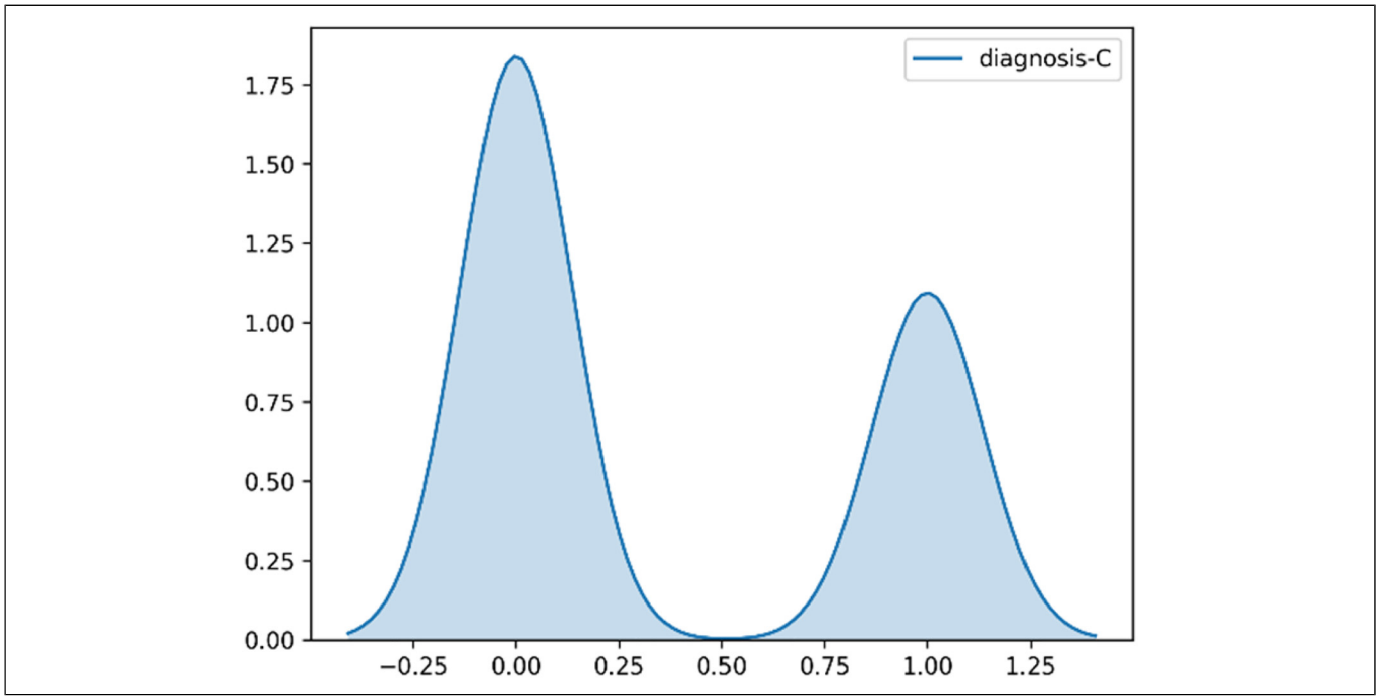

Figure 3. Kernel density estimation of diagnosisbiao categories.

After conducting a normality test, it became apparent that the feature data did not follow a normal distribution. As a result, we utilized the Spearman correlation coefficient to analyze the relationship between each attribute and the breast cancer diagnostic data. This led to the creation of a heat map, which displays the correlation coefficients and *p*-values. Table 1 presents the correlations at various Spearman correlation coefficients. Additionally, Figures 4 and 5 depict the Spearman correlation

coefficient matrix and P-value matrix for the feature data within the dataset. Any feature having correlation coefficients below 0.7 are excluded, while any exceeding this threshold are identified, such as "perimeter_worst," "radius_worst," "area_worst," "concave points_worst," "concave points_mean," "perimeter_mean," "area_mean," and "concavity_mean." The correlation coefficients and *p*-values for these 8 features are displayed in Table 2. They exhibit a strong correlation with the data of breast cancer diagnosis and hence have been chosen for further analysis and visualization.

Data Visualization

The data was visualized using a feature subset constructed by high correlation filtering employing the Spearman correlation coefficient. American English spelling was not used. Initially, a scatterplot was plotted to analyze the data (Figures 6 to 9). The figures illustrate the two-dimensional (2D) distribution of the feature subset, providing insight into the attribute pairs' linear separability between benign and malignant tumors. Technical abbreviations will be defined when introduced. Regular author and institution formatting were maintained, along with conventional academic sectioning, style guides, and citation formats. The language is formal and objective, with no hedging or filler words. Grammatical correctness was ensured, and precise word choices were made. The figures illustrate the 2D distribution of the feature subset, providing insight into the attribute pairs' linear separability between benign and malignant tumors. Concave points_worst and perimeter_worst, in addition to concave points_mean and radius_worst, show clear linear differentiability. However, the attributes perimeter_mean, area_worst, radius_mean and area_mean are more closely distributed, making it difficult to distinguish their linear differentiability solely from 2D scatterplots. To investigate these attributes further, we created three-dimensional (3D) scatterplots using the 8 selected data points.

3D scatter plots reveal patterns of linear separability in the spatial distribution of benign and malignant breast cancer data. (Figures 10 and 11) Specifically, these patterns are observed for attributes such as perimeter_mean, area_worst, radius_mean and area_mean. This finding emphasizes the validity of using feature subsets created through high relevance filtering in training classification algorithms.

Table 1. Pearson's Correlation Coefficient Reference Table.

Pearson correlation coefficient range	Sample correlation strength
0.8–1.0	Very strong correlation
0.6–0.8	Strong correlation
0.4–0.6	Moderately correlation
0.2–0.4	Weak correlation
0.0–0.2	Very weak correlation or correlation

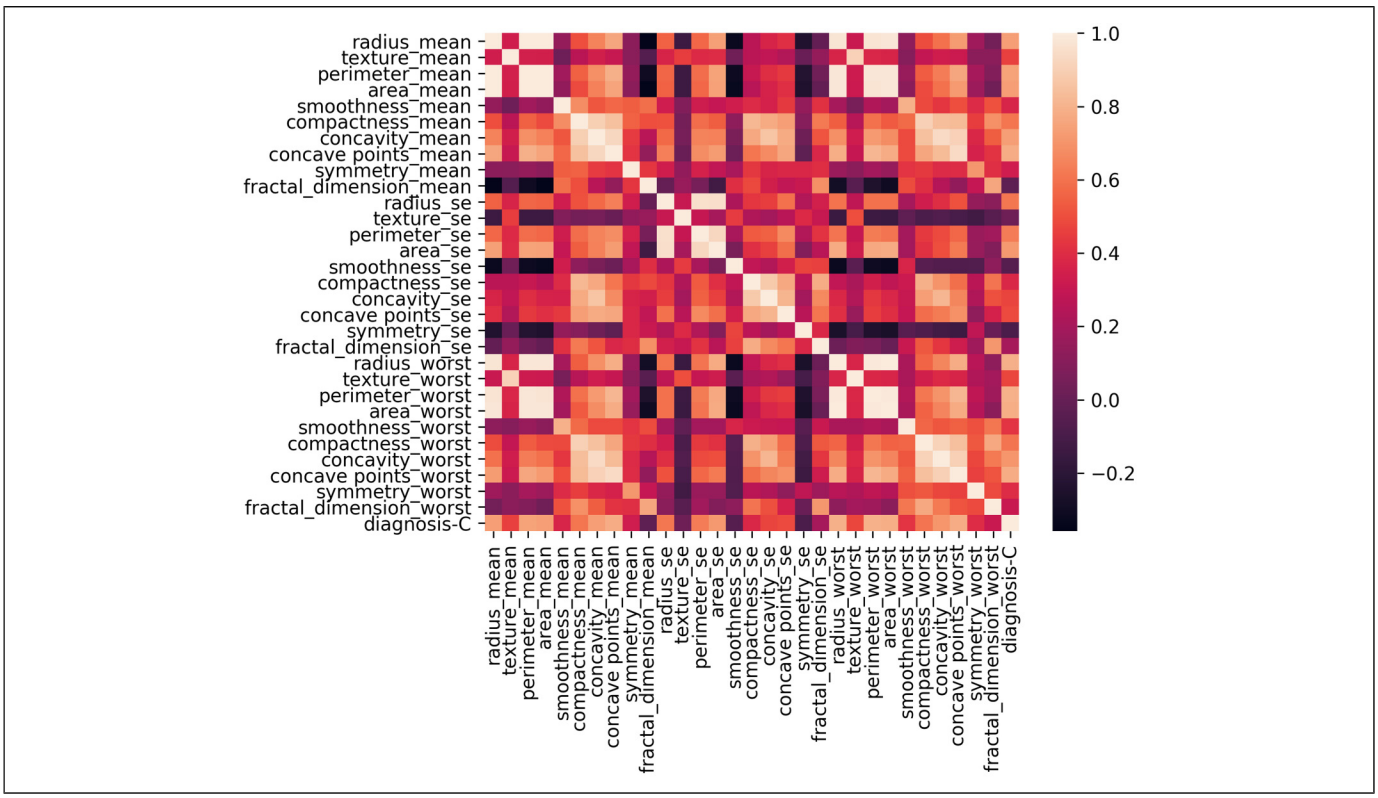

Figure 4. Heat map of breast cancer characterization correlations.

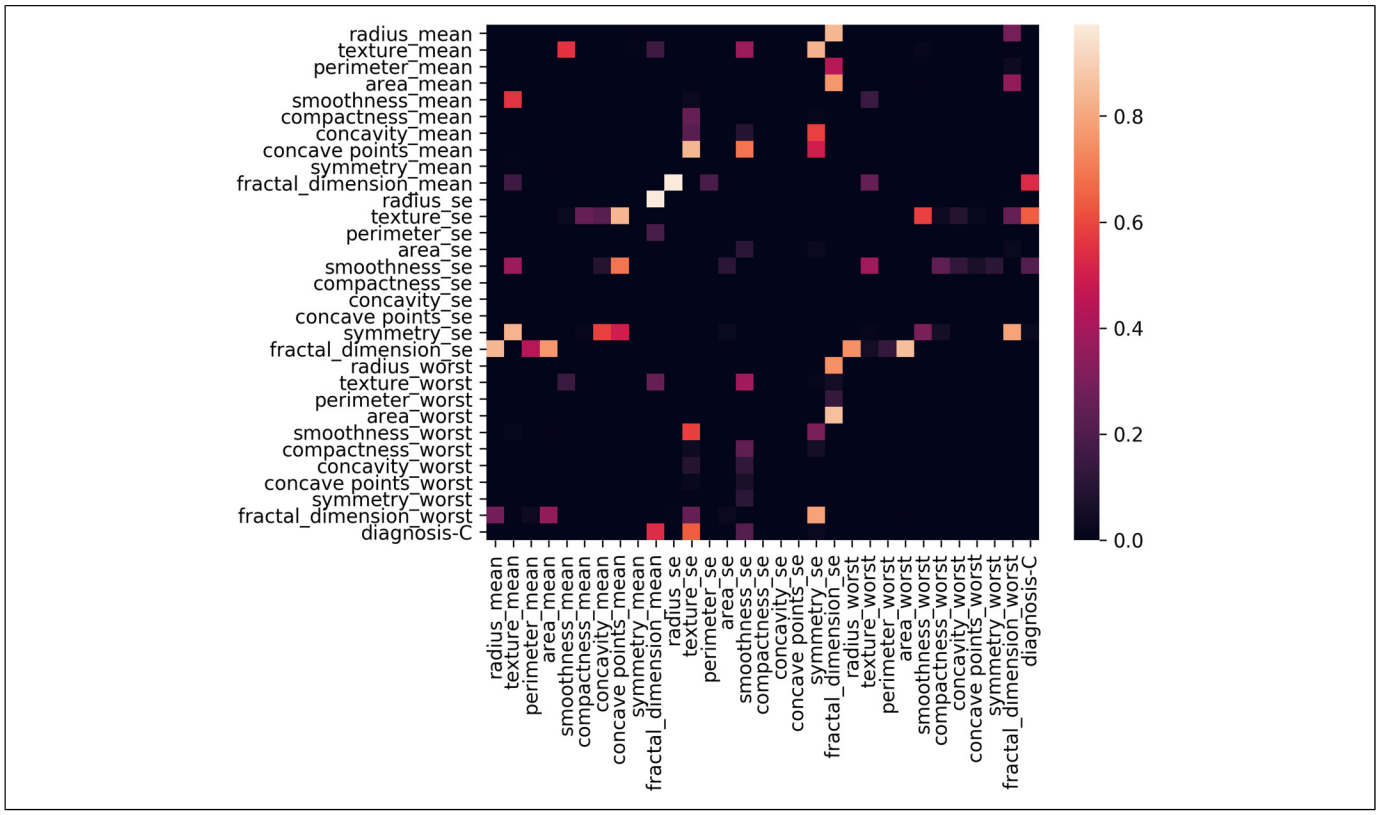

Figure 5. P-value matrix of characteristic data.

Table 2. Spearman's Correlation Coefficient and *P*-Value for 8 Feature Data.

Feature	Spearmanr	*P* value
perimeter_worst	0.796319	$6.742652 \times 10^{-126}$
radius_worst	0.787933	$1.677735 \times 10^{-121}$
area_worst	0.786902	$5.641911 \times 10^{-121}$
concave points_worst	0.781674	$2.387021 \times 10^{-118}$
concave points_mean	0.777877	$1.736699 \times 10^{-116}$
perimeter_mean	0.748496	$3.162795 \times 10^{-103}$
area_mean	0.734122	2.162221×10^{-97}
concavity_mean	0.733308	4.509903×10^{-97}

Wilcoxon Rank sum Test

To determine the statistical relationship between the data and the features, we selected the Wilcoxon rank sum test after conducting a normality test. This allowed us to comprehend the discrepancies in the distribution of the various breast cancer categories of the subset data. In this study, we present the distribution histogram for the 8 feature subset data, as illustrated in Figure 12. In addition, Figure 13 illustrates the distribution of the feature subset data for different types of breast cancer. The parameter values and *P*-values of the Wilcoxon rank sum test for the feature subset data are also displayed in Table 3. Our findings suggest that the P-values of the feature subset are significantly less than 0.01, leading to the rejection of the samples. The initial hypothesis was that the data originated from the same distribution. However, it has been identified that the feature subset data, obtained through screening, differs from this distribution.

Data Segmentation

Subsequently, 80% of the dataset was allocated for algorithm training, while the remaining 20% was used for testing algorithm performance. The training set consisted of 284 instances labeled as benign (type B) and 171 instances labeled as malignant (type M), while the test set comprised 73 instances of type B and 41 instances of type M. To mitigate the impact of dissimilar data distributions on model training speed, both the training and test data were normalized using equation (1). The box plot after normalization is completed is shown in Figures 14 and 15.

After normalizing the data, the distribution of the training data is adjusted to be within the same range. Furthermore, an analysis of the box plot identifies certain abnormal data points within the datasets. In order to handle these outliers, the following methods are utilized:

- First calculate the first quartile (Q1), median, and third quartile (Q3).
- Interquartile range (IQR) = Q3−Q1, data upper limit = Q3 + 1.5*IQR, 0 data lower limit = Q1–1.5*IQR.
- Data value > upper data limit or data value < lower data limit is anomalous data.

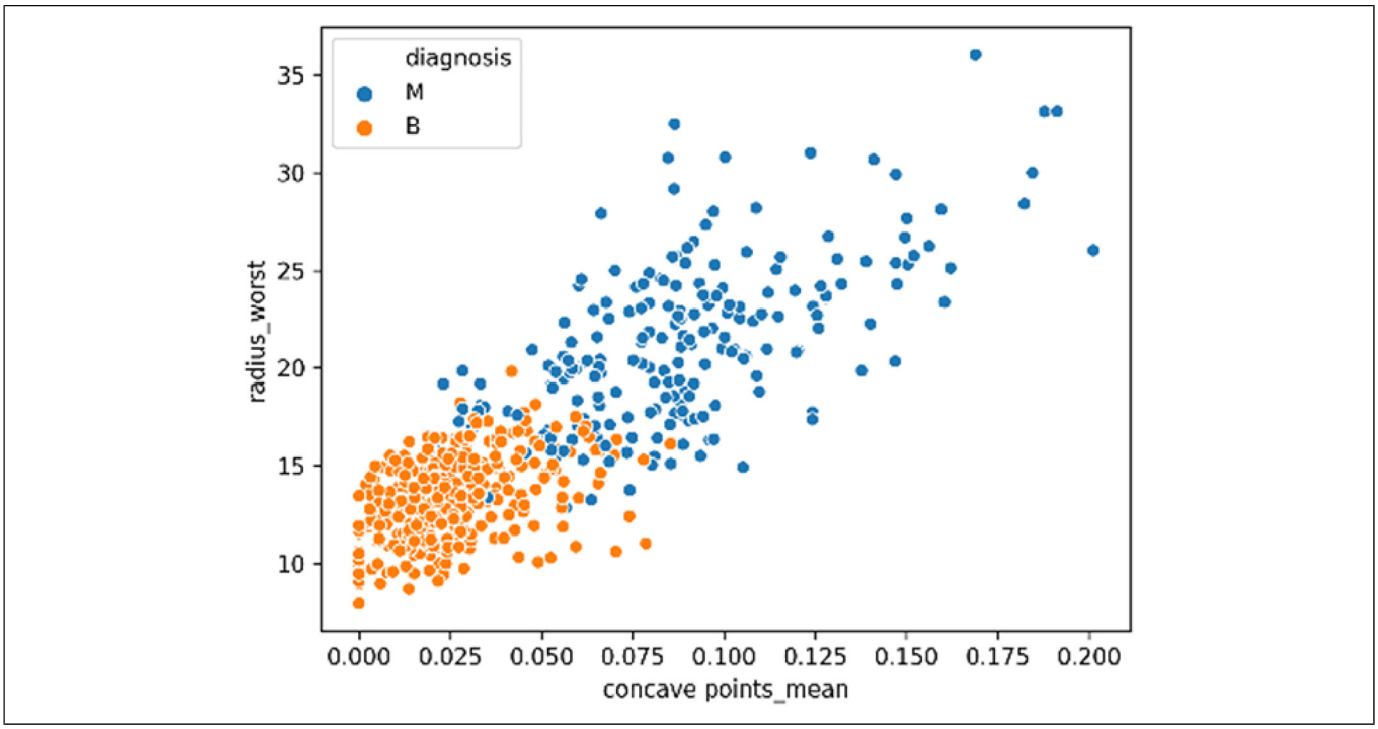

Figure 6. Concave points_worst,perimeter_worst scattering graph.

