Kandhasamy *et al*. *J Transl Genet Genom*. 2025;9:11-34
DOI: 10.20517/jtgg.2024.82

Journal of Translational Genetics and Genomics

Original Article

Open Access

Machine learning framework for breast cancer detection with feature selection with L2 ridge regularization: insights from multiple datasets

Premalatha Kandhasamy[1], **Duraisamy Prabha Devi**[1], **Sivakumar Kandhasamy**[2]

[1]Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode 638401, Tamil Nadu, India.
[2]Department of Biomedical Engineering, Karpaga Vinayaga College of Engineering and Technology, Palayanoor 603308, Tamil Nadu, India.

Correspondence to: Dr. Premalatha Kandhasamy, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode 638401, Tamil Nadu, India. E-mail: kpl_barath@yahoo.co.in

How to cite this article: Kandhasamy P, Devi DP, Kandhasamy S. Machine learning framework for breast cancer detection with feature selection with L2 ridge regularization: insights from multiple datasets. *J Transl Genet Genom.* 2025;9:11-34. https://dx.doi.org/10.20517/jtgg.2024.82

Received: 18 Oct 2024 **First Decision:** 14 Dec 2024 **Revised:** 17 Dec 2024 **Accepted:** 10 Jan 2025 **Published:** 22 Jan 2025

Academic Editor: Laura Ottini **Copy Editor:** Fangling Lan **Production Editor:** Fangling Lan

Abstract

Aim: This study aims to investigate and apply effective machine learning techniques for the early detection and precise diagnosis of breast cancer. The analysis is conducted using various breast cancer datasets, including Breast Cancer Wisconsin, Breast Cancer Diagnosis, NKI Breast Cancer, and SEER Breast Cancer datasets. The primary focus is on identifying key features and utilizing preprocessing methods to enhance classification accuracy.

Methods: The datasets undergo several preprocessing steps, such as label encoding for categorical variables, linear regression for handling missing values, and Robust scaler normalization for data standardization. To address class imbalance, Tomek Link SMOTE is employed to improve dataset representation. Significant features are selected through L2 Ridge regularization, helping to pinpoint the most important predictors of breast cancer. A range of machine learning models, including decision tree, random forest, support vector machine (SVM), neural network, K-nearest neighbor, naïve bayes, extreme gradient boost (XGBoost), and AdaBoost, are applied for classification tasks. The performance of these models is assessed using metrics such as accuracy, precision, recall, F1-score, and the Kappa statistic. Additionally, the models' effectiveness is further evaluated using the receiver operating characteristic curve and precision-recall curve.

Results: The XGBoost model achieved the best performance on both the breast cancer Wisconsin and diagnosis datasets. The SVM model reached 100% accuracy on the NKI breast cancer dataset, while the random forest model performed optimally on the SEER breast cancer dataset. The feature selection process through L2 Ridge regularization was crucial in enhancing the performance of these models.

Conclusions: This work emphasizes the critical role of machine learning in improving breast cancer detection. By applying a combination of preprocessing techniques and classification models, the study successfully identifies significant features and boosts model performance. These findings contribute to the development of more accurate diagnostic tools, ultimately enhancing patient outcomes.

Keywords: Breast cancer, machine learning, feature selection, data preprocessing, Tomek Link SMOTE, L2 ridge regularization

INTRODUCTION

Breast cancer is a heterogeneous and multifactorial disease that typically begins in the milk ducts or lobules of the breast. Its development is driven by a combination of genetic alterations, hormonal imbalances, and environmental exposures, which lead to abnormal cell proliferation and the potential for malignant cells to spread to other organs. The capacity for metastasis makes early detection and intervention crucial for effective management. As one of the leading causes of death and disability globally, breast cancer places considerable pressure on healthcare systems. Despite progress in diagnostic methods and treatments, the disease's high prevalence and significant impact on patients' lives highlight the urgent need for novel approaches in both detection and therapy to improve patient outcomes and alleviate the burden on healthcare infrastructures.

Breast cancer is a significant global health concern, accounting for approximately 25% of all cancer cases among women and remaining one of the leading causes of cancer-related mortality worldwide. The increasing incidence of breast cancer necessitates improved screening methods, diagnostic techniques, and treatment options to enhance early detection and improve patient outcomes. Timely identification of breast cancer can substantially reduce mortality rates and increase the likelihood of successful treatment. The integration of machine learning (ML) techniques into breast cancer diagnosis presents a promising approach to address the challenges associated with traditional diagnostic methods. These techniques can analyze large and complex datasets, uncover patterns, and make predictions that may not be readily apparent through conventional statistical methods. Several publicly available breast cancer datasets, such as the Breast Cancer Wisconsin, Breast Cancer Diagnosis, NKI Breast Cancer, and SEER Breast Cancer Dataset, offer valuable resources for developing and validating ML models.

Effective preprocessing of these datasets is crucial for enhancing the performance of machine learning algorithms. This study employs various preprocessing techniques, including label encoding to convert categorical variables into numerical values, linear regression to handle missing data, and Robust scalar normalization for feature scaling. To mitigate class imbalance - a common issue in cancer datasets - Tomek Link SMOTE is applied, allowing for a more equitable representation of classes and improving model performance.

Feature selection is another vital aspect of this research, as identifying the most relevant features can lead to more accurate and interpretable models. L2 Ridge regularization is utilized to highlight the key predictors of breast cancer, providing insights that can inform clinical decision making.

In this study, we explore a range of machine learning models, including decision trees, random forests, support vector machines (SVM), Neural networks, K-nearest neighbors (KNN), Naïve bayes, extreme gradient boosting (XGBoost), and AdaBoost. The performance of these models is evaluated using several metrics, including accuracy, precision, recall, F1-score, and kappa constant, alongside the ROC curve and Precision-Recall curve for a comprehensive assessment of their effectiveness.