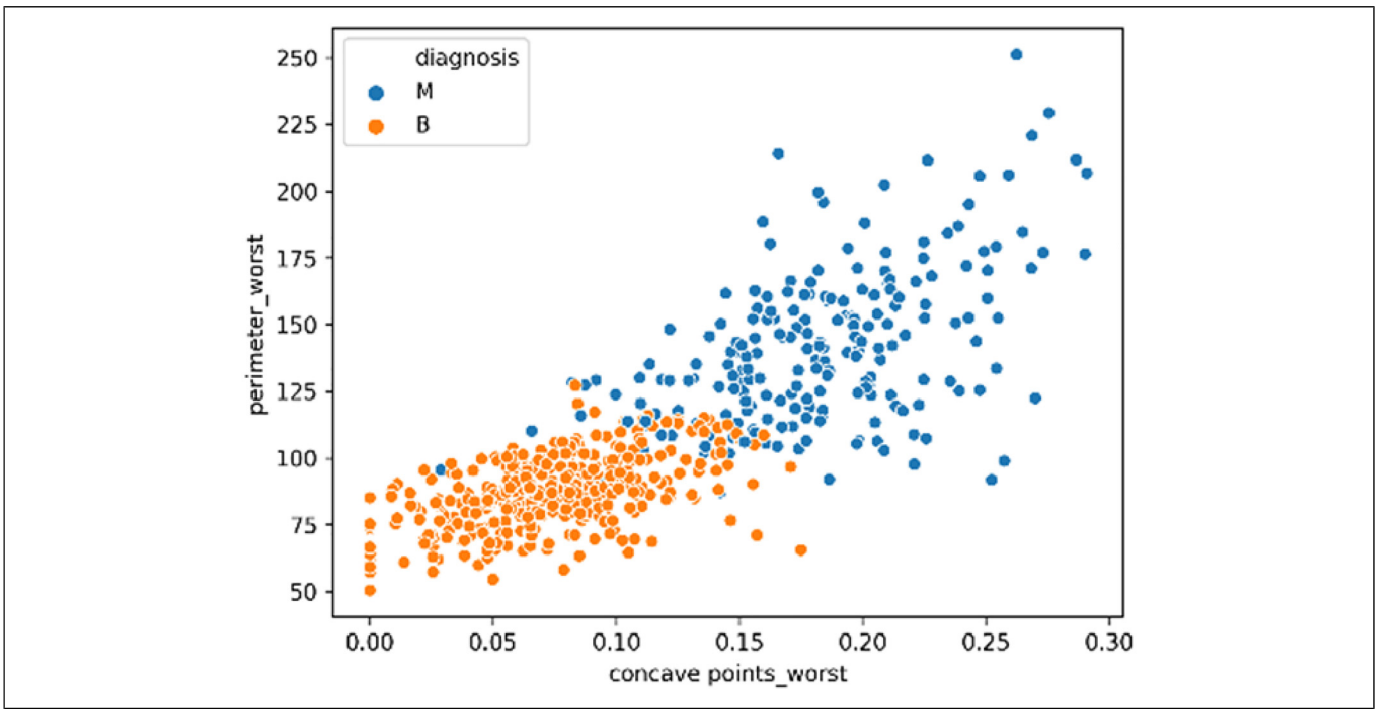

Figure 7. Concave points_mean, radius_worst scattering graph.

- Replace anomalous data with median data.

Upon processing the anomalous data, the resulting dataset, which incorporates normalization and removal of outliers, is suitable for the training and testing of machine learning models.

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

Z is the normalized value, x is the initial value, μ is the mean, and the σ standard deviation.

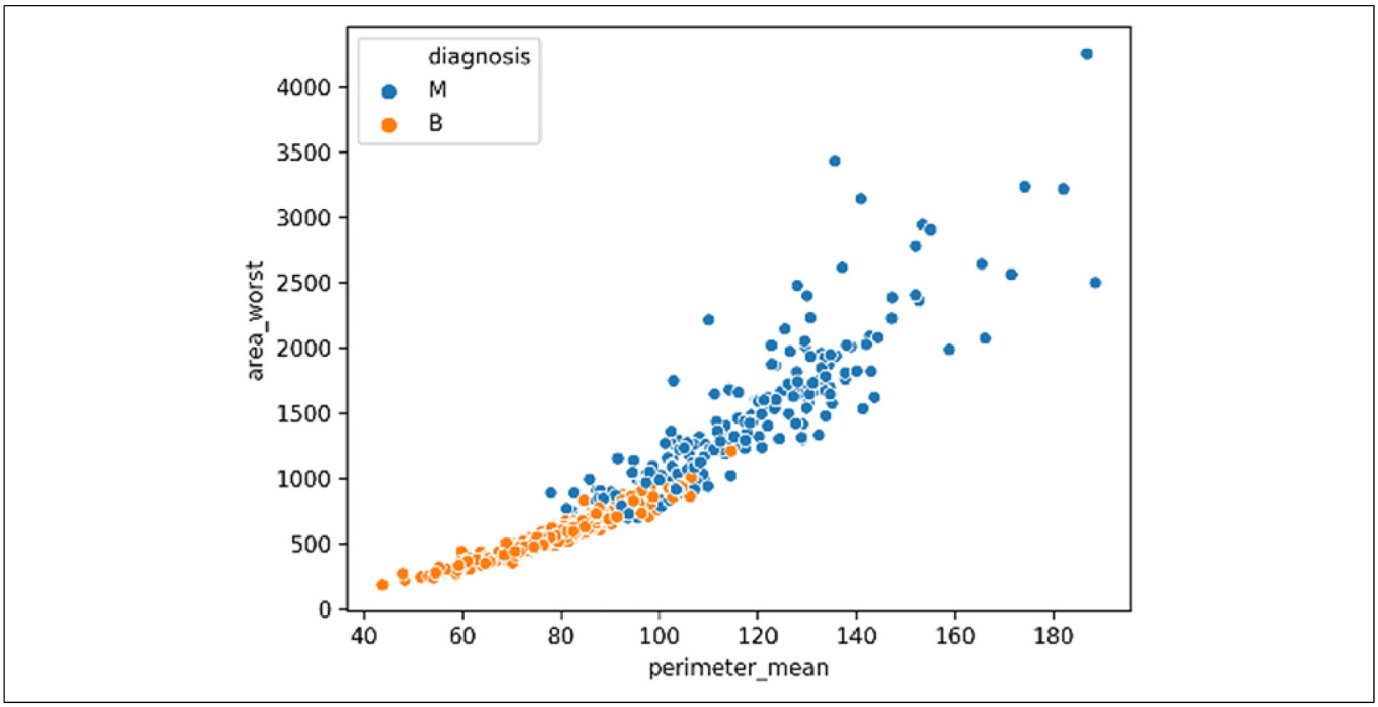

Figure 8. Perimeter_mean,area_worst scattering graph.

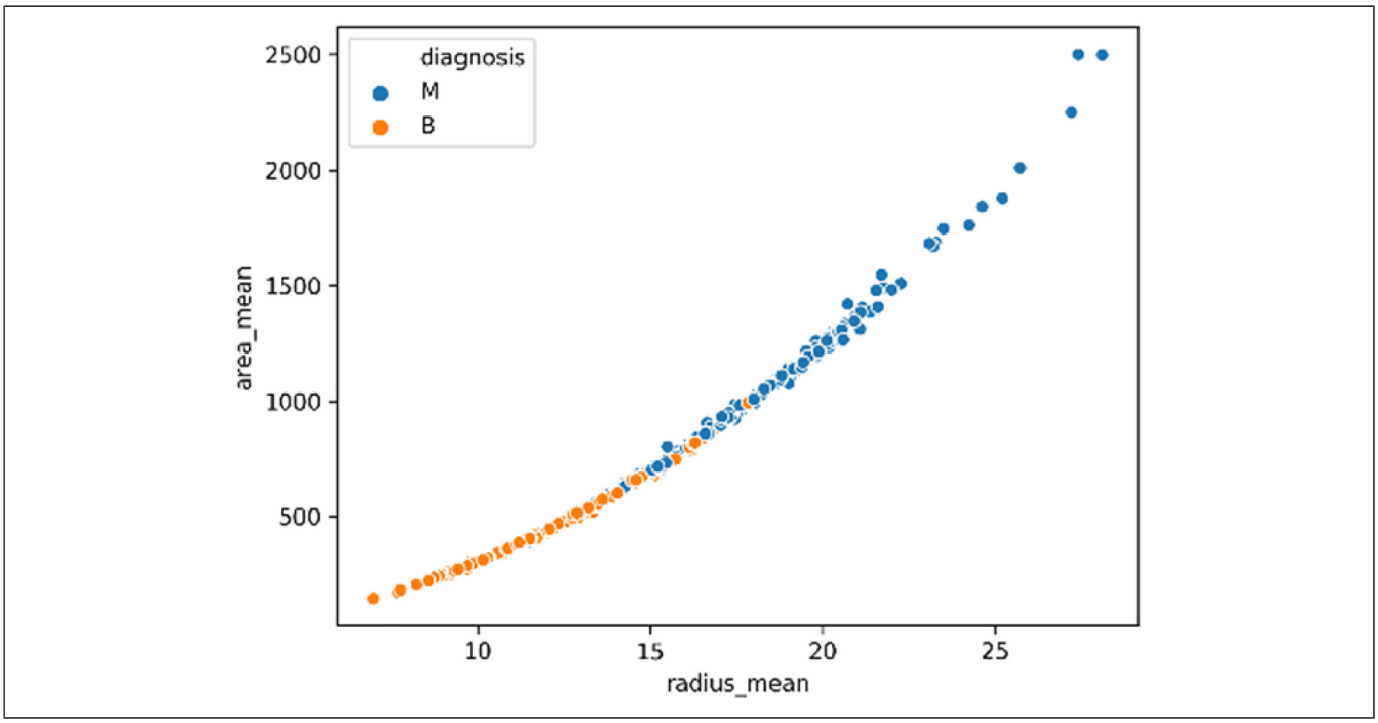

Figure 9. Radius_mean,area_mean scattering graph.

Build Machine Learning Models

We analyze the given data through the utilization of various machine learning models, namely decision tree models, random gradient descent models, random forest models, K-NN models, SVM models, logistic regression models, and AdaBoost models. The subsequent section provides a concise overview of the training process associated with these machine-learning models.

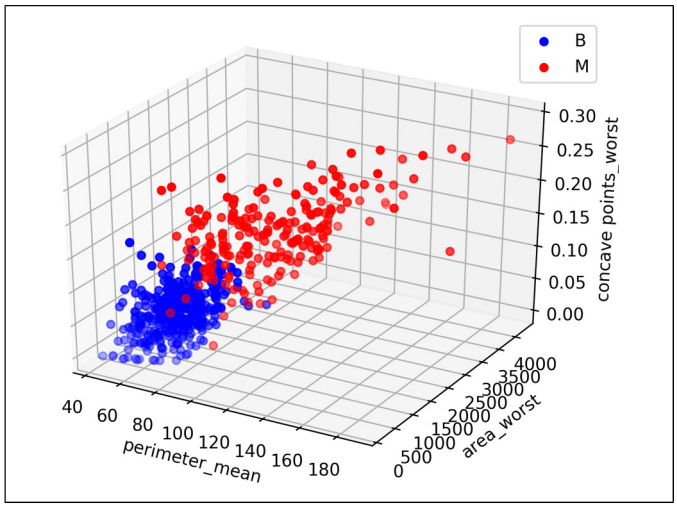

Figure 10. Perimeter_mean、area_worst、concave points_worst scattering graph.

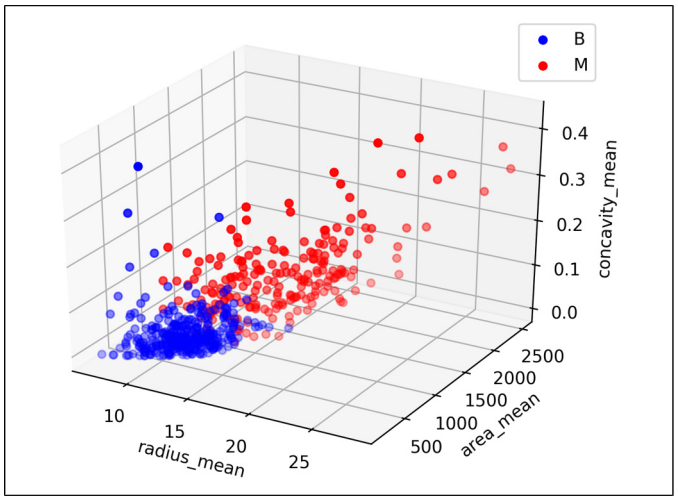

Figure 11. Radius_mean、area_mean、concavity_mean scattering graph.

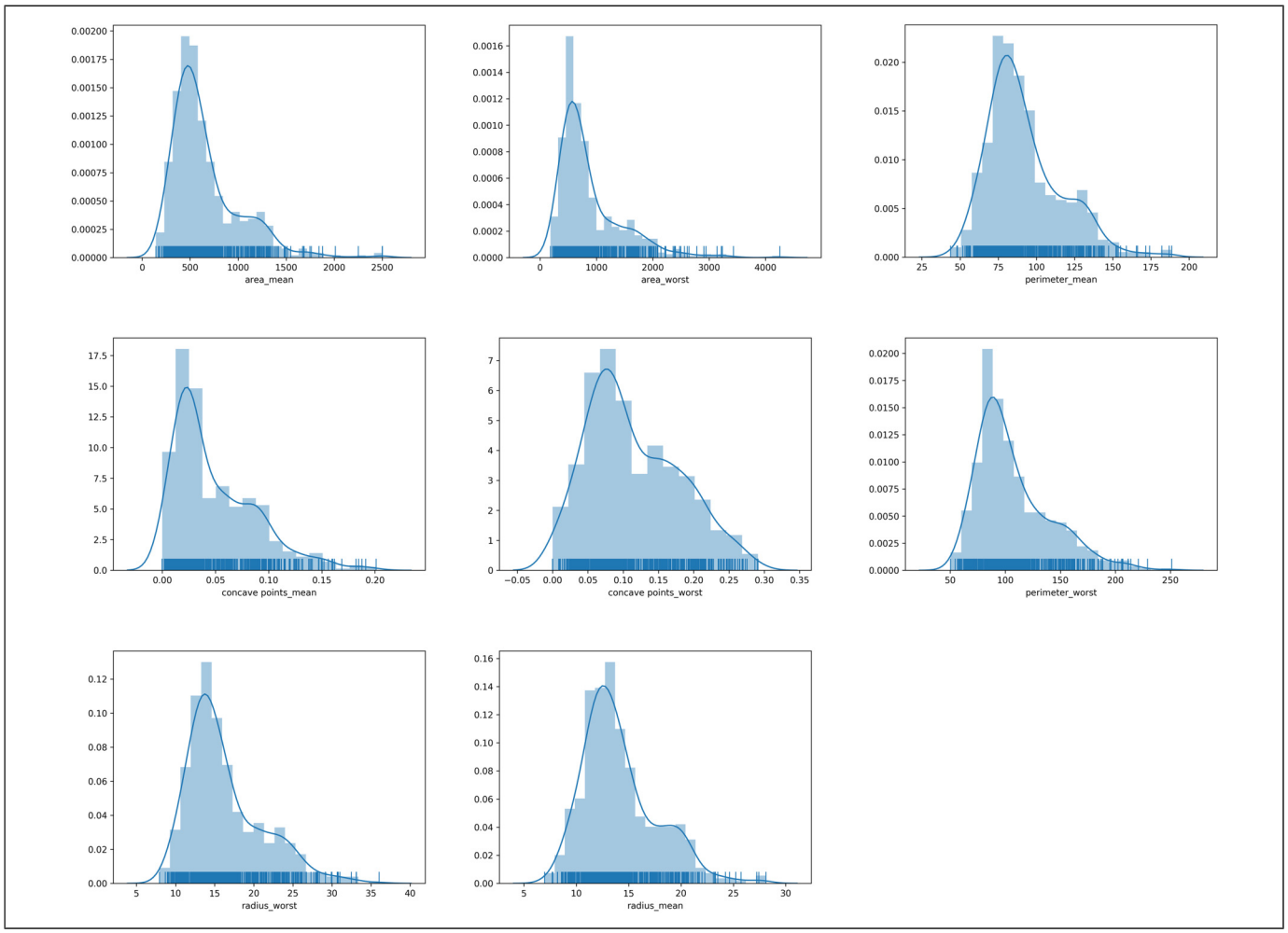

Figure 12. Distribution of feature subset data.

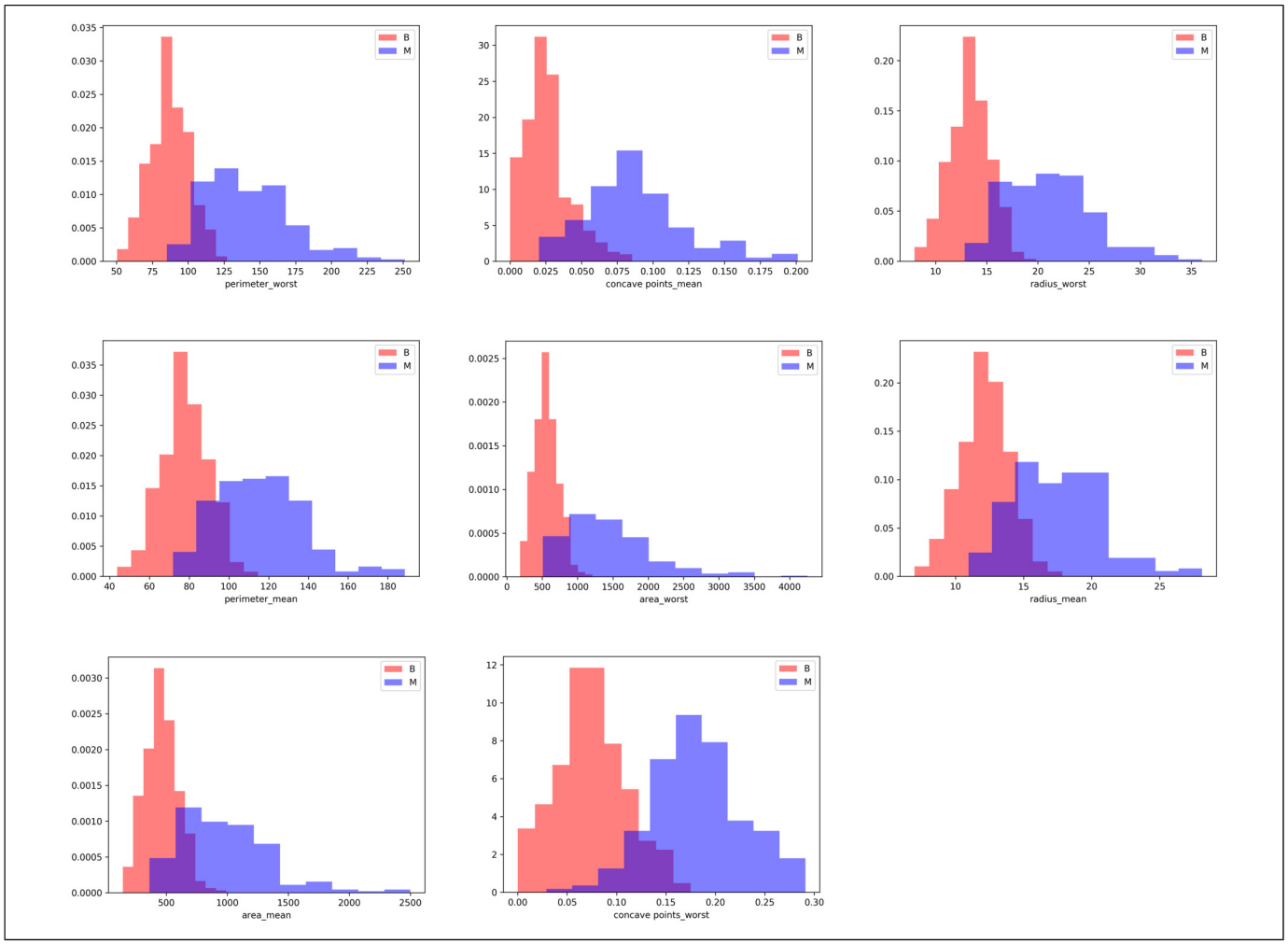

Figure 13. Histogram of the data distribution of the subset of characteristics of different categories of breast cancer.

Table 3. Table of Results of Wilcoxon Rank sum Test for Subset of Features.

Feature	Statistic	*P*-value
concave points_worst	7.0230087465450115	$2.171406127996621 \times 10^{-12}$
perimeter_worst	29.201894775379632	$1.8347146160092246 \times 10^{-187}$
concave points_mean	7.0230087465450115	$2.171406127996621 \times 10^{-12}$
radius_worst	29.201894775379632	$1.8347146160092246 \times 10^{-187}$
perimeter_mean	29.201894775379632	$1.8347146160092246 \times 10^{-187}$
area_worst	29.201894775379632	$1.8347146160092246 \times 10^{-187}$
radius_mean	29.201894775379632	$1.8347146160092246 \times 10^{-187}$
area_mean	29.201894775379632	$1.8347146160092246 \times 10^{-187}$

Decision Tree Algorithm. The decision tree algorithm is a supervised machine learning method that overcomes the data distribution requirements associated with classical algorithms such as general linear models, generalized linear models, and Bayesian classifiers.[19] Unlike these traditional algorithms, the decision tree algorithm can establish logical connections between input and output data without being limited by data distributions. As a result, it can make predictions for new data based on the discovered rules or logical connections. In our study, we utilized the DecisionTreeClassifier from the scikit-learn library to construct a decision tree classification model. Subsequently, we conducted tests on the model by adjusting the max_depth parameter using the data that was selected and extracted based on the features. Figure 16 illustrates the training error and testing error of the decision tree algorithm at different max_depth parameters. By doing this, we were able to determine the optimal max_depth parameter that resulted in the lowest test error.

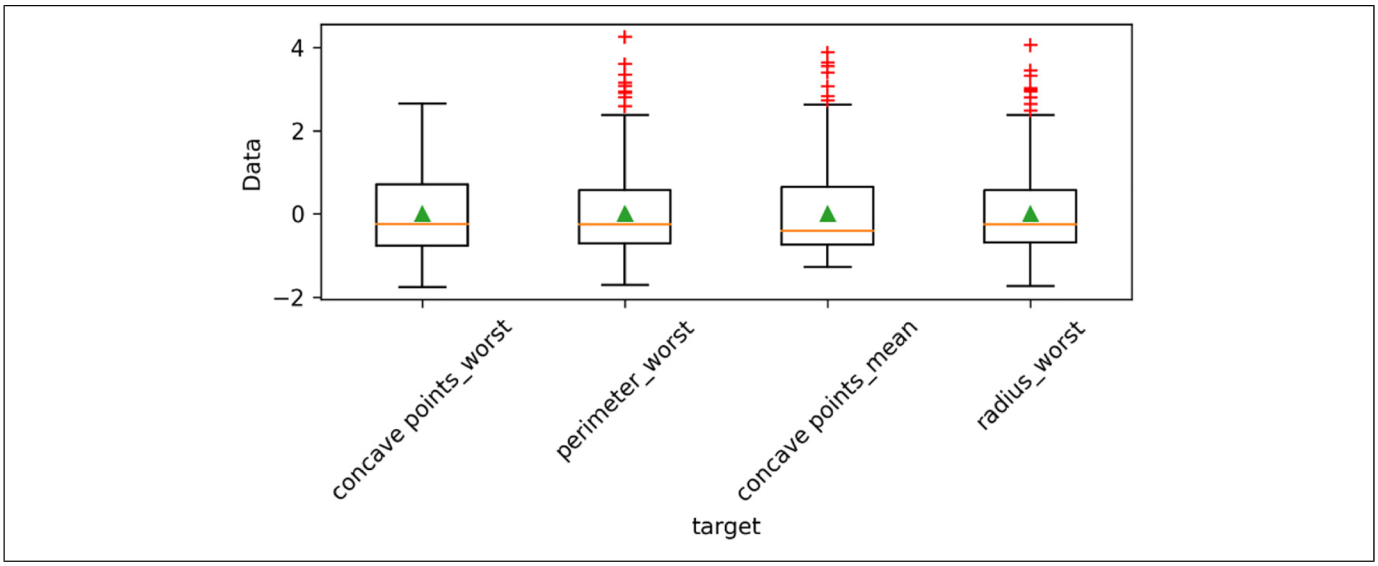

Figure 14. Box plot of feature data after data normalization.

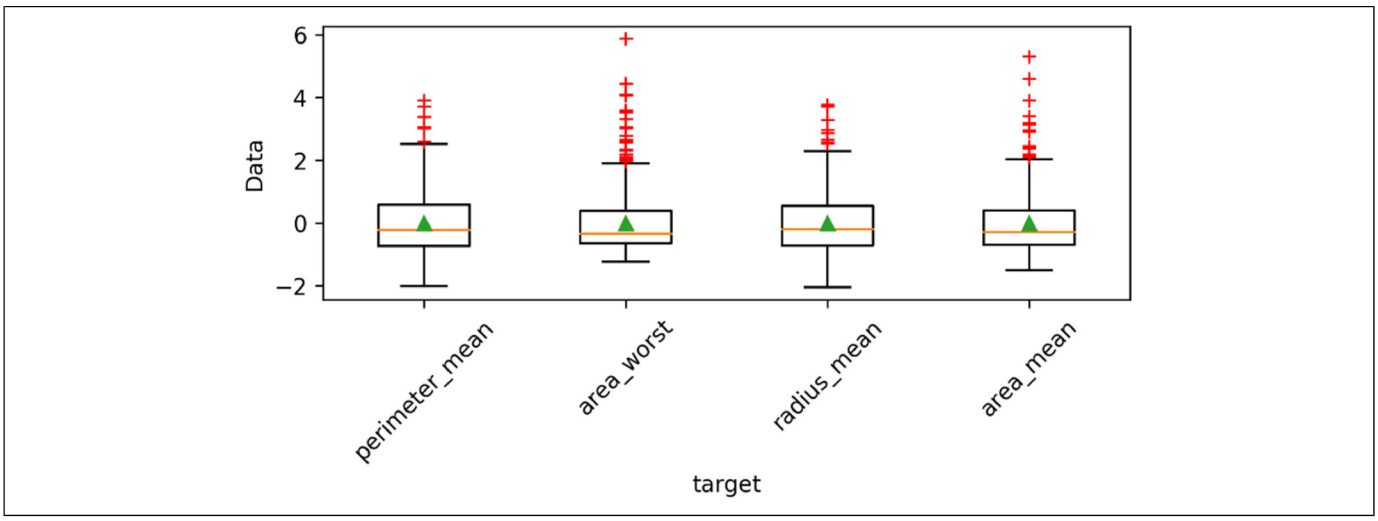

Figure 15. Box plot of feature data after data normalization.

Random Forest Algorithm. The random forest algorithm is an ensemble method that combines multiple decision tree algorithms, typically using the bagging or pasting training approaches.[20] Instead of searching for the optimal features when splitting nodes, it searches for the best features among a randomly selected subset of features. This increases the diversity among the decision trees, leading to better generalization performance.

We utilized the RandomForestClassifier from the scikit-learn library to construct a random forest classification model. The feature-selected datasets were used to train the model, enabling us to examine the training and testing errors of the random forest model for different numbers of base classifiers (n_estimators). The training and testing errors of the random forest model using different number of base classifiers (n_estimators) in the random deep forest model are presented in Figure 17. From Figure 17, it can be observed that Random Forest achieves the best training and testing error when the number of base classifiers is 40.

Other Machine Learning Algorithms Trained. We trained the Decision Tree and Random Forest algorithms and selected additional machine learning algorithms for training based on the characteristics of the data distribution and their applicability. These algorithms include the SGD algorithm, K-NN algorithm, SVM algorithm, Logistic regression algorithm, and AdaBoost algorithm. Here are the characteristics of these algorithms:

- The SGD algorithm minimizes the cost function by iteratively adjusting the parameters,[21] leading to faster convergence and higher accuracy.
- K-NN is a classification and regression method first proposed by Cover T and Hart P in 1967.[22] The K-NN

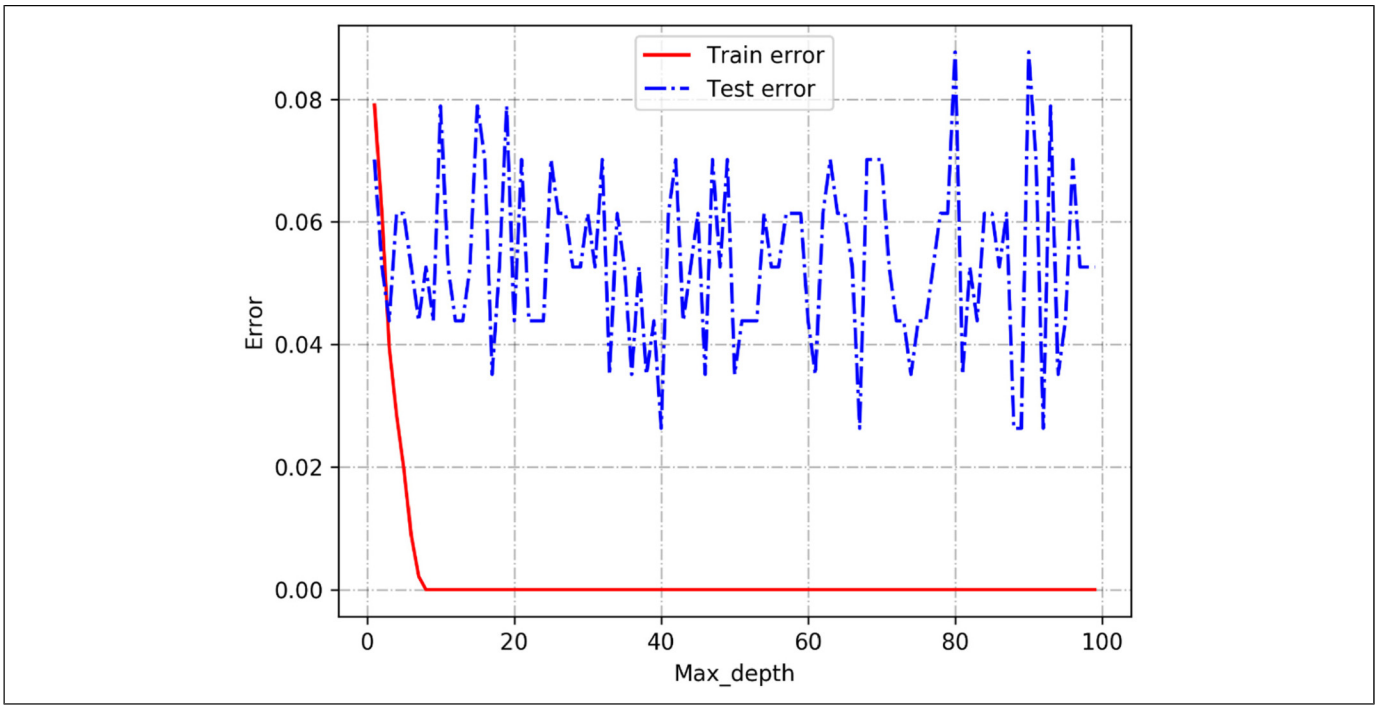

Figure 16. Decision tree max depth-error diagram.

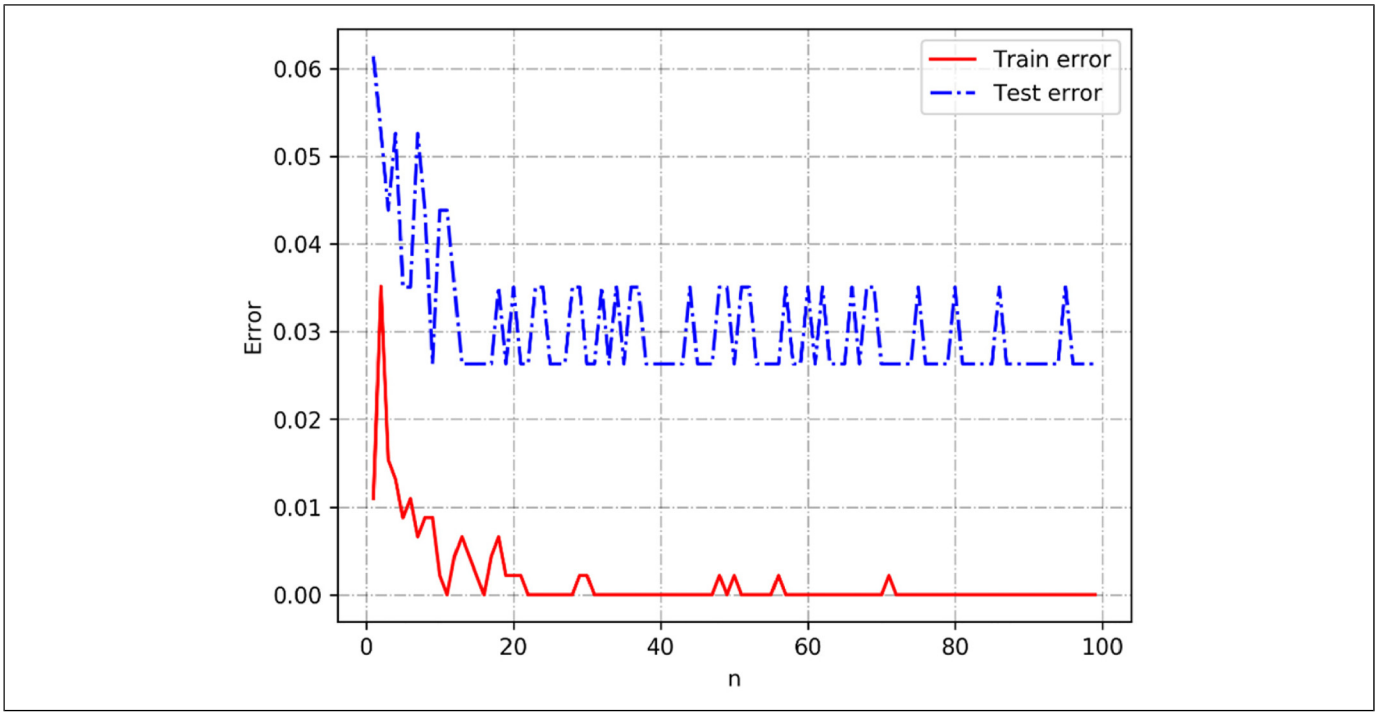

Figure 17. Random forest model number of classifiers - error plot.

method shows good performance in cases where the spatial dimensions of the input variables are small and the actual category boundaries are irregular.