The findings from this research aim to contribute to the ongoing efforts in breast cancer identification, ultimately fostering advancements in diagnostic technologies and improving patient outcomes through the application of machine learning.

Breast cancer is one of the most prevalent and harmful tumors affecting women, often arising from a combination of lifestyle choices, environmental factors, and genetic predispositions. Studies indicate that 5%-10% of breast cancer cases are linked to hereditary genetic mutations associated with family history[1]. Despite advancements in diagnosis and treatment, breast cancer continues to pose a major health challenge, with approximately 30% of women in the USA affected annually. According to the 2024 Breast Cancer Statistics report, 1 in 8 women is likely to develop breast cancer during their lifetime, with about 61% of cases detected at localized stages and 70%-80% involving invasive ductal carcinoma. Triple-negative breast cancer remains common, accounting for 10%-15% of cases[2]. Symptoms include unexplained breast swelling, nipple discharge, and persistent discomfort. Non-invasive breast cancers, such as ductal carcinoma *in situ* (DCIS), involve abnormal cell growth within ducts, while invasive types like infiltrative ductal carcinoma spread into breast tissues and beyond[3].

Early detection of breast cancer is vital in reducing mortality rates and slowing the progression of carcinoma cells. Diagnostic methods such as biopsy, mammography, ultrasonography, and thermography are effective but often underutilized due to limited access and high costs in certain areas, leading to higher death rates in some communities[4]. Systematic analysis was introduced to classify cells as cancerous or non-cancerous, addressing these challenges.

In recent years, artificial intelligence has become increasingly important in the medical field, where accurate diagnosis is essential for effective treatment. Machine learning and deep learning techniques are critical for screening severe diseases such as breast cancer. Advances in molecular biotechnology and imaging have enabled reliable diagnostic methods, including deep learning with biomarkers from hematoxylin and eosin images to detect and localize tumor regions[5].

Another way to identify them is through medical imaging along with machine learning process. Medical imaging is instrumental in the detection of defects in various organs of the body, such as the lungs[6], brain[7], and stomach[8]. Through these images, it is significant to perform feature selection to optimize its performance and improve the diagnostic accuracy in mammograms[9]. Convolutional neural networks (CNN) have proven highly effective in analyzing medical images by filtering out noise, removing imaging artifacts, and enhancing low-contrast features.

Khan *et al.*, (2020) introduced a CNN-based method, CNNI-BCC, achieving 90.5% accuracy in breast cancer classification using data from 221 patients[10]. This deep learning model operates without human intervention to classify cancer types. Similarly, Al-Antari *et al.* (2018) developed a CAD system employing a deep belief network (DBN) to assist radiologists in diagnosing breast cancer from digital mammography images[11]. The system uses two ROI extraction techniques - small ROIs from detected masses and whole mass ROIs - extracting 347 features for classification with methods such as quadratic and linear discriminant analysis and neural networks. The DBN-based CAD system demonstrated superior accuracy, achieving 92.86% and 90.84% for the respective ROI approaches.

The research was carried out on the Wisconsin Breast Cancer Dataset, in which five machine learning algorithms were adopted: random forest algorithm (RF), SVM, logistic regression (LR), KNN, and C4.5 decision tree. The main objective of this research was to predict and diagnose breast cancer with respect to accuracy, confusion matrix, and precision. Among these five ML algorithms, SVM outperformed well and obtained an accuracy of 97.2%[12].

Bone marrow carcinomatosis is a rare complication of breast cancer. A case involved a woman in her seventies with stage IV estrogen receptor-positive invasive lobular carcinoma, whose hematologic abnormalities improved with letrozole treatment but later required a switch to palbociclib and fulvestrant due to disease progression[13].

Breast cancer, a leading cause of cancer mortality, has been linked to progranulin (PGRN) as a potential biomarker. Studies suggest PGRN's role in increased cancer risk, clinicopathological features, and drug resistance. Targeting PGRN and related pathways, such as sortilin (SORT1), may offer innovative strategies for early detection and treatment[14].

A medical IoT-based diagnostic system[15] was proposed to effectively identify malignant and benign breast cancer cases. This method uses ANN and CNN with optimized hyperparameters for classification using particle swarm optimization (PSO), where SVMs and multi-layer perceptron (MLPs) act as baseline classifiers for comparison. This achieved the highest accuracy of 98.5% using CNN and 99.2 % using ANN.

A study using the invasive ductal carcinoma (IDC) dataset evaluated various ML algorithms, including decision tree, random forest, and light gradient boosting (LGB). The LGB algorithm achieved the highest accuracy of 95%, with precision, recall, and F1 scores of 94.86%, 94.32%, and 94.57%, respectively. This method could assist healthcare providers in making better decisions, improving treatment, and enhancing outcomes for breast cancer patients[16].

Advancements in machine learning and deep learning have enabled early breast cancer diagnosis. A study using the SEER Database applied preprocessing techniques, handling missing values with random forest classifiers for categorical variables and random forest regressors for continuous variables. Significant features were selected using Variance Threshold and Principal Component Analysis. These features were then classified using decision tree (DT), Naïve Bayes, AdaBoost, gradient boosting classifier (GBC), and XGBoost, with metrics such as accuracy, recall, precision, F1 score, sensitivity, and specificity. Among all, the Decision Tree achieved the highest accuracy of 98%[17].

Although all these achievements toward breast cancer diagnosis resulted decently with good accuracy, they present potential limitations in terms of precision, recall and F1 score, kappa co-efficient, sensitivity, and specificity. This study addresses these challenges and analyzes various machine learning techniques by turning an imbalanced dataset into a balanced one using SMOTE technique. This helps to achieve superior performance in breast cancer detection.

Problem statement

Breast cancer continues to pose a significant threat to women's health globally, with rising incidence rates and associated mortality. Traditional diagnostic methods, including mammography and clinical examinations, often suffer from limitations such as false positives, false negatives, and a lack of comprehensive analysis of individual risk factors. These challenges underscore the urgent need for more accurate, efficient, and accessible diagnostic tools.