- SVM is a supervised learning algorithm developed in 1992 by Boser, Guyon, and Vapnik based on statistical learning theory.[23] The SVM algorithm has shown strong performance in solving classification and regression problems with small sample sizes, nonlinearity, and high dimensionality. The 2D scatterplot and 3D scatterplot show that the feature subset has some linear divisibility in spatial distribution, so we try to use the SVM algorithm for breast cancer classification and diagnosis task.

- Logistic regression algorithms are widely used to assess the probability of state data belonging to a certain category,[24] and based on the characteristics of the data distribution presented by the data visualization, we use the Logistic regression algorithm to train the data.
- The AdaBoost algorithm is a boosting method in integrated learning that combines several weak classifiers to create a strong classifier. During runtime, the algorithm focuses more on instances that were previously underfitted and misclassified, thus improving the performance of the predictive model.[25]

Finally, we created and trained our machine learning model using the sklearn[26] tool in the Python environment.

Machine Learning Algorithm Interpretability. In medicine, models or systems that lack clear explanations for their decisions pose difficulty and the opaque medical practice of AI algorithms obstructs clinicians from evaluating model inputs and parameters accurately.[27] Failure to understand the decision-making process may infringe on patients' rights to informed consent and automation.[28] Therefore, AI interpretability plays a crucial role in applying AI models to healthcare organizations for research in the field of AI + Medicine. Digital Health, a subjournal of the Lancet, suggests that the use of explainable frameworks could help to align model performance with clinical guidelines objectives. Therefore, enabling better adoption of AI models in clinical practice. Transparent algorithms or explanatory approaches can also make the adoption of AI systems less risky for clinical practitioners.[29]

From the perspective of model relevance, interpretable AI its can be classified into model-independent interpretation methods and model-specific interpretation methods.[30] The main model-related interpretation methods are: Feature-related interpretation, sample-based interpretation and agent model interpretation.[31] In our investigation, we predominantly employ feature-based and sample-based interpretation methods to probe the interpretability of our algorithms. During the feature-based interpretation method, we utilize statistical methods to examine the data's distribution, the correlation between the feature data and the target data, and the dissimilarities in the data's distribution among the diverse categories of breast cancer.

Results

First, we assess the algorithm's overall classification performance by calculating its accuracy, precision, and F1 score. Additionally, we analyze the confusion matrix of the algorithm on both the test and training datasets to evaluate its specific classification performance. This evaluation provides valuable insights and suggestions for optimizing the algorithm. The accuracy, precision, F1 score, recall, and confusion matrices of each algorithm are presented below 40:

Decision Tree Model Training Results

The decision tree algorithm training set and test set confusion matrices are shown in Figures 18 and 19, and the accuracy,

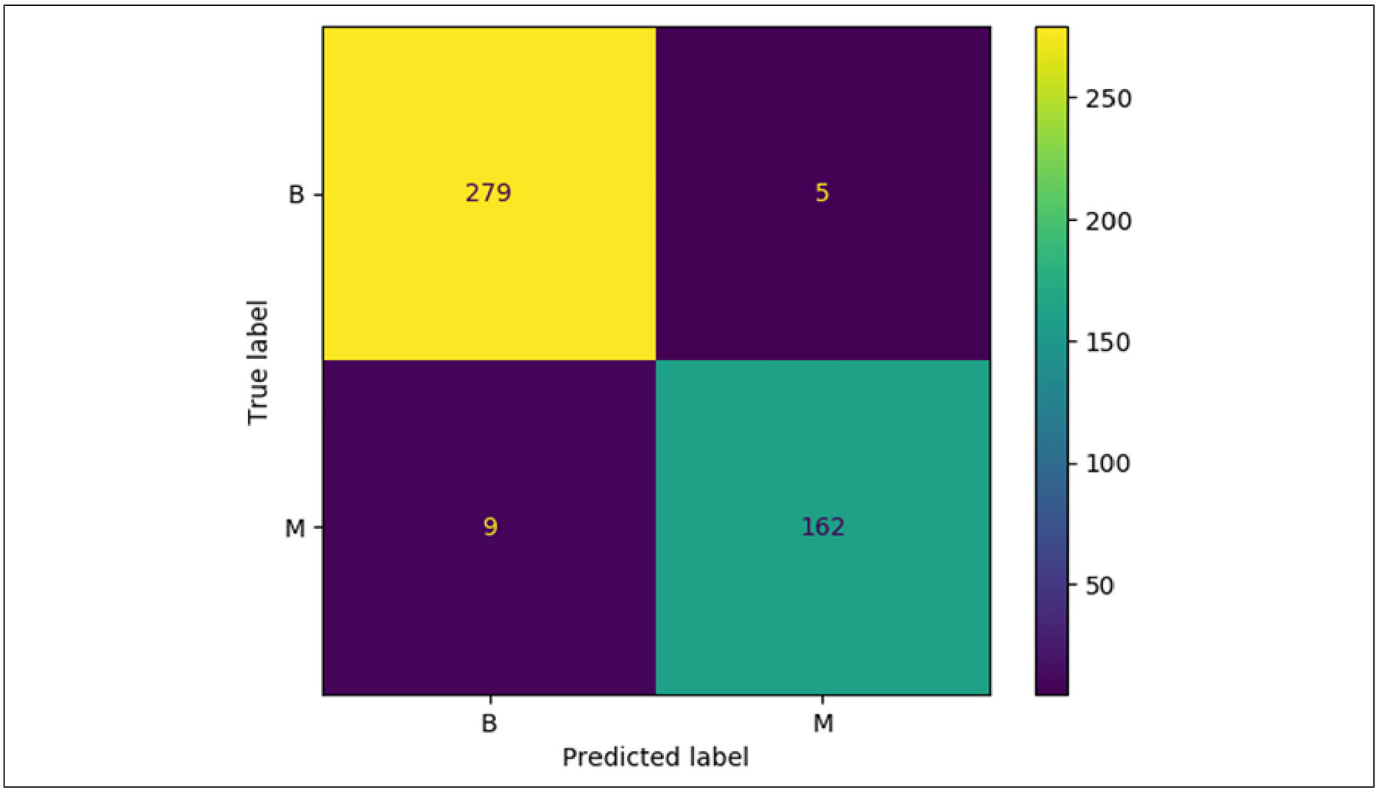

Figure 18. Decision tree model confusion matrix (training set).

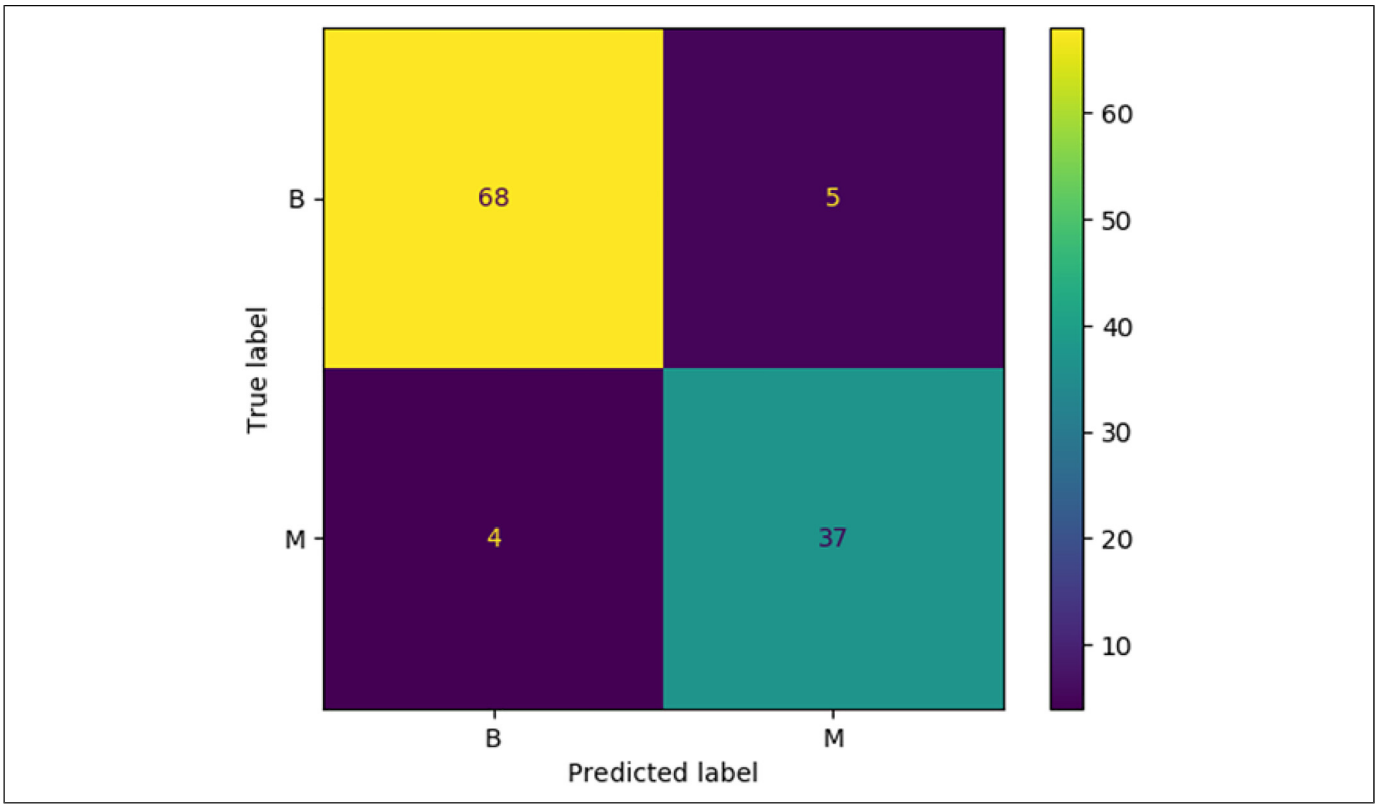

Figure 19. Decision tree model confusion matrix (test set).

Table 4. Classification Results of Decision Tree Algorithm Test set.

	Precision	Recall	F1-score	Support
B	0.94	0.93	0.94	73
M	0.88	0.90	0.89	41
Accuracy			0.92	114
Macro avg	0.91	0.92	0.91	114
Weighted avg	0.92	0.92	0.92	114

Table 5. Classification Results of Gradient Descent Algorithm Test set.

	Precision	Recall	F1-score	Support
B	0.96	1.00	0.98	73
M	1.00	0.93	0.96	41
Accuracy			0.97	114
Macro avg	0.98	0.96	0.97	114
Weighted avg	0.97	0.97	0.97	114

precision, recall and F1 scores are shown in Table 4. The training set's confusion matrix indicated a precision of 0.96875, a recall of 0.98239, an F1 score of 0.97552, and an accuracy of 0.96923. For the test set, the respective precision, recall, F1 score, and accuracy values for the "B" and "M" categories are presented in Table 4. The table reveals that the decision tree algorithm exhibits weaker classification performance for "M" data compared to "B" data. Class "B" data showcases superior performance, with precision, recall, F1 score, and accuracy all exceeding 0.9. Conversely, "M" data exhibits lower precision and F1 score values. This discrepancy is likely influenced by the smaller quantity of "M" data, resulting in limited performance for the decision tree model in classifying "M" data.

SGD Training Results

The results of the test set for the SGD algorithm are shown in Table 5. From the training set confusion matrix, it is observed that the gradient descent model achieved a precision of 0.94612, a recall of 0.98943, an F1 score of 0.96729, and an overall accuracy of 0.95824. The classification performance of the model on the test set is presented in Table 5. Table 5 demonstrates that the gradient descent model performed exceptionally well on the test set, achieving a recall of 1.00 for class 'B' data, surpassing that of the decision tree model (0.93). Moreover, the model attained a recall of 0.93 for class 'M' data, outperforming the decision tree model (0.90). These findings indicate that the gradient descent model exhibited superior classification performance compared to the decision tree model.

Random Forest Training Results

The random forest algorithm training set and test set confusion matrices are shown in Figures 20 and 21. According to the training set confusion matrix, the random forest model

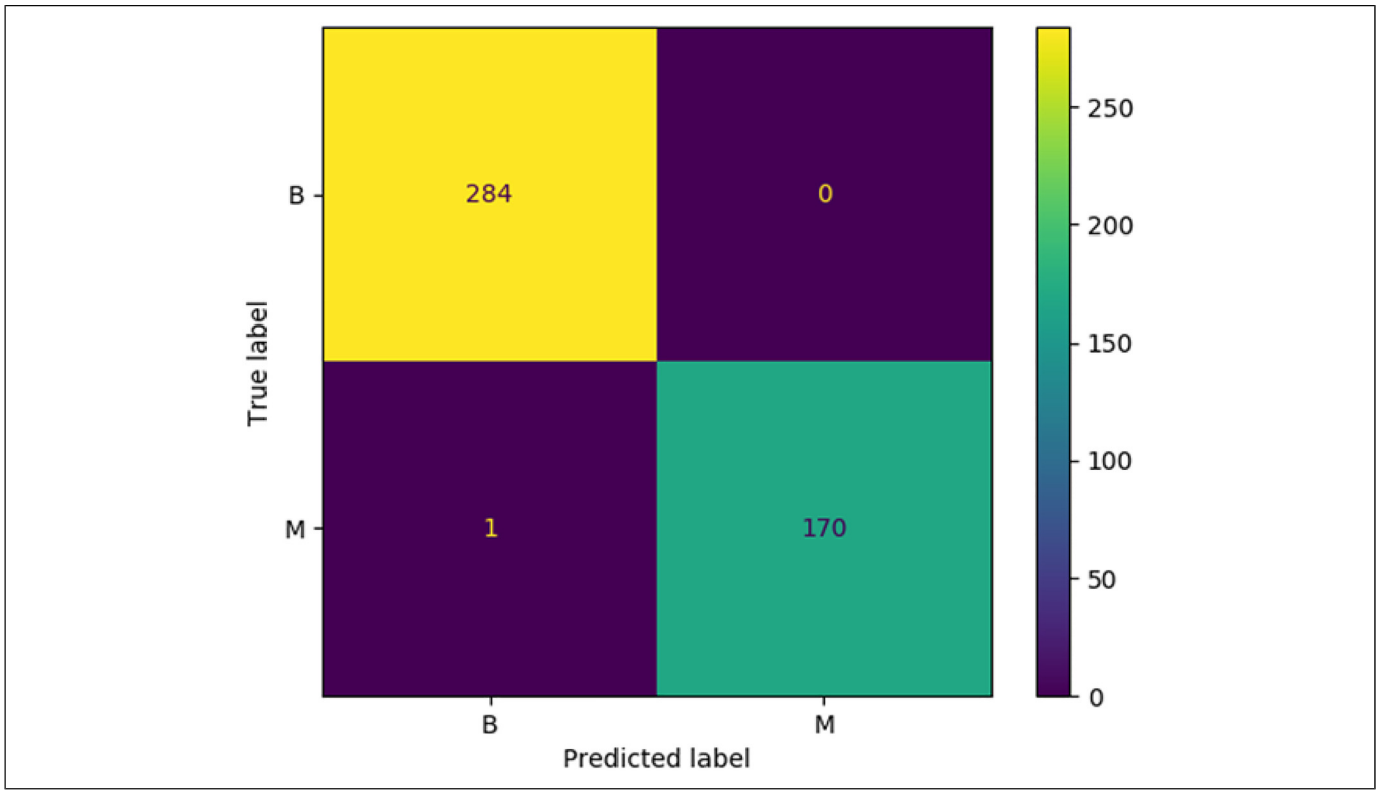

Figure 20. Confusion matrix of random forest model (training set).

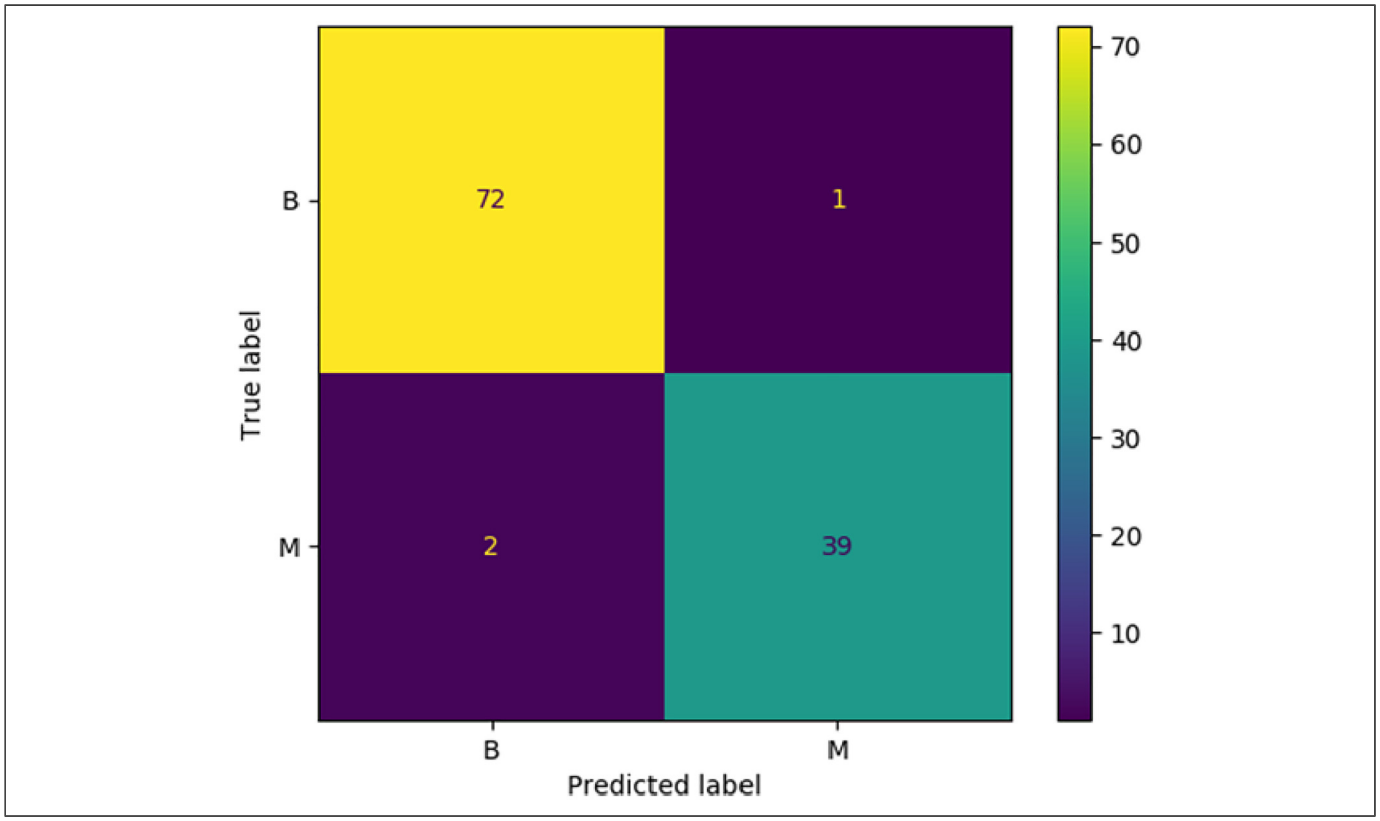

Figure 21. Confusion matrix of random forest model (test set).

demonstrates a precision of 0.99649, recall of 1.00, F1 score of 0.99824, and accuracy of 0.99780. The precision, recall, F1 score, and accuracy of the test datasets are shown in the following Table 6.

Table 6, it can be observed that the Random Forest algorithm achieves a recall of 0.95 in class "M" data classification, which is on par with the gradient descent algorithm. However, in class "B" data classification, the recall of the random forest model is 0.99, which is lower than the gradient descent model's 1.00. The average precision of the model is 0.97. Therefore, the classification accuracy of Random Forest is better than Decision Tree Algorithm and SGD algorithm.

K-NN Algorithm Training Results

The K-NN algorithm training set and test set confusion matrices are shown in Figures 22 and 23, and the accuracy, precision, recall and F1 scores are shown in Table 7. From the figure, we observe that the precision for the K-NN model on the training set is 0.94039, the recall is 1.00, the F1 score is 0.96928, and the accuracy is 0.96043. The data table reveals that the K-NN model has superior classification performance for class "B" data with a recall of 1.00, surpassing that of the decision tree model (0.93), gradient descent model (0.97), and random forest model (0.99). The data table reveals that the K-NN model has superior classification performance for class "B" data with a recall of 1.00, surpassing that of the decision tree model (0.93), gradient descent model (0.97), and random forest model (0.99). It can be inferred that K-NN model is the most effective in classifying class "B" data. Furthermore, the K-NN model demonstrates an accuracy of 0.95614 but a recall of 0.88 for class "M" data. Subsequently, its recall values fall short when compared to those of the decision tree model (0.90), the gradient descent model (0.93), and the random forest model (0.95). This outcome infers that the efficiency of classifying class "M" data can still be elevated within the K-NN model.

Table 6. Classification Results of the Test set of Random Forest Algorithm.

	Precision	Recall	F1-score	Support
B	0.97	0.99	0.98	73
M	0.97	0.95	0.96	41
Accuracy			0.97	114
Macro avg	0.97	0.97	0.97	114
Weighted avg	0.97	0.97	0.97	114

Support Vector Machine Algorithm Training Results

The SVM algorithm training set and test set confusion matrices are shown in Figures 24 and 25, and the accuracy, precision, recall, and F1 scores are shown in Table 8. The SVM model has a precision of 0.94256, a recall of 0.98239, an F1 score of 0.96206, and an accuracy of 0.95164 on the training set,

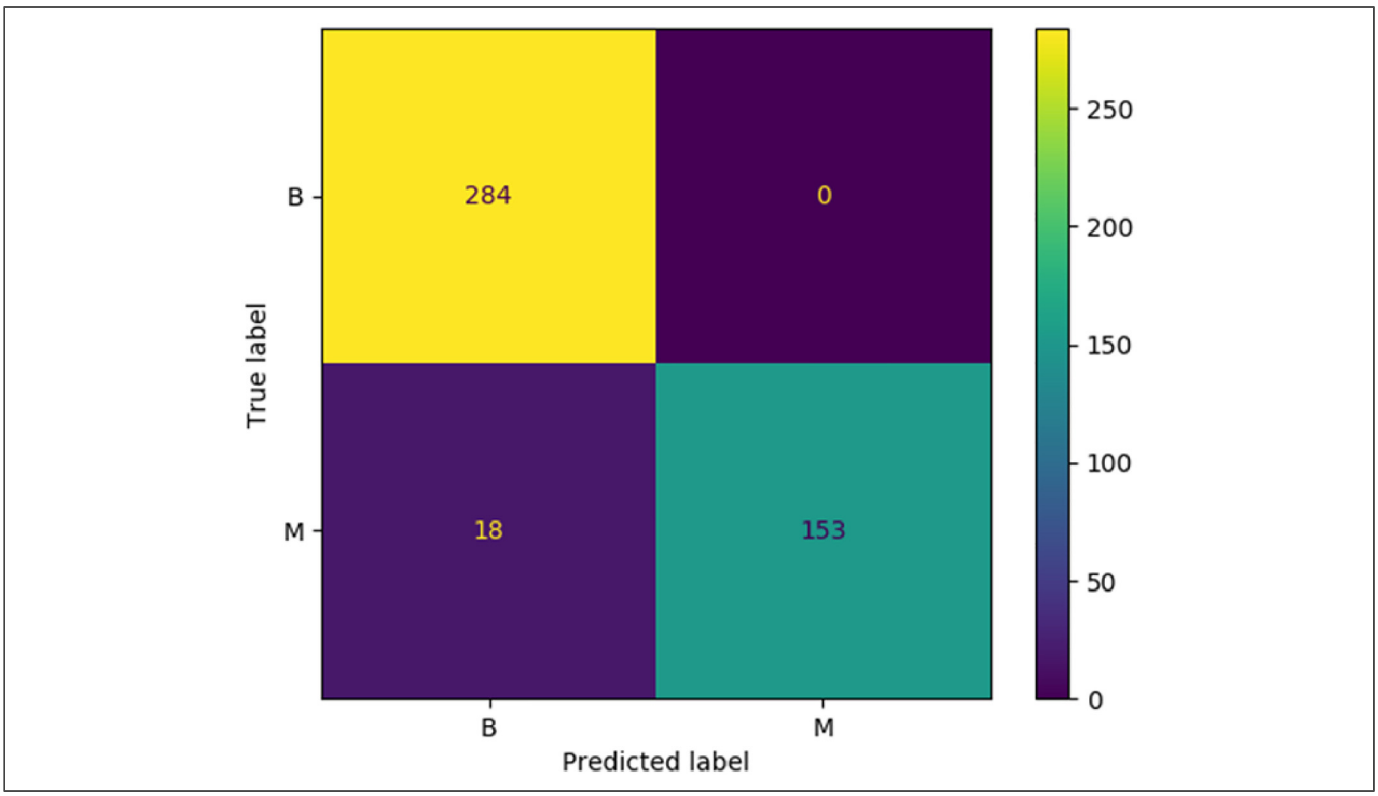

Figure 22. Confusion matrix of K-nearest neighbor model (train set).

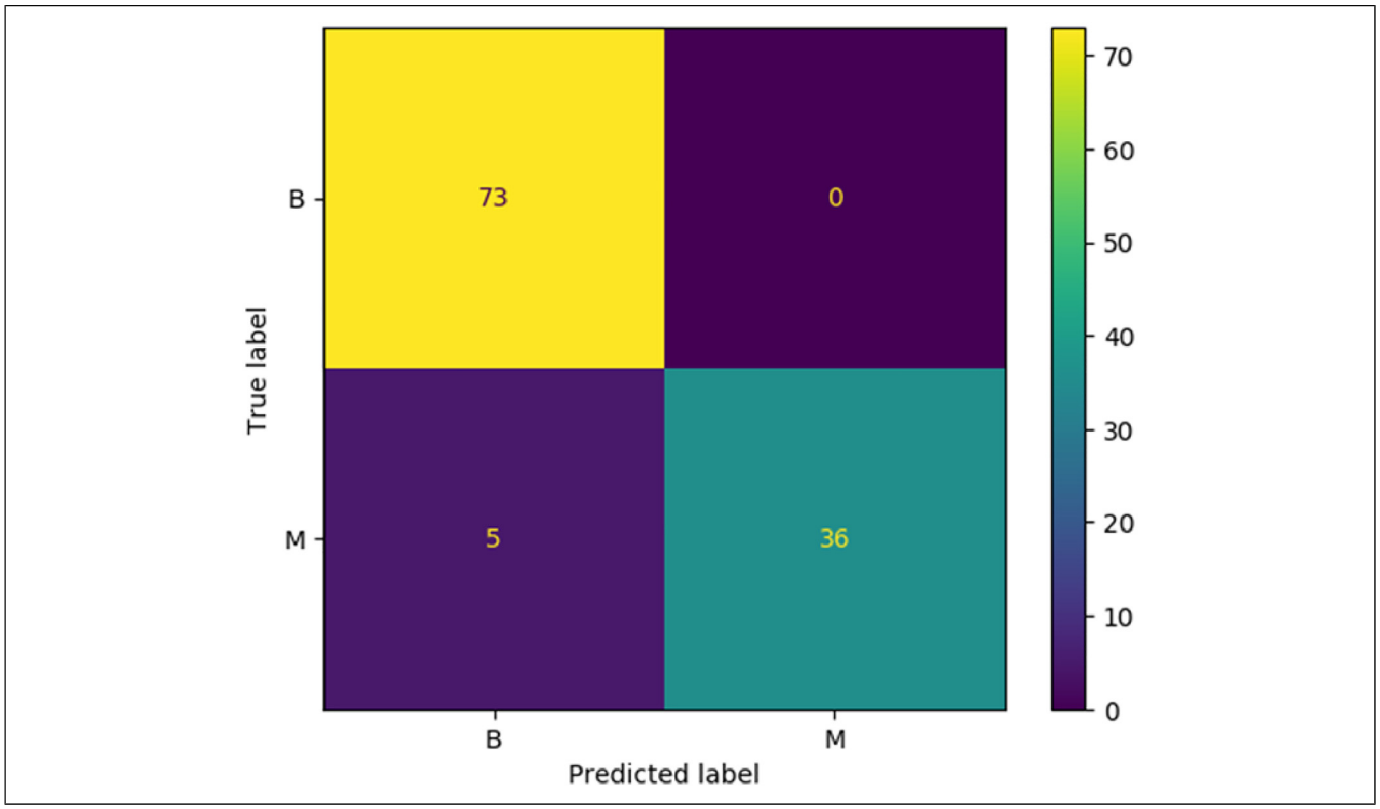

Figure 23. Confusion matrix of K-nearest neighbor model (test set).

Table 7. Classification Results of K-Nearest Neighbor Algorithm Test set.

	Precision	Recall	F1-score	Support
B	0.94	1.00	0.97	73
M	1.00	0.88	0.94	41
Accuracy			0.96	114
Macro avg	0.97	0.94	0.95	114
Weighted avg	0.96	0.96	0.96	114

and the SVM algorithm performs better on the training dataset. Figure 25 shows the confusion matrix obtained by the SVM algorithm after testing on the test dataset, through Python toolkit, we calculate the detailed data of the test set of confusion evidence as shown in Table 8. We find that the SVM model outperforms the Decision Tree, Gradient Descent, and Random Forest models in terms of recall for class "B" data classification, achieving a recall of 1.0, which is comparable to the classification performance of the K-NN model. For class "M" data classification, the SVM model has a recall of 0.93, which is comparable to the performance of the Random Forest and Gradient Descent model.

Logistic Regression Algorithm Training Results

The Logistic regression algorithm train set and test set confusion matrices are shown in Figures 26 and 27, and the accuracy, precision, recall, and F1 scores are shown in Table 9. The training precision of the Logistic regression is 0.95804, the recall is 0.96478, the F1 score is 0.96140, and the accuracy is 0.95164. The classification performance parameters of the Logistic algorithm in the test set computed from Figure 27 are obtained as shown in Table 9. From Table 9, it can be seen that the Logistic model obtained the same classification performance as the SVM and K-NN model with a precision of 1.0 and an F1 score of 0.97 for class "B" data, achieving the best performance in class "B" data classification. The recall of the Logistic model in class "M" data is better than that of the decision tree model (recall 0.90), K-NN model (recall 0.88), gradient descent model (recall 0.93), random forest model (recall 0.95), and SVM model (recall 0.93). Logistic algorithm obtained the best classification performance compared to decision tree algorithm, SGD algorithm, K-NN Algorithm, random forest algorithm, and SVM algorithm, so we used logistic algorithm as a base classifier to train AdaBoost algorithm.

AdaBoost-Logistic Algorithm Training Results

The AdaBoost-Logistic algorithm training set and test set confusion matrices are shown in Figures 28 and 29, and the accuracy, precision, recall, and F1 scores are shown in Table 10. Calculated from Figure 28, the precision of Ada-Logistic model in the training set is 0.95390, the recall is 0.94718, the F1 score is 0.95053, and the accuracy is 0.93846; and the

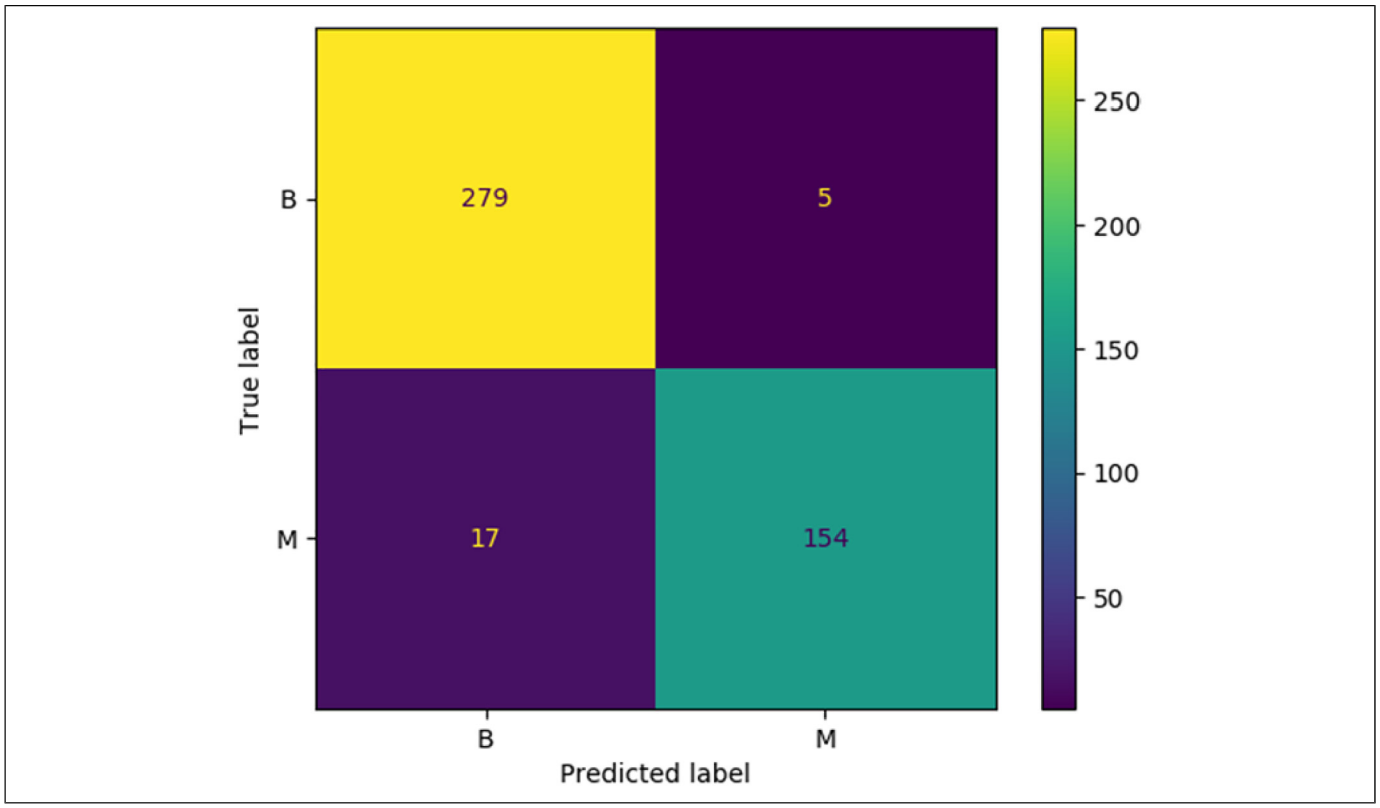

Figure 24. Support vector machine (SVM) model confusion matrix (train set).