Despite the availability of extensive datasets related to breast cancer, the effective utilization of these data for early diagnosis remains a formidable challenge. Many existing approaches do not fully leverage machine learning techniques to analyze complex datasets, which can lead to suboptimal identification of breast cancer cases. Additionally, issues such as class imbalance, missing values, and the need for robust feature selection further complicate the development of effective predictive models.

This work aims to address these challenges by employing advanced machine learning methodologies and rigorous data preprocessing techniques to enhance the identification of breast cancer. Specifically, we focus on improving the accuracy and reliability of predictive models through the integration of multiple machine learning algorithms, comprehensive feature selection, and effective handling of data inconsistencies. By tackling these issues, the study seeks to contribute to the development of a more reliable and effective diagnostic framework that can facilitate early detection of breast cancer and ultimately improve patient outcomes

Research contributions

This work makes several significant contributions to the field of breast cancer identification, aiming to enhance diagnostic accuracy and promote the use of machine learning techniques in clinical practice:

1. Comprehensive analysis of multiple datasets: by utilizing well-known breast cancer datasets such as Breast Cancer Wisconsin, Breast Cancer Diagnosis, NKI Breast Cancer, and SEER Breast Cancer Dataset, this research provides a thorough examination of diverse data sources. This allows for a more robust understanding of the factors influencing breast cancer diagnosis.

2. Data preprocessing techniques: the application of rigorous preprocessing methods - including label encoding for categorical variables, linear regression for imputing missing values, Robust scalar normalization for feature scaling, and Tomek Link SMOTE for class imbalance - enhances the quality of the datasets and prepares them for effective machine learning modeling. These methodologies serve as a guideline for future research in handling similar data challenges.

3. Feature selection using L2 ridge regularization: the study employs L2 Ridge regularization for feature selection, facilitating the identification of critical predictors of breast cancer. This not only improves model interpretability but also offers valuable insights into the biological and clinical relevance of the identified features, which can inform further research and clinical practices.

4. Evaluation of multiple machine learning models: by comparing the performance of various machine learning algorithms - including Decision Trees, Random Forests, SVM, Neural Networks, KNN, XGBoost, and AdaBoost - this research contributes to the understanding of which models are most effective for breast cancer identification. The comprehensive evaluation using metrics such as accuracy, precision, recall, F1-score, and Kappa constant provides a solid foundation for future studies.

5. Implementation of robust performance measures: the inclusion of ROC curves and precision-recall curves as additional performance metrics offers a more nuanced view of classifier performance, enabling better decision making in clinical settings. This holistic assessment supports the deployment of machine

learning models in real-world scenarios, where the trade-off between sensitivity and specificity is crucial.

6. Impact on clinical practice: ultimately, this research aims to bridge the gap between machine learning advancements and clinical application. By demonstrating the potential of machine learning techniques in enhancing breast cancer identification, the findings contribute to the development of more effective diagnostic tools that can lead to timely interventions and improved patient outcomes.

METHODS

Label encoder

A label encoder[17] is a preprocessing technique used to convert categorical data into numerical values, which are better suited for machine learning algorithms. Label encoding assigns a unique integer to each category or label in the dataset. This approach is useful when the categorical variables do not have an inherent order, but they still need to be transformed into a format that models can work with.

Replace the missing values by linear regression

Replacing missing values using linear regression[18] is an imputation technique where missing data in a dataset are estimated based on the linear relationship between the target feature and other predictor features in the dataset.

Normalization using robust scalar

Normalization using robust scaler[19] is a data preprocessing technique that helps transform features in a dataset so that they are centered around the median and have a specific range. This method is particularly useful for datasets with outliers, as it mitigates their influence on the scaling process.

Oversampling using Tomek links and SMOTE

Tomek links and synthetic minority over-sampling technique (SMOTE) are commonly used methods to address class imbalance issues in classification tasks. Their combination is widely applied to improve the balance in datasets by increasing minority class instances and eliminating unclear samples from the majority class.

SMOTE[20] creates new synthetic samples for the minority class by interpolating between existing data points within that class. Tomek links[21] identifies pairs of data points, one from the majority class and the other from the minority class, that are nearest neighbors to each other. These pairs are considered ambiguous or noisy and are often removed to clarify class boundaries.

Using both techniques together forms an effective preprocessing strategy:

- SMOTE first oversamples the minority class by generating synthetic data points.

- Tomek links is then applied to remove noisy or ambiguous points from the majority class.

This combined process, referred to as SMOTE + Tomek, involves:

1. Oversampling the minority class using SMOTE to increase its representation in the dataset.

2. Applying Tomek links to the oversampled dataset to eliminate overlapping and ambiguous samples from the majority class, resulting in cleaner class boundaries and a more balanced dataset.

This approach helps improve model performance by addressing both class imbalance and boundary clarity between the classes.

Feature selection using L2 ridge regularization

Ridge feature selection[21], also known as Ridge Regression or Tikhonov regularization, is a method that helps to prevent overfitting in regression models by introducing a regularization term to the cost function. Unlike traditional feature selection methods that directly remove features, Ridge Regression shrinks the coefficients of less important features, reducing their impact but keeping all features in the model. This technique is particularly useful when there is multicollinearity (correlation among predictor variables) in the dataset.

In ordinary least squares (OLS) regression, the coefficients the β are obtained by minimizing the sum of squared residuals. The solution is given by:

$$\beta = (X^T X)^{-1} X^T y \tag{1}$$

where X is the matrix of input features, y is the vector of target values, and β is the vector of coefficients (including β_j).

In Ridge Regression, the cost function is modified to include a penalty term proportional to the square of the coefficients:

$$Cost \ = \ \sum_{i=1}^{n} (y_i - \ x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2}$$

where y_i is the actual value, β_j is the coefficient for feature j, λ the regularization parameter (the higher λ, the more significant the regularization effect), and p is the number of features.

In Ridge Regression (L2 regularization), feature selection is performed implicitly by shrinking the coefficients of less important features, but it does not set them exactly to zero.

Steps to rank features in Ridge Regression:

1. Train the Ridge Regression model with your dataset.

2. Extract the coefficients of the model (β_j)

3. Sort the coefficients by their absolute values (after standardization of features)

4. Rank the features based on the magnitude of their coefficients.

The features are not explicitly selected in the traditional sense because Ridge does not shrink coefficients to zero, unlike Lasso regression. Instead, Ridge penalizes the size of the coefficients, shrinking them toward zero but retaining all features in the model. This means that all variables remain part of the model, albeit with reduced influence depending on the penalty term. If a selection of features is needed using Ridge, an additional step is required. In this work, the features are ranked based on the magnitude of their coefficients. Features with larger absolute coefficients are typically considered more important.

RESULTS

This section presents the results obtained from the various machine learning models applied to breast cancer datasets, along with a comprehensive analysis of their performance. The preprocessing steps, including label encoding, missing value imputation via linear regression, Robust scalar normalization, and the application of Tomek Link SMOTE for balancing the classes, were critical in ensuring the quality of the data. Following these preprocessing techniques, we trained and evaluated multiple machine learning algorithms - decision tree, random forest, SVM, neural network, KNN, XGBoost, and AdaBoost. Each model's performance was assessed using key metrics, including accuracy, precision, recall, F1-score, and the Kappa constant, as well as additional measures such as the ROC curve and Precision-Recall curve. The analysis highlights the strengths and weaknesses of each model, providing insights into their applicability for breast cancer identification. These results not only validate the effectiveness of machine learning approaches in this context but also emphasize the importance of robust preprocessing and feature selection in achieving optimal performance. Table 1 shows the datasets used in the present work.

Correlation matrix

The correlation matrix is a valuable tool for identifying relationships between different attributes (or features) in a dataset. It quantifies how the values of two variables move together. In the context of breast cancer identification, a correlation matrix can help understand the relationships between clinical or genomic attributes and how these correlations might influence the outcomes predicted by machine learning models.

A correlation matrix is a table where each element represents the correlation coefficient between two variables. The value of the correlation coefficient ranges from -1 to 1:

- 1 indicates a perfect positive correlation: as one attribute increases, the other increases proportionally.

- -1 indicates a perfect negative correlation: as one attribute increases, the other decreases proportionally.

- 0 indicates no correlation: changes in one attribute do not affect the other.

For example, a visual correlation matrix for a breast cancer dataset includes attributes such as tumor size, cell texture, compactness, symmetry, and fractal dimension. The matrix would be color-coded, with shades of blue representing negative correlations, shades of red indicating positive correlations, and white or light hues denoting near-zero correlation. The diagonal of the matrix will have a correlation of 1, as each variable is perfectly correlated with itself.

Machine learning models

The machine learning models utilized in this study are Decision Tree[22], Random Forest[23] Support Vector Machine[24], Neural Networks[25], K-Nearest Neighbor[26], Naïve Bayes[27] Extreme Gradient Boosting (XGBoost)[28], and AdaBoost[29].

Confusion matrix

A confusion matrix is a fundamental tool in machine learning for evaluating the performance of classification models. It provides a tabular representation of the actual versus predicted classifications, allowing for the visualization of the model's performance across different classes. In a binary classification scenario, the confusion matrix typically consists of four components:

Table 1. Dataset description

S. no.	Dataset	No. of instances	No. of features	No. of categorical features	No. of class labels	No. of features selected through RIDGE	Dataset link details
1	Breast cancer Wisconsin	569	32	1	Malignant, benign	11	Kaggle dataset
2	Breast cancer diagnosis	286	13	6	No-recurrence-event, recurrence-events	3	UCI
3	NKI breast cancer	272	1,568	0	Alive, dead	653	Data world
4	SEER breast cancer dataset	4,023	16	10	Alive, dead	6	Kaggle

- True positives (TP): the number of instances correctly predicted as positive.

- True negatives (TN): the number of instances correctly predicted as negative.

- False positives (FP): the number of instances incorrectly predicted as positive instances (Type I error).

- False negatives (FN): the number of instances incorrectly predicted as negative instances (Type II error).

Figure 1 depicts the confusion matrix for binary classes.

From this matrix, the following important performance metrics are derived:

Accuracy: This metric indicates the overall correctness of the model's predictions and is calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

Accuracy provides a general sense of how well the model performs, but it may not be sufficient for imbalanced datasets where one class is more prevalent than the other.

Precision: Also known as positive predictive value, precision measures the proportion of true positive predictions among all positive predictions. It is defined as:

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

High precision indicates a low false positive rate, which is crucial in scenarios where the cost of false positives is high.

Recall: Also referred to as sensitivity or true positive rate, recall measures the proportion of actual positive instances that were correctly identified. It is calculated as:

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

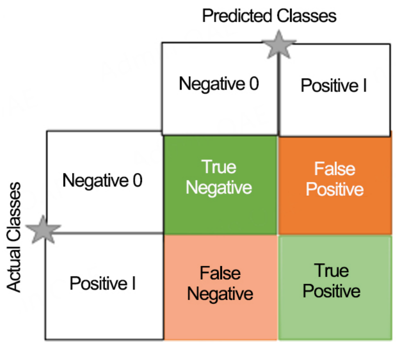


Figure 1. Confusion matrix.

High recall is important in contexts where missing positive cases is critical, such as in medical diagnoses.

F1-Score: The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both concerns. It is particularly useful in situations where there is an uneven class distribution. The F1-score is given by:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

This metric is beneficial when both precision and recall need to be optimized.

Kappa Constant (Cohen's Kappa): This statistic measures the agreement between predicted and actual classifications, accounting for chance agreement. It is calculated as:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{7}$$

where p_0 is the observed accuracy and p_e is the expected accuracy by chance. Kappa values range from -1 to 1, with values closer to 1 indicating strong agreement.