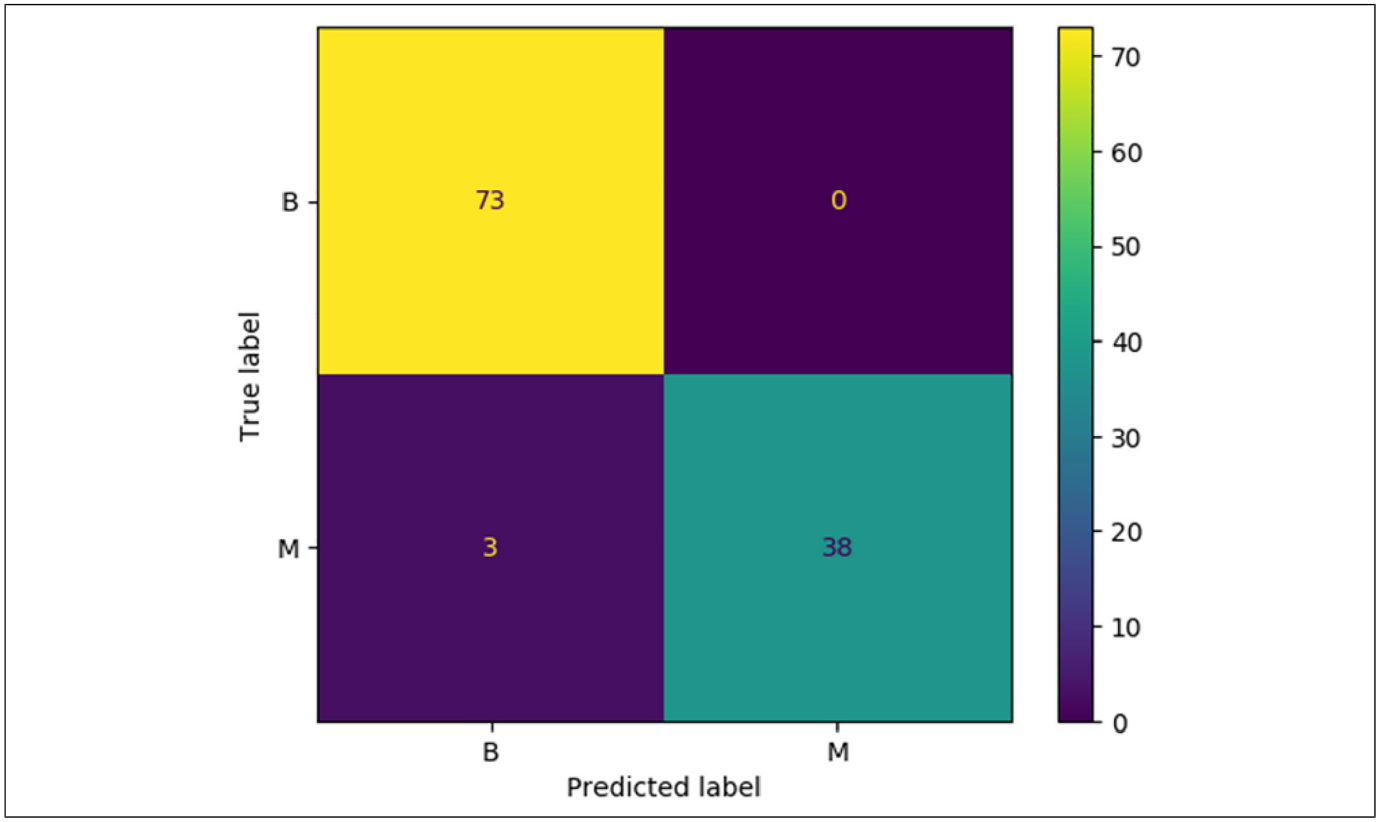

Figure 25. Support vector machine (SVM) model confusion matrix (test set).

precision, recall, f1 score, and accuracy of the AdaLogistic model in the test set calculated from Figure 29 are shown in Table 10. From Table 10, we can see that the AdaBoost-Logistic model achieves the best performance in class "B" data classification with a precision of 0.99, a recall of 1.0 and an F1 score of 0.99. Similarly, the AdaBoost-Logistic algorithm was tested on class "M" data categorization with a precision of 1.00, a recall of 0.98 and an F1 score of 0.99, also achieving superior performance on class "M" data. Notably, the average precision, recall, and F1 score all reached 0.99, the best performance among the 7 machine learning models.

Discussion

Algorithm Performance Comparison

Based on the model training results in the "Results" section, we summarize the training results to graphically compare the performance of different models; Table 11 shows the summarized results, and Figure 30 shows the visualization of the summarized results.

According to Table 11 and Figure 30, the training phase of the AdaBoost-Logistic regression model indicates a training error rate of 0.06153, a testing error rate of 0.00877, a precision of 0.98648, a recall rate of 1.00000, an F1 score of 0.99319, and an accuracy of 0.99122. These results highlight the superiority of this model over the other 6 models in the breast cancer dataset. Conversely, the decision tree algorithm exhibits a training error rate of 0.03076, a testing error rate of 0.07894, a precision of 0.94444, a recall rate of 0.93150, an F1 score of 0.93793, and an accuracy of 0.92105. Compared to the other 6 models, its performance on the breast cancer datasets is relatively weak.

Table 8. Classification Results of SVM Algorithm on the Test set.

	Precision	Recall	F1-score	Support
B	0.96	1.00	0.98	73
M	1.00	0.93	0.96	41
Accuracy			0.97	114
Macro avg	0.98	0.96	0.97	114
Weighted avg	0.97	0.97	0.97	114

Receiver Operating Characteristic Curve and Precision-Recall Curve of the Model

The receiver operating characteristic (ROC) curve, originally developed for military applications, has gained significant popularity in the medical and machine learning fields. It serves as a valuable tool for evaluating the classification performance of machine learning models in binary classification tasks. The area under the curve (AUC), a commonly used metric, represents the area beneath the ROC curve and can be employed to predict precision in binary classification models.[32] The

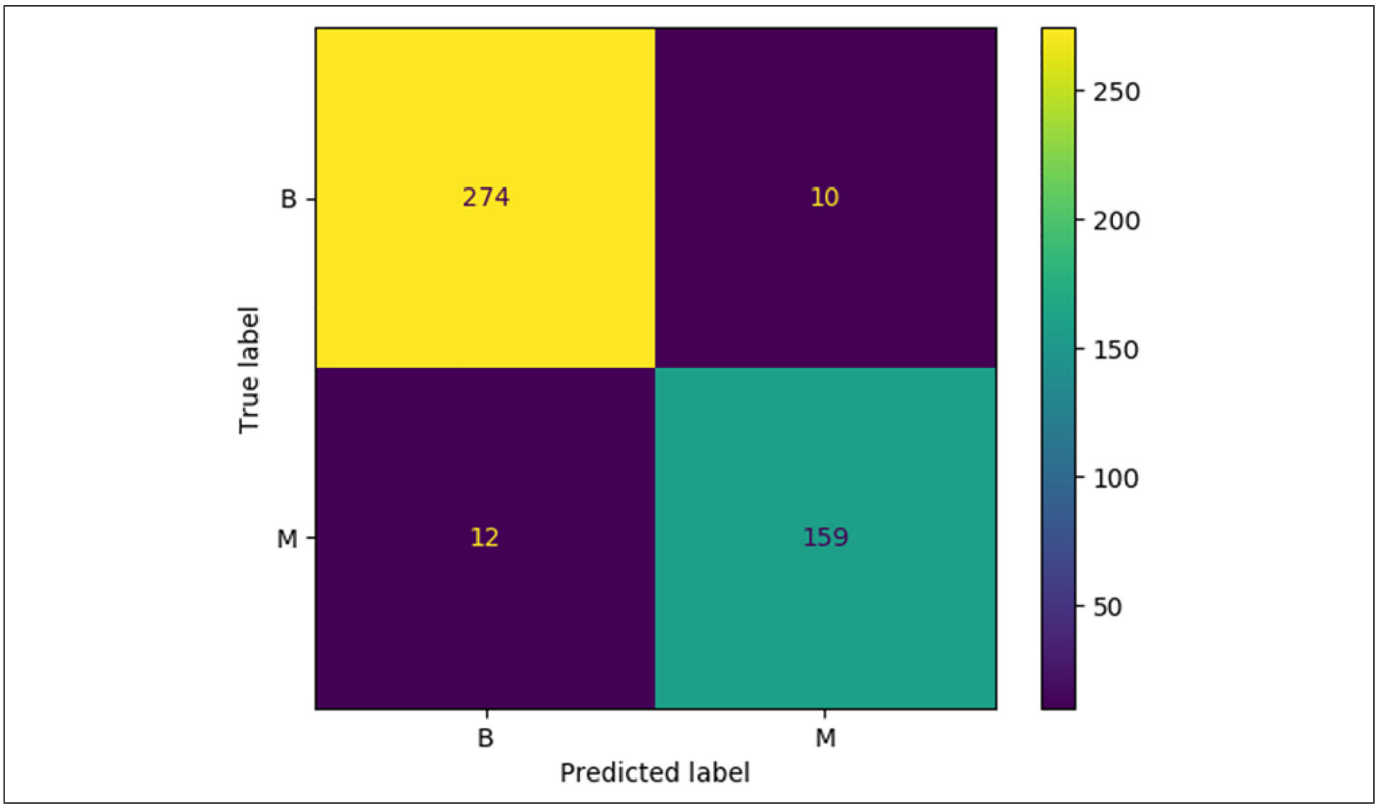

Figure 26. Confusion matrix of logistic model (train set).

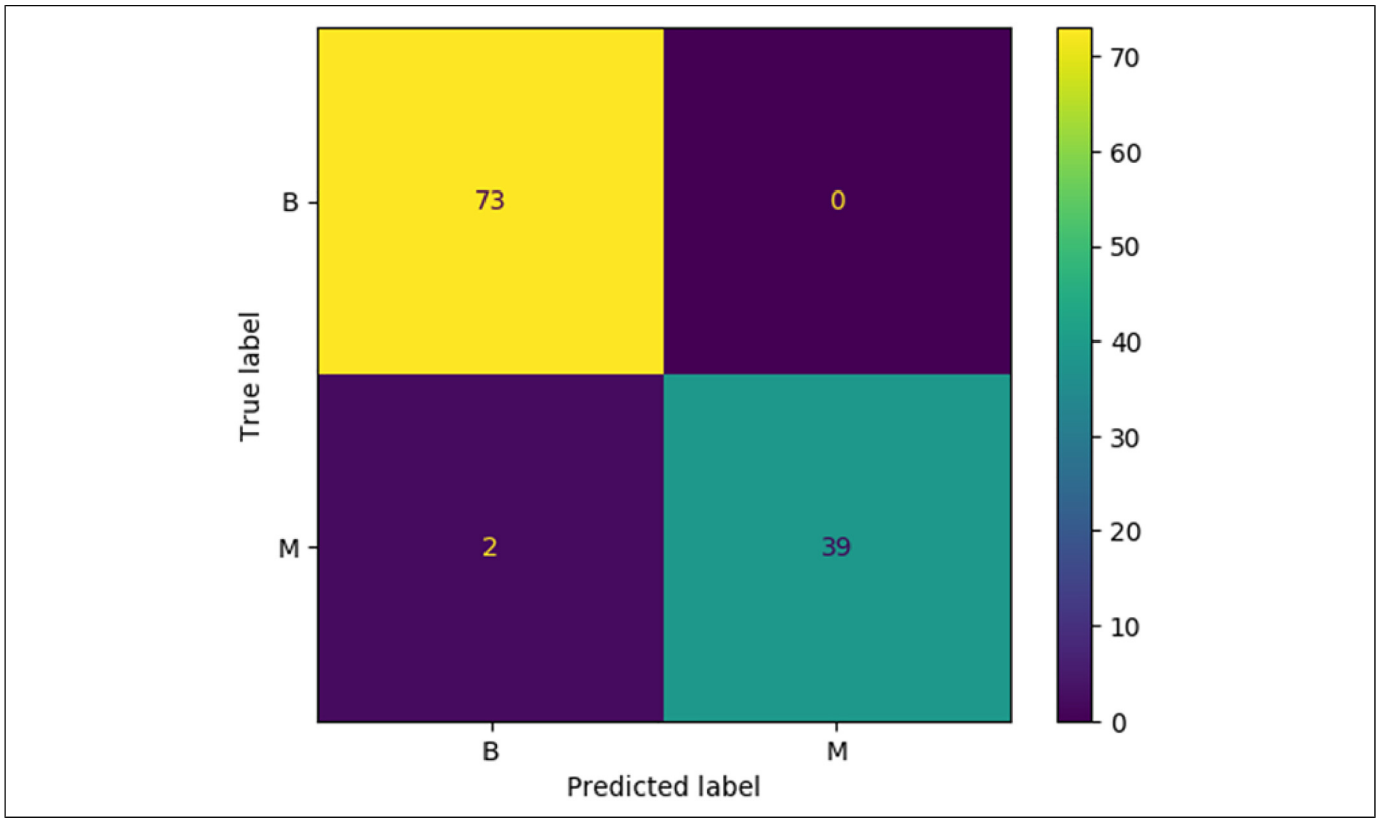

Figure 27. Confusion matrix of logistic model (test set).

Table 9. Classification Results of Logistic Test set.

	Precision	Recall	F1-score	Support
B	0.97	1.00	0.99	73
M	1.00	0.95	0.97	41
Accuracy			0.98	114
Macro avg	0.99	0.98	0.98	114
Weighted avg	0.98	0.98	0.98	114

AUC value ranges from 0 to 1, with a higher value indicating a higher correct rate for the model. On the other hand, the precision-recall (PR) curve visualizes the accuracy and recall of a binary classification model under varying thresholds, this allows developers to observe changes in accuracy and recall, aiding in the selection of the optimal threshold for training the best classification model.[33] Therefore, a higher accuracy under the threshold indicates a more effective classification performance of the model. Figures 31 and 32 present the PR curves and ROC curves for each mode

The PR curve reveals several significant findings. Firstly, the AdaBoost-Logistic regression model achieves an accuracy of 0.991228, surpassing the accuracy of the other 6 models in the breast cancer dataset. On the other hand, the decision tree model exhibits a lower accuracy of 0.921053 during training, indicating relatively poorer classification accuracy. Moreover, comparing the AUC values from the ROC curves of different models provides additional insights. The AdaBoost-Logistic algorithm, SGD algorithm, and SVM all demonstrate an AUC of 0.989642, indicating an identical correctness rate when applied to the Wisconsin Breast Cancer dataset. In contrast, the decision tree algorithm exhibits a substantially lower AUC of 0.900936 compared to the other 6 models, indicating a poorer correctness rate for the Wisconsin Breast Cancer dataset.

The AdaBoost-Logistic regression algorithm emerges as the highest-performing model among the 7 models, demonstrating superior precision, recall, F1 score, accuracy, training error, testing error, and AUC metrics. To further assess its performance, we utilized the test dataset to generate predictions using the integrated AdaBoost-Logistic regression model. These predictions were then visually represented in both 2D and 3D plots alongside the actual values Figures 33 and 34.

The data visualization results illustrate that integrating the Logistic regression algorithm with the AdaBoost algorithm produces superior performance on the test set. Additionally, the model predictions align well with the test dataset.

Interpretable AI

Regarding data interpretability, we investigate 3 key aspects: the distribution of feature data, the correlation between feature data and target data, and the distribution difference

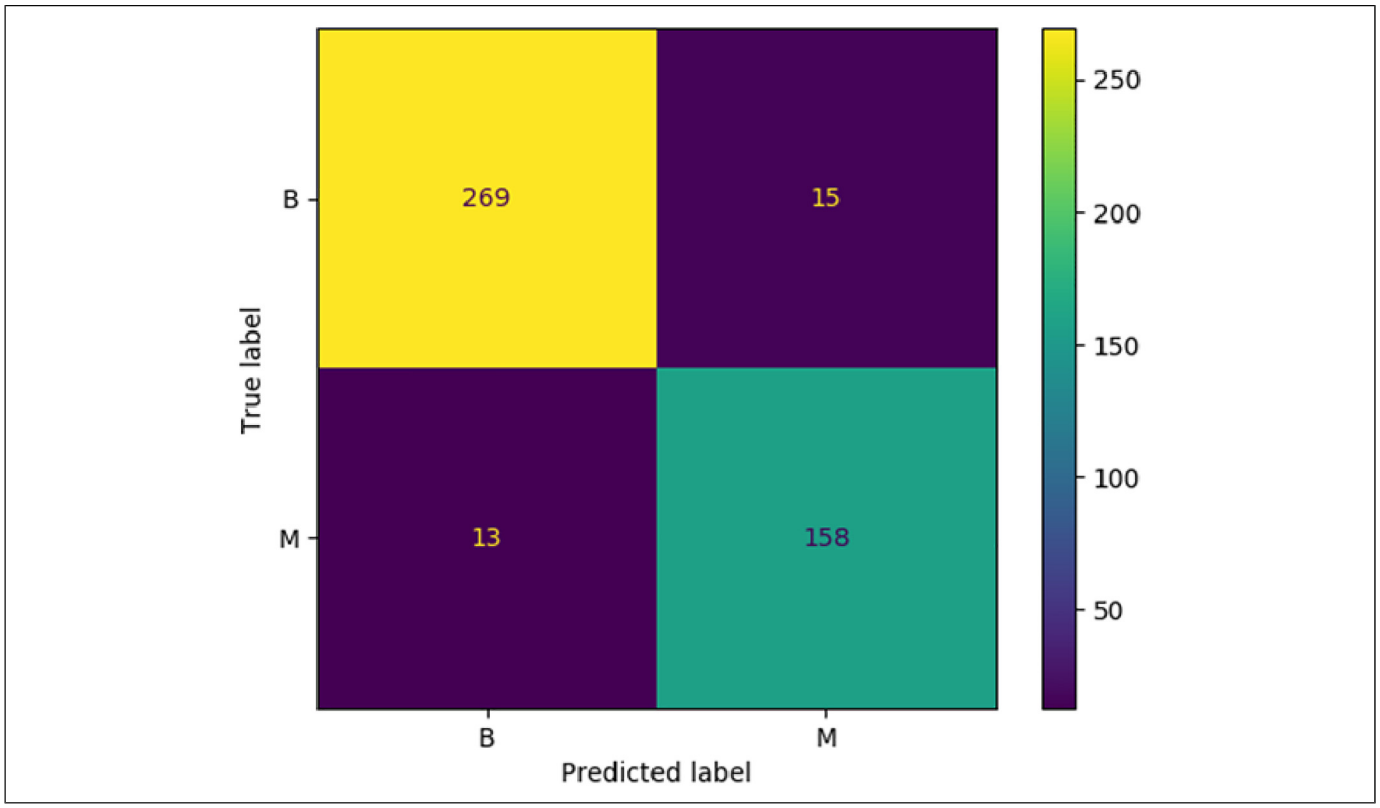

Figure 28. Adaboost_logistic confusion matrix (training set).