Receiver Operating Characteristic (ROC) Curve: The ROC curve is a graphical representation that illustrates the diagnostic ability of a binary classifier as its discrimination threshold is varied. The curve plots the True Positive Rate (TPR or Recall) against the false positive rate at various threshold settings. The area under the ROC curve (AUC-ROC) quantifies the overall performance of the model, with a value of 1 indicating perfect classification and a value of 0.5 representing no discriminative power.

Precision-Recall Curve: This curve is another important tool for evaluating the performance of a classifier, particularly in imbalanced datasets. It plots precision against recall for different threshold values, providing a more informative view of the trade-offs between these two metrics.

These metrics, derived from the confusion matrix, provide a comprehensive evaluation of the model's performance in breast cancer identification, enabling a deeper understanding of its strengths and limitations in clinical contexts.

Breast cancer Wisconsin dataset

Figure 2 presents the correlation matrix for the Breast Cancer Wisconsin dataset. Figure 3A and B display the dataset before and after applying Tomek-link SMOTE sampling, respectively. Figure 4 illustrates the selected features along with their corresponding coefficients ranked by importance. The top features selected - radius_worst, area_worst, compactness_mean, radius_se, concavity_mean, perimeter_mean, texture_worst, area_mean, concave_point_mean, concave_points_worst and radius_mean - are applied to machine learning models. Figures 5-7 depict the performance metrics, ROC curve, and precision-recall curve, respectively. For this dataset, XGBoost surpasses the other machine learning models, achieving an accuracy of 96% and an area under the curve (AUC) value of 1.

Breast cancer diagnosis dataset

Figure 8 presents the correlation matrix for the Breast Cancer Wisconsin dataset. Figure 9A and B show the dataset before and after applying Tomek-link SMOTE sampling, respectively. Figure 10 displays the selected features, ranked by importance, along with their coefficients. The top features identified are deg_malig, node_caps, and menopause, which were utilized in machine learning models for classification. Figures 11-13 provide visualizations of the performance metrics, ROC curve, and precision-recall curve, respectively. Among the models applied, XGBoost achieved the best results, with an accuracy of 72% and an AUC of 0.83.

NKI breast cancer data

Figure 14 illustrates the dataset before and after sampling using Tomek-link SMOTE. Figures 15-18 display the performance metrics, ROC curve, and precision-recall curve, respectively. Among the models tested, SVM outperformed the others, achieving 100% accuracy.

SEER breast cancer dataset

Figure 18 shows the correlation matrix for the Breast Cancer Wisconsin dataset. Figure 19A and B illustrate the dataset before and after applying Tomek-link SMOTE sampling, respectively. Figure 20 highlights the selected features, ranked by importance, along with their corresponding coefficients. The most significant features identified are progesterone status, N Stage, 6th stage, Race, A Stage, and Grade, which were used for classification in the machine learning models. Figures 21-23 provide visual representations of the performance metrics, ROC curve, and precision-recall curve, respectively. Of the models tested, Random Forest delivered the highest performance, achieving 92% accuracy and an AUC of 0.98.

Incorporating confidence intervals (CIs) into the analysis involves estimating a range within which the true value of a metric, such as accuracy or mean squared error, is likely to fall, typically with a 95% confidence level. Cross-validation is used to measure confidence intervals. The steps to measure CI using cross-validation are

1. Perform k-fold cross-validation.

2. Calculate the metric for each fold.

3. Compute the standard deviation of the metrics and use it to derive the CI using the formula: $CI = \bar{x} \pm Z.\frac{\sigma}{\sqrt{n}}$ where $\bar{x}$ is the mean of the metric, σ is the standard deviation, n is the sample size, and Z corresponds to the confidence level (1.96% for 95%).

Table 2 provides the confidence intervals (CIs) for all classifier models across the four datasets.

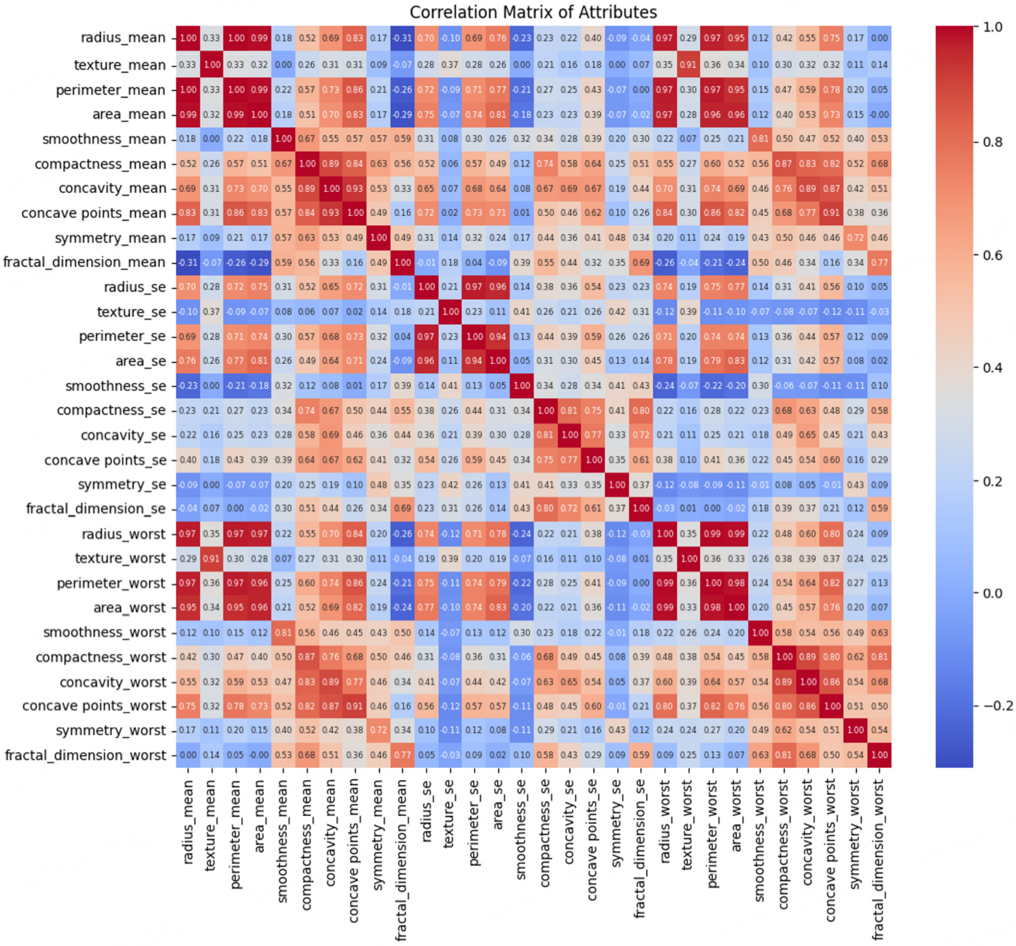


Figure 2. Correlation matrix.

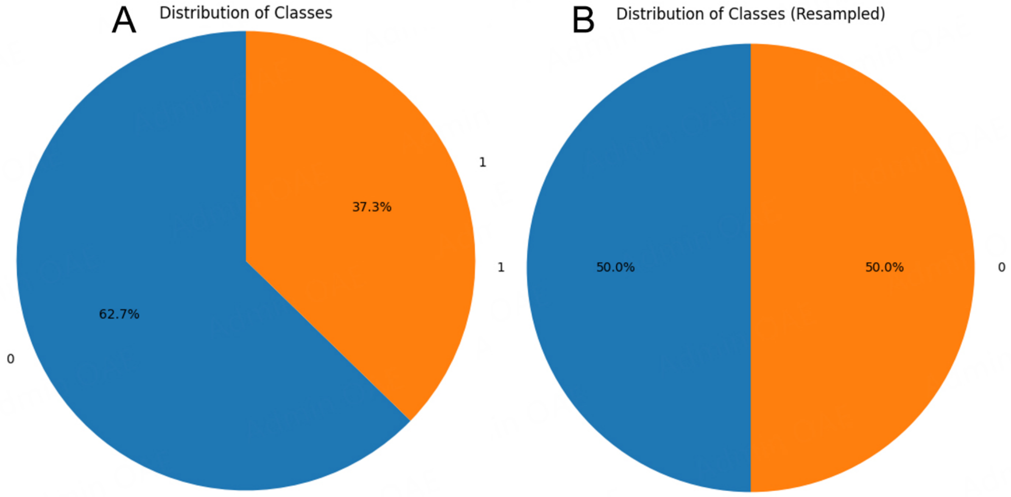


Figure 3. Breast cancer Wisconsin dataset before and after sampling.

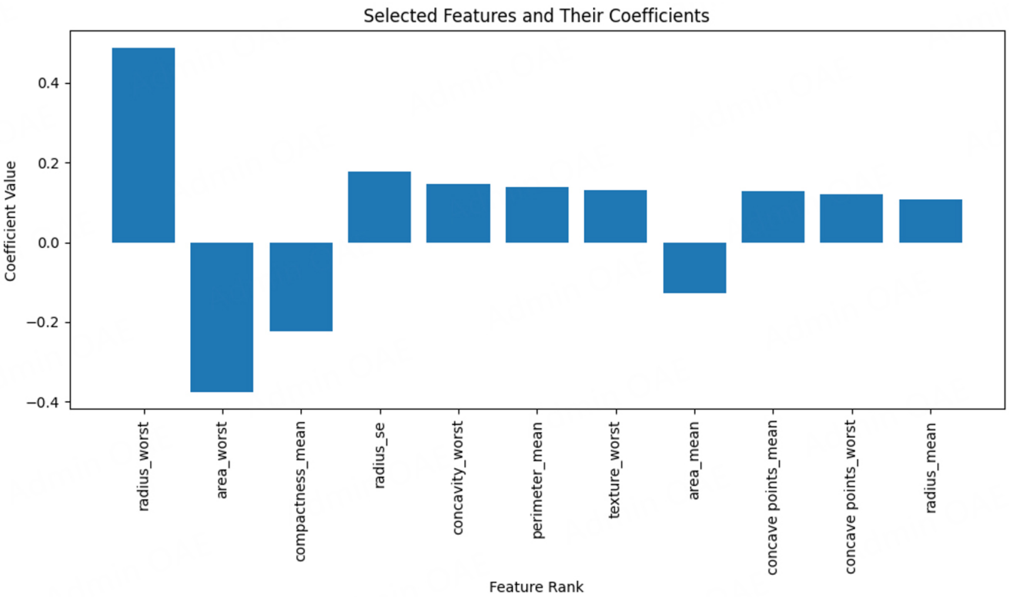


Figure 4. Features selected with their coefficients.