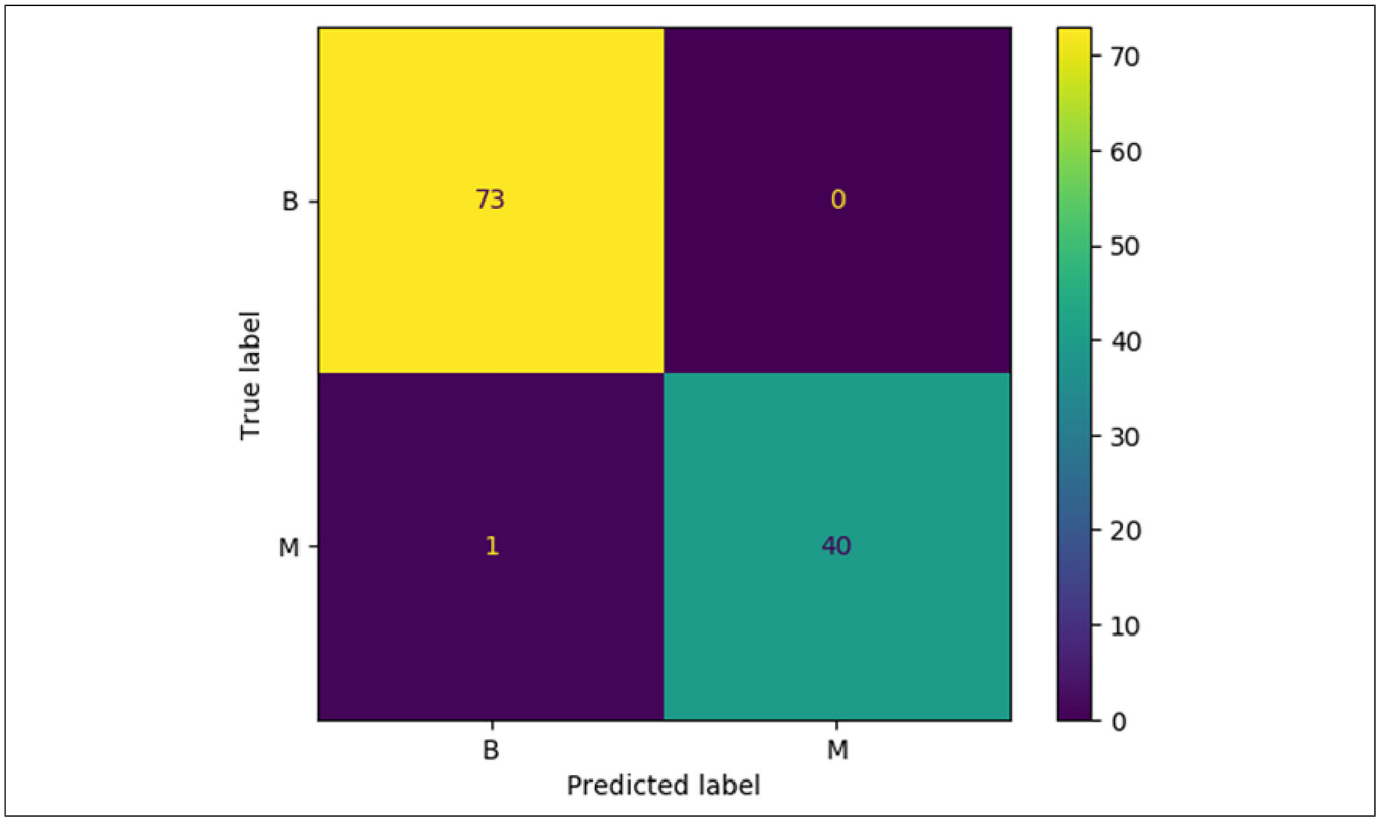

Figure 29. Adaboost_logistic confusion matrix (test set).

between feature data. As previously outlined, the statistical results demonstrate that the feature data do not conform to normal distribution ($P < .01$) per the normality test of feature data by using the scipy library Normaltest toolkit within a Python environment. The analysis of the Spearman correlation coefficient demonstrates a strong correlation between the screened feature subset data and the target data, with all correlation coefficients above 0.7. Additionally, the Wilcoxon rank sum test shows that the distribution of feature subset data is different across various breast cancer categories, with a P value less than .01. The aforementioned findings suggest a strong correlation between the feature subset data and the target data, while also highlighting noticeable differences in the distribution of feature subset data across different breast cancer categories. High-quality feature subset data thus facilitate the algorithm in learning patterns within the data.

Table 10. Classification Results of AdaBoost-Logistic Test set.

	Precision	Recall	F1-score	Support
B	0.99	1.00	0.99	73
M	1.00	0.98	0.99	41
Accuracy			0.99	114
Macro avg	0.99	0.99	0.99	114
Weighted avg	0.99	0.99	0.99	114

Comparison with Other Methods

To demonstrate the exceptional precision of our AdaBoost-Logistic algorithm for breast cancer detection, we conducted a comparative analysis with other contemporary techniques. The ensuing results have been presented in the table below.

As demonstrated in Table 12, our proposed AdaBoost-Logistic regression algorithm outperforms other methods in

Table 11. Model Training Results.

Model name	Training error	Test error	Precision	Recall	F1 score	Accuracy
Decision tree	0.03076	0.07894	0.94444	0.93150	0.93793	0.92105
Random forest	0.00219	0.02631	0.97297	0.98630	0.97959	0.97368
K-nearest neighbors	0.03956	0.04385	0.93589	1.00000	0.96688	0.95614
Logistic regression	0.04835	0.01754	0.97333	1.00000	0.98648	0.98245
Stochastic gradient descent	0.04175	0.02631	0.96052	1.00000	0.97986	0.97368
Support vector machines	0.04835	0.02631	0.96052	1.00000	0.97986	0.97368
AdaBoost-Logistic	0.06153	0.00877	0.98648	1.00000	0.99319	0.99122

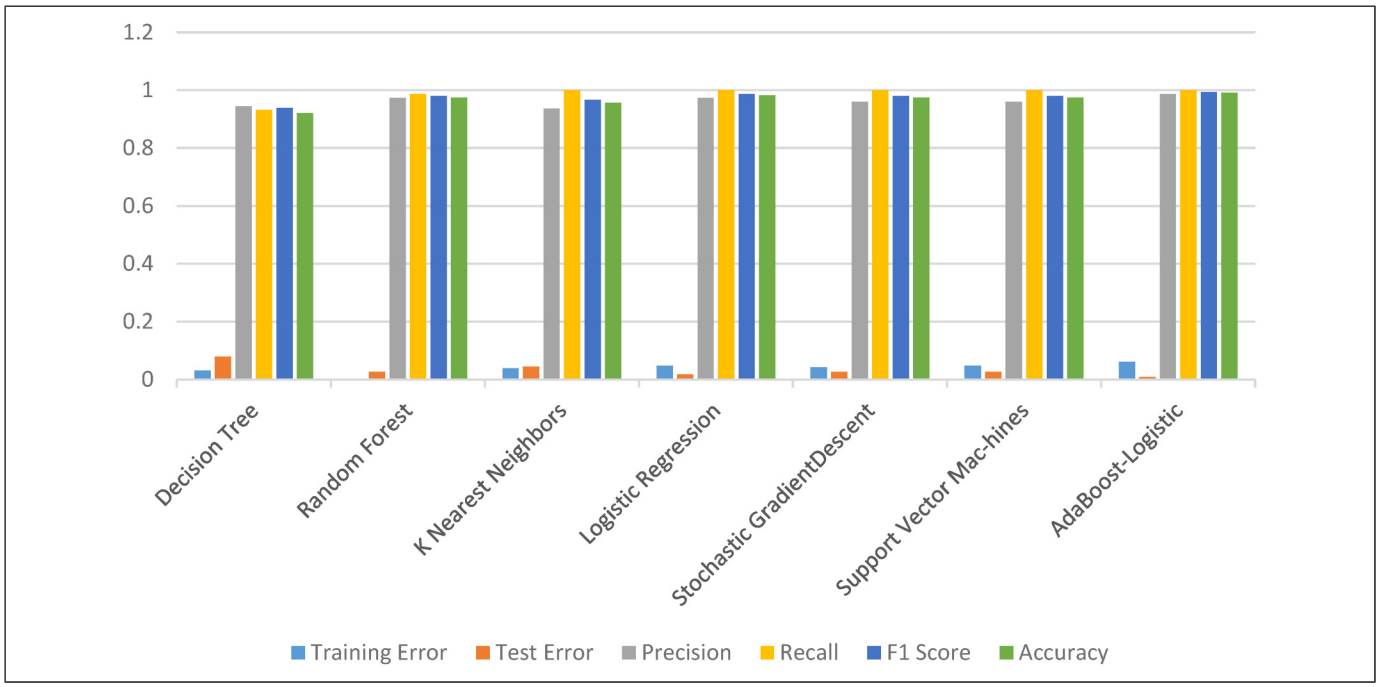

Figure 30. Bar graph of model training results.

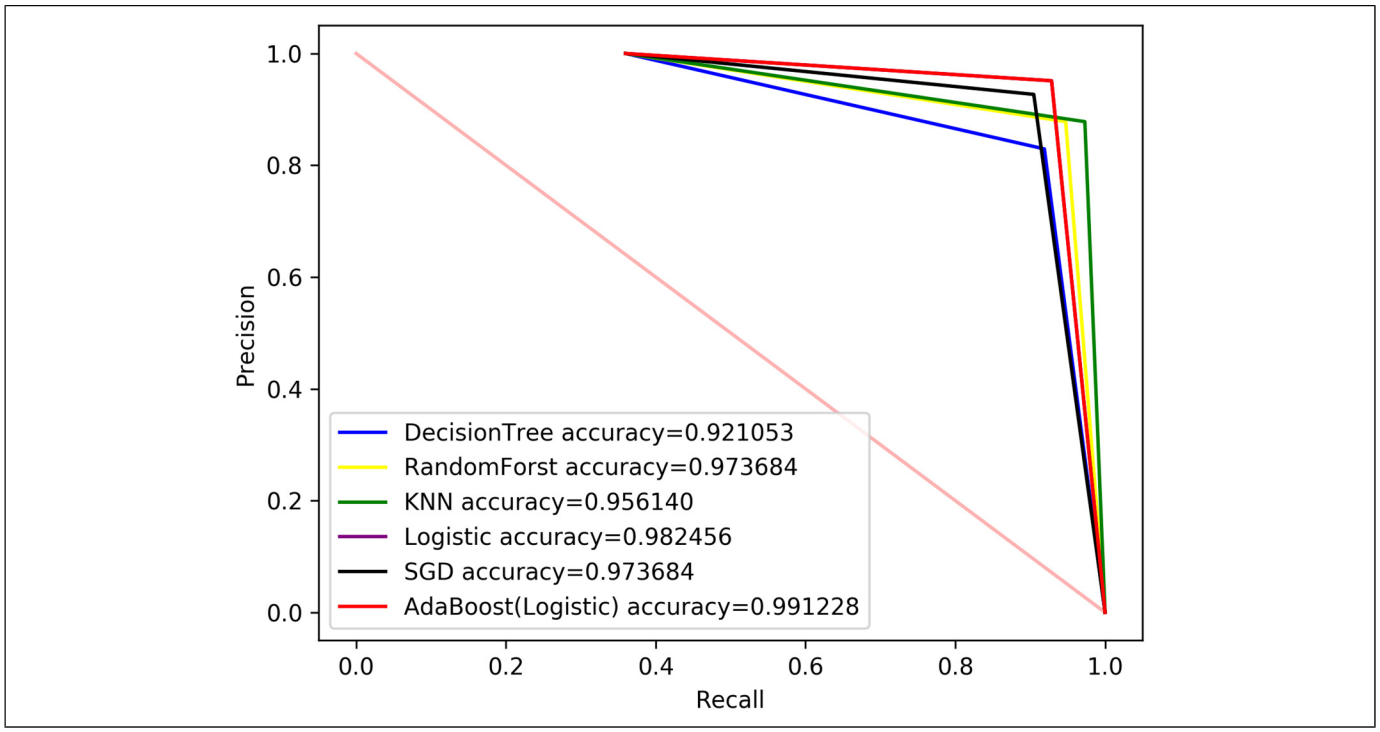

Figure 31. Model precision-recall (PR) curve.

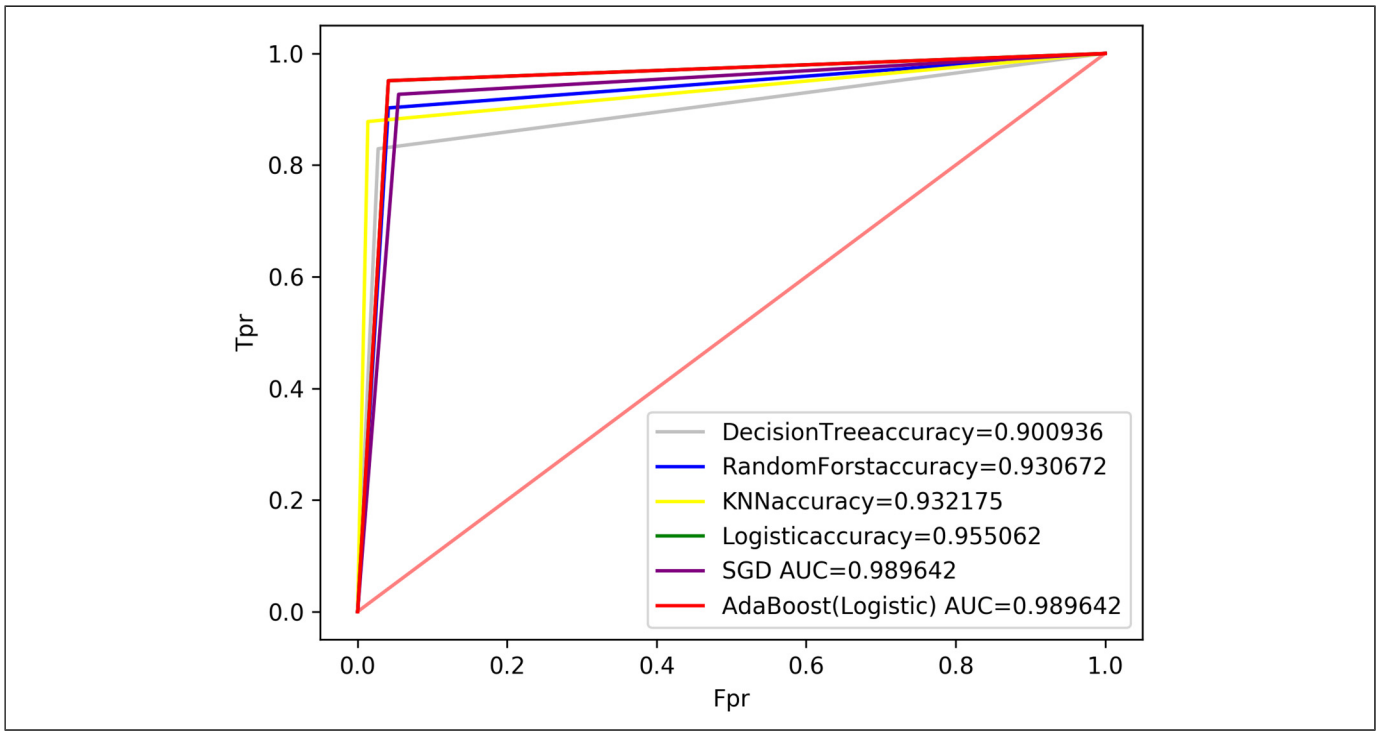

Figure 32. Model receiver operating characteristic (ROC) curve.