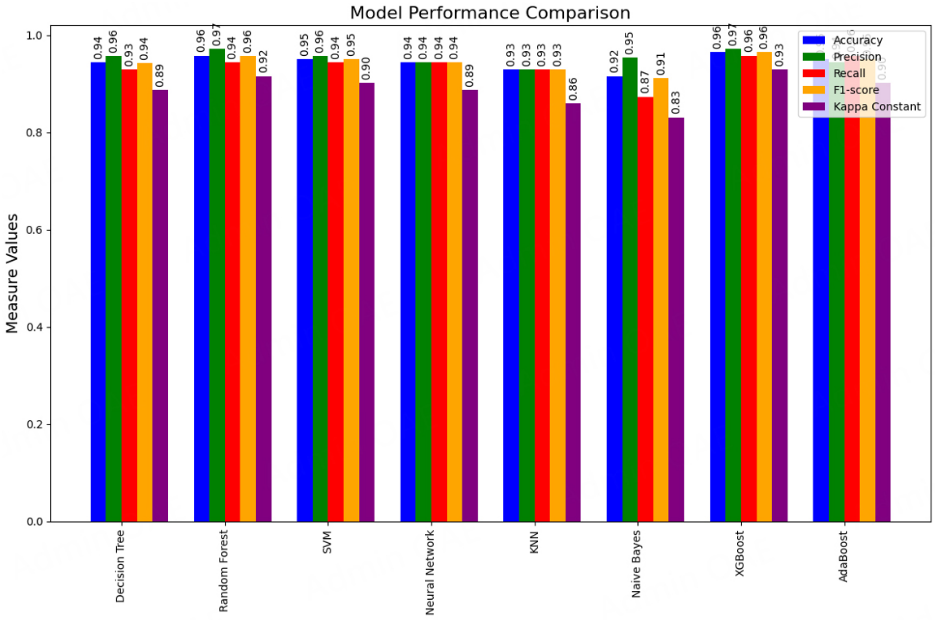


Figure 5. Performance measures.

DISCUSSIONS

This work successfully demonstrates the potential of machine learning techniques for improving the identification of breast cancer through the analysis of multiple datasets. By employing advanced preprocessing methods, including label encoding, linear regression for missing values, Robust scalar normalization, and Tomek Link SMOTE for class imbalance, we established a robust foundation for effective model training.

Table 2. Measuring confidence interval with cross entropy

Dataset	Measures	Decision tree	Random forest	SVM	XGB	AdaBoost	Neural networks	k-NN	Naïve bayes
Breast cancer Wisconsin	Mean CV score:	0.9366	0.9662	0.9704	0.9732	0.3913	0.9634	0.9676	0.9451
	Standard deviation:	0.0154	0.0150	0.0113	0.0103	0.0179	0.0053	0.0072	0.0163
	95% confidence interval	(0.9231, 0.9501)	(0.9530, 0.9794)	(0.9605, 0.9803)	(0.9642, 0.9823)	(0.9598, 0.9783)	(0.9588, 0.9680)	(0.9613, 0.9739)	(0.9308, 0.9594)
Breast cancer diagnosis	Mean CV score:	0.6994	0.7279	0.6591	0.7359	0.7042	0.6672	0.6646	0.6328
	Standard deviation:	0.0962	0.0957	0.1501	0.0966	0.0883	0.0890	0.0772	0.1433
	95% confidence interval	(0.6150, 0.7837)	(0.6440, 0.8118)	(0.5275, 0.7907)	(0.6512, 0.8206)	(0.6268, 0.7817)	(0.5891, 0.7452)	(0.5969, 0.7323)	(0.5072, 0.7584)
NKI_cleaned	Mean CV score:	0.8795	0.9154	0.9667	0.9256	0.9231	0.9077	0.6821	0.6974
	Standard deviation:	0.0224	0.0310	0.0192	0.0297	0.0363	0.0274	0.0971	0.1002
	95% confidence interval	(0.8599, 0.8991)	(0.8882, 0.9425)	(0.9498, 0.9835)	(0.8996, 0.9517)	(0.8913, 0.9549)	(0.8837, 0.9317)	(0.5969, 0.7672)	(0.6096, 0.7852)
SEER	Mean CV score:	0.4533	0.4696	0.7391	0.4448	0.3913	0.6702	0.5650	0.7059
	Standard deviation:	0.0433	0.0514	0.1793	0.0399	0.0179	0.1691	0.1564	0.1574
	95% confidence interval	(0.4154, 0.4913)	(0.4246, 0.5146)	(0.5819, 0.8963)	(0.4098, 0.4797)	(0.3757, 0.4070)	(0.5219, 0.8184)	(0.4279, 0.7021)	(0.5679, 0.8438)

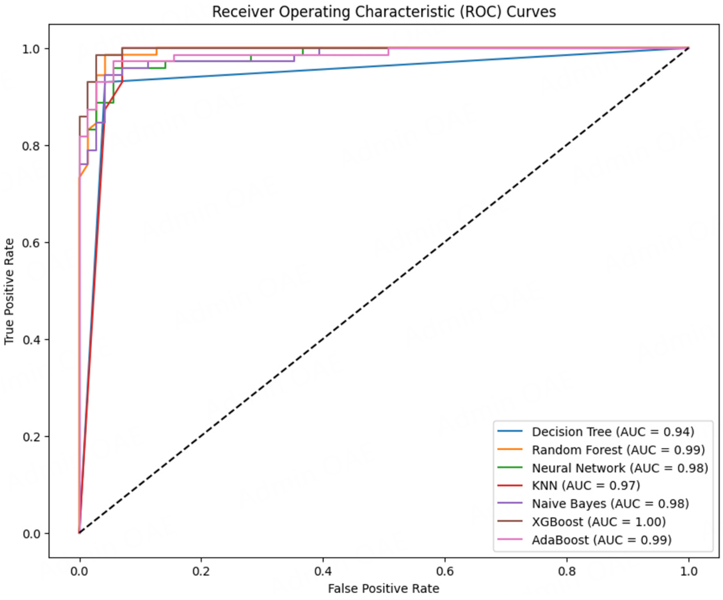


Figure 6. ROC curve.