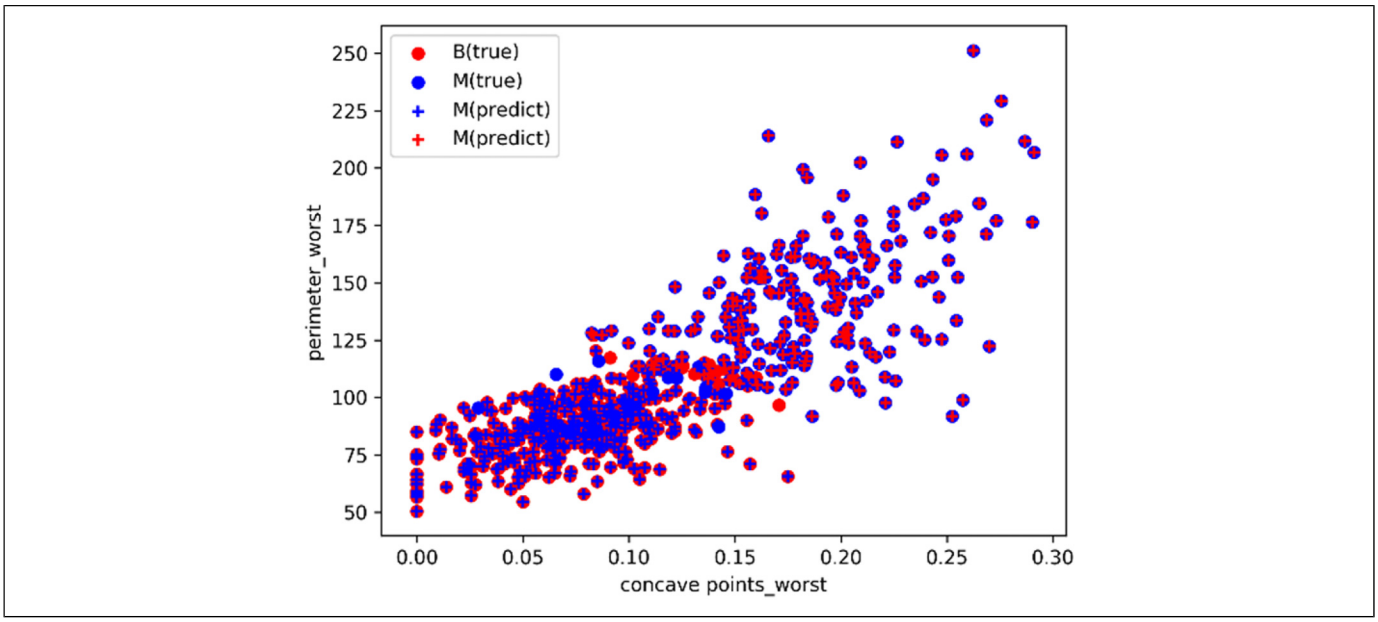

Figure 33. Two-dimensional scatter plot of model predictions.

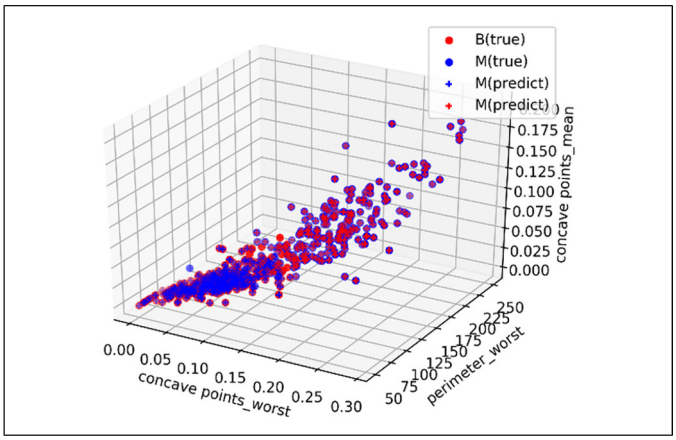

Figure 34. Three-dimensional scatterplot of model predictions.

terms of accuracy, thus making it more effective in the diagnosis of both benign and malignant breast cancers.

Conclusions

In this research, statistical methods were employed to assess the normality of the data in the Wisconsin breast cancer dataset. Spearman correlation analysis was then utilized to compute the correlation coefficients between the labeled data and the feature data. Subsequently, the feature data with high correlation with the labeled data was filtered to form a feature subset. Finally, the distributions of the feature subset data were examined for the various types of breast cancers using the Wilcoxon rank-sum test. The accuracy of 7 machine learning models was predictively trained and assessed using feature

Table 12. Comparison Table with the Latest Methodology.

Author name	Reference	Year	Model/method	Best observed accuracy
Hazra et al	34	2016	Support Vector Machine (SVM: using 19 features)	94.423%
Osman A. H. et al	35	2017	SVM	95.23%
Wang et al	36	2018	SVM-based ensemble learning	96.67%
Abdar et al	37	2018	Nested Ensemble 2-MetaClassifier (K = 5)	97.01%
Mushtaq et al	38	2019	KNN with multiple distances (Correlation $K = 2$)	91.00%
Rajaguru & Chakravarthy	39	2019	KNN Euclidean distance	95.61%
Durgalakshmi & Vijayakumar	40	2019	SVM	73%
Khan et al	41	2020	SVM	97.06%
Al-Azzam & Shatnawi	42	2021	LR with area under curve	96%
Abdur Rasool et al	43	2022	Polynomial SVM	99.03%
The model we propose		2023	**AdaBoost-Logistic**	**99.12%**

*Bolded text indicates what we have done.

subset data. Evaluation findings indicate that the AdaBoost-Logistic algorithm has excellent performance in diagnosing breast cancer, with an accuracy of up to 0.991228. This technique is expected to play a crucial role in breast cancer screening and assisted diagnosis, substantially decreasing patient waiting periods and enhancing healthcare organizations' efficiency.

The algorithm has practical significance for clinicians as it can be deployed on medical devices, reducing waiting time for breast cancer diagnosis and enabling them to quickly obtain predictions. These predictions can guide clinicians in developing appropriate treatment plans. Additionally, for patients, the algorithm's predictions can minimize wait times, prevent deterioration, and facilitate timely access to treatment. Furthermore, the integration of online learning technologies allows the model to continuously learn from new data, enhancing its accuracy and generalizability based on the information provided by patients. Lastly, the features extracted through the high relevance filtering approach demonstrate excellent performance in the machine learning model, offering valuable insights for further investigation of their role in breast cancer pathogenesis, disease progression, and diagnosis.

The proposed method offers various benefits. It starts by examining the correlation between input feature data and output variables through high-correlation filtering. Moreover, the Wilcoxon rank sum test investigates the distributional differences between different characteristics of breast cancer. A key advantage of this approach is the reduction of feature size, leading to the faster training of models. Secondly, the methods of high correlation filtering and Wilcoxon rank sum testing can be applied to various datasets, independent of specific machine learning algorithms. Furthermore, the set of features produced by these methods is highly interpretable and can be used in diverse datasets. Finally, the AdaBoost integration algorithm was used in combination with Logistic regression to develop a reliable classifier, which improves the overall stability and generalization performance of the machine learning model.

The AdaBoost logistic regression model, which we propose in this study, has certain limitations. Initially, the model was trained on a dataset with only 569 instances, which limits its generalization performance when applied to large-scale clinical data with limited training samples. Moreover, although the interpretability of our work on feature data is superior, further investigation is required to enhance the interpretability of the AdaBoost-Logistic algorithm. Furthermore, the efficacy and consistency of our suggested AdaBoost-Logistic regression framework in practical implementation scenarios have not undergone assessment. Online learning, a framework for designing predictive models that process data only once, offers theoretical performance guarantees regardless of statistical assumptions of the data source, alongside high computational efficiency.[44] Therefore, in future work, we plan to deploy the models on Baidu AI Studio and implement online learning techniques. This approach will enable continuous learning from new breast cancer clinical data, thereby improving the accuracy and generalization performance of the model.

Confirm

All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are openly available in [UCI Machine Learning Repository] at [https://doi.org/10.24432/C5DW2B], reference number [10.24432/C5DW2B]. Aethics statement (including the Committee Approval Number) for animal and human studies 2023 Lunar Review No. (38).

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Guizhou Medical University High-level Talent Start-up Fund, grant number XBH J[2021]022.

ORCID iD

Sheng Zhou https://orcid.org/0000-0001-8979-0157

References

1. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA Cancer J Clin.* 2023;73(1):17-48.
2. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer J Clin.* 2021;71(3):209-249.
3. Giaquinto AN, Sung H, Miller K, et al. Breast cancer statistics, 2022. *CA Cancer J Clin.* 2022;72(6):524-541.
4. Zheng R, Zhang S, Zeng H, et al. Cancer incidence and mortality in China, 2016. *J Natl Cancer Center*. 2022;2(1):1-9.
5. Adami HO, Persson I, Ekbom A, et al. The aetiology and pathogenesis of human breast cancer. *Mutat Res/Fundam Mol Mech Mutagenesis*. 1995;333(1–2):29-35.
6. He Z, Chen Z, Tan M, et al. A review on methods for diagnosis of breast cancer cells and tissues. *Cell Prolif.* 2020;53(7):e12822.
7. Haque MN, Tazin T, Khan MM, et al. Predicting characteristics associated with breast cancer survival using multiple machine learning approaches. *Comput Math Methods Med.* 2022;1(1):1-12.
8. Zhang Y-D, Wang S-H, Liu G, Yang J. Computer-aided diagnosis of abnormal breasts in mammogram images by weighted-type fractional Fourier transform. *Advances in Mechanical Engineering*. 2016;8(2):1687814016634243.
9. Wu J, Hicks C. Breast cancer type classification using machine learning. *J Pers Med.* 2021;11(2):61.
10. Zhang Y, Wu X, Lu S, Wang H, Phillips P, Wang S. Smart detection on abnormal breasts in digital mammography based on contrast-limited adaptive histogram equalization and chaotic adaptive real-coded biogeography-based optimization. *SIMULATION*. 2016;92(9):873-885.

11. Monirujjaman Khan M, Islam S, Sarkar S, et al. Machine learning based comparative analysis for breast cancer prediction. *J Healthc Eng.* 2022;1(1):4365855.
12. Kumar M, Singhal S, Shekhar S, et al. Optimized stacking ensemble learning model for breast cancer detection and classification using machine learning. *Sustainability*. 2022;14(21):13998.
13. Aamir S, Rahim A, Aamir Z, et al. Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques. *Comput Math Meth Med*. 2022;1(1):1-13.
14. Wang S, Rao RV, Chen P, et al. Abnormal breast detection in mammogram images by feed-forward neural network trained by jaya algorithm. *Fundam Inform.* 2017;151(1–4):191-211.
15. Ly A, Marsman M, Wagenmakers E-J. Analytic posteriors for Pearson's correlation coefficient. *Stat Neerl.* 2018;72(1):4-13.
16. Sun Y, Ding S, Zhang Z, et al. An improved grid search algorithm to optimize SVR for prediction. *Soft Comput.* 2021;25(25):5633-5644.
17. Ramakrishna MT, Venkatesan VK, Izonin I, Havryliuk M, Bhat CR. Homogeneous Adaboost Ensemble Machine Learning Algorithms with Reduced Entropy on Balanced Data. *Entropy* . 2023;25(2):245.
18. 飞桨AI Studio-人工智能学习实训社区, Available online: https://aistudio.baidu.com/aistudio/index
19. Hunt EB, Marin J, Stone PJ. *Experiments in induction*. psycnet.apa.org; 1966.
20. Breiman L. Random forests. *Mach Learn.* 2001;45(45):5-32.
21. Li XL. Preconditioned stochastic gradient descent. *IEEE Trans On Neural Netw Learn Syst*. 2017;29(5):1454-1466.
22. Gong C, Su Z, Wang P, et al. Evidential instance selection for K-nearest neighbor classification of big data. *Int J Approx Reason.* 2021;138(138):123-144.
23. Wang X, Huang F, Cheng Y. Computational performance optimization of support vector machine based on support vectors. *Neurocomputing*. 2016;211(211):66-71.
24. Cramer JS. The origins of logistic regression (December 2002)[R]. Tinbergen Institute Working Paper.
25. Dinakaran S, Ranjit Jeba Thangaiah P. Ensemble method of effective adaboost algorithm for decision tree classifiers. *Int J Artif Intell Tools.* 2017;26(3):1750007.
26. https://scikit-learn.org, Available online:https://scikit-learn.org/stable/index.html
27. Vellido A. Societal issues concerning the application of artificial intelligence in medicine. *Kidney Dis*. 2019;5(1):11-17.
28. Kundu S. AI In medicine must be explainable. *Nat Med*. 2021; 27(8):1328.
29. Sandeep R. Explainability and artificial intelligence in medicine. *Lancet Dig Health*. 2022;4(4):e214-e215.
30. MOLNAR C.Interpretable machine learning. 2020.
31. Yoon CH, Torrance R, Scheinerman N. Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? *J Med Ethics*. 2022;48(9):581-585.
32. Omar L, Ivrissimtzis I. Using theoretical ROC curves for analysing machine learning binary classifiers. *Pattern Recognit Lett.* 2019;128(128):447-451.
33. Hughes-Oliver JM. Population and empirical PR curves for assessment of ranking algorithms. arXiv preprint arXiv:1810.08635, 2018.
34. Hazra A, Mandal SK, Gupta A. Study and analysis of breast cancer cell detection using naïve Bayes, SVM and ensemble algorithms. *Int J Comput Appl*. 2016;145(2):39-45.
35. Osman AH. An enhanced breast cancer diagnosis scheme based on two-step-SVM technique. *Int J Adv Comput Sci Appl*. 2017;8(4):158-165.
36. Wang H, Zheng B, Yoon SW, Ko HS. A support vector machine-based ensemble algorithm for breast cancer diagnosis. *Eur J Oper Res*. 2018;267(2):687-699.
37. Abdar M, Yen NY, Hung JCS. Improving the diagnosis of liver disease using multilayer perceptron neural network and boosted decision trees. *J Med Biol Eng*. 2018;38(38):953-965.
38. Mushtaq Z, Yaqub A, Sani S, Khalid A. Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets. *J Chin Inst Eng*. 2020;43(1):80-92.
39. Rajaguru H. Analysis of decision tree and K-nearest neighbor algorithm in the classification of breast cancer. *Asian Pac J Cancer Prev APJCP*. 2019;20(12):3777.
40. Durgalakshmi B, Vijayakumar V. Feature selection and classification using support vector machine and decision tree. *Comput Intell*. 2020;36(4):1480-1492.
41. Khan F, Khan MA, Abbas S, et al. Cloud-based breast cancer prediction empowered with soft computing approaches. *J Healthc Eng*. 2020;1(1):8017496.
42. Al-Azzam N, Shatnawi I. Comparing supervised and semi-supervised machine learning models on diagnosing breast cancer. *Ann Med Surg*. 2021;62(2021):53-64.
43. Abdur R, Chayut B, Luo T, et al. Improved machine learning-based predictive models for breast cancer diagnosis. *Int J Environ Res Public Health.* 2022;19(6):3211.
44. Cesa-Bianchi N, Orabona F. Online learning algorithms. *Annu Rev Stat Appl.* 2021;8(8):165-190.

Annex 1. Results of the Eigen normality Test.

Feature	Normal test result	*P*-value
radius_mean	73.17938185797058	$1.286172249506454 \times 10^{-16}$
texture_mean	42.962593202972435	$4.685882796961145 \times 10^{-10}$
perimeter_mean	80.33371758248074	$3.595463394731772 \times 10^{-18}$
area_mean	191.6778773687591	$2.3860403182400484 \times 10^{-42}$
smoothness_mean	28.365736423550842	$6.925619033399656 \times 10^{-07}$
compactness_mean	113.11820358995082	$2.7333433490329536 \times 10^{-25}$
concavity_mean	141.50770215049243	$1.8706515779247436 \times 10^{-31}$
concave points_mean	101.08583583151744	$1.120700717286823 \times 10^{-22}$
symmetry_mean	58.93994598508814	$1.5898397363101684 \times 10^{-13}$
fractal_dimension_mean	146.6528543983162	$1.4280298636963606 \times 10^{-32}$
radius_se	422.6146599814736	$1.6997795475056234 \times 10^{-92}$
texture_se	211.4900711530144	$1.1899112222853113 \times 10^{-46}$
perimeter_se	464.4216817724931	$1.4194273365493062 \times 10^{-101}$
area_se	658.800107474458	$8.777570438431168 \times 10^{-144}$
smoothness_se	318.8677914591451	$5.7377134655009676 \times 10^{-70}$
compactness_se	235.2825097603231	$8.110584649898849 \times 10^{-52}$
concavity_se	636.7669018057265	$5.343479450682037 \times 10^{-139}$
concave points_se	187.85195713200977	$1.6160909617897005 \times 10^{-41}$
symmetry_se	289.7292894347901	$1.2192258903277784 \times 10^{-63}$
fractal_dimension_se	515.1754842858727	$1.3522748824773011 \times 10^{-112}$
radius_worst	91.75004501524197	$1.1932484193696656 \times 10^{-20}$
texture_worst	22.833847273896634	$1.1007610762075063 \times 10^{-05}$
perimeter_worst	96.55541166709088	$1.0795897623829055 \times 10^{-21}$
area_worst	222.9650485535912	$3.8349073670520574 \times 10^{-49}$
smoothness_worst	20.183781887596698	$4.141402593544656 \times 10^{-05}$
compactness_worst	165.16516096105937	$1.3640836264499759 \times 10^{-36}$
concavity_worst	108.26393236265899	$3.095891935387191 \times 10^{-24}$
concave points_worst	33.7274729514053	$4.744301759090373 \times 10^{-08}$
symmetry_worst	179.09714551333752	$1.2869124116718329 \times 10^{-39}$
fractal_dimension_worst	212.0842519784993	$8.840764735693279 \times 10^{-47}$