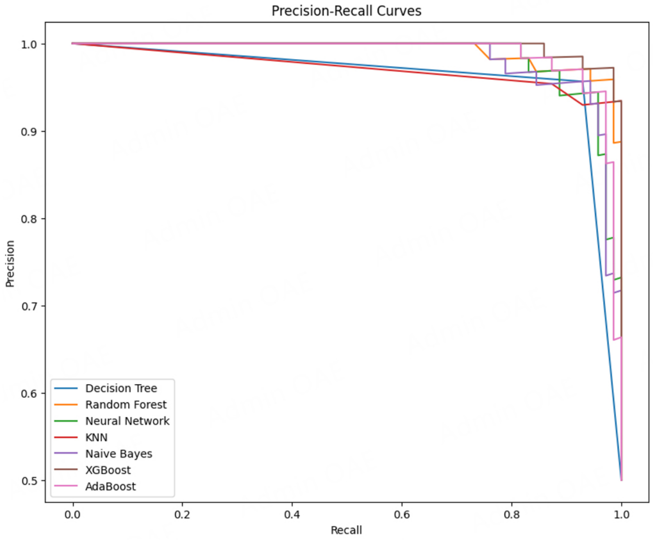
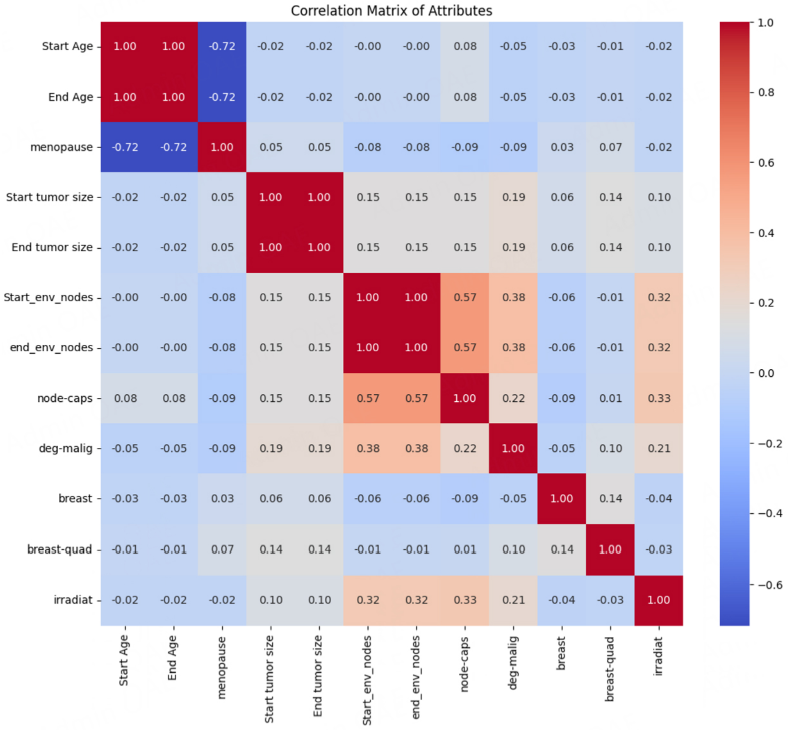


Figure 7. Precision recall curve.



	Start Age	End Age	menopause	Start tumor size	End tumor size	Start_env_nodes	end_env_nodes	node-caps	deg-malig	breast	breast-quad	irradiat
Start Age	1.00	1.00	-0.72	-0.02	-0.02	-0.00	-0.00	0.08	-0.05	-0.03	-0.01	-0.02
End Age	1.00	1.00	-0.72	-0.02	-0.02	-0.00	-0.00	0.08	-0.05	-0.03	-0.01	-0.02
menopause	-0.72	-0.72	1.00	0.05	0.05	-0.08	-0.08	-0.09	-0.09	0.03	0.07	-0.02
Start tumor size	-0.02	-0.02	0.05	1.00	1.00	0.15	0.15	0.15	0.19	0.06	0.14	0.10
End tumor size	-0.02	-0.02	0.05	1.00	1.00	0.15	0.15	0.15	0.19	0.06	0.14	0.10
Start_env_nodes	-0.00	-0.00	-0.08	0.15	0.15	1.00	1.00	0.57	0.38	-0.06	-0.01	0.32
end_env_nodes	-0.00	-0.00	-0.08	0.15	0.15	1.00	1.00	0.57	0.38	-0.06	-0.01	0.32
node-caps	0.08	0.08	-0.09	0.15	0.15	0.57	0.57	1.00	0.22	-0.09	0.01	0.33
deg-malig	-0.05	-0.05	-0.09	0.19	0.19	0.38	0.38	0.22	1.00	-0.05	0.10	0.21
breast	-0.03	-0.03	0.03	0.06	0.06	-0.06	-0.06	-0.09	-0.05	1.00	0.14	-0.04
breast-quad	-0.01	-0.01	0.07	0.14	0.14	-0.01	-0.01	0.01	0.10	0.14	1.00	-0.03
irradiat	-0.02	-0.02	-0.02	0.10	0.10	0.32	0.32	0.33	0.21	-0.04	-0.03	1.00

Figure 8. Correlation matrix.

The application of L2 Ridge regularization for feature selection allowed us to identify key predictors associated with breast cancer, providing valuable insights that could inform both clinical practice and future research. Our evaluation of various machine learning models - such as Decision Trees, Random Forests, Support Vector Machines, Neural Networks, K-Nearest Neighbors, Naïve Bayes, Extreme Gradient Boosting, and AdaBoost - yielded valuable performance metrics, including accuracy, precision, recall, F1-score, and Kappa constant. The additional assessment through ROC and Precision-Recall curves further elucidated the strengths and weaknesses of each model in clinical contexts.

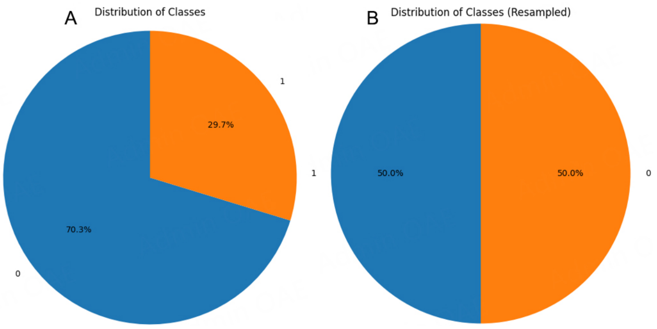


Figure 9. Breast cancer Wisconsin dataset before and after sampling.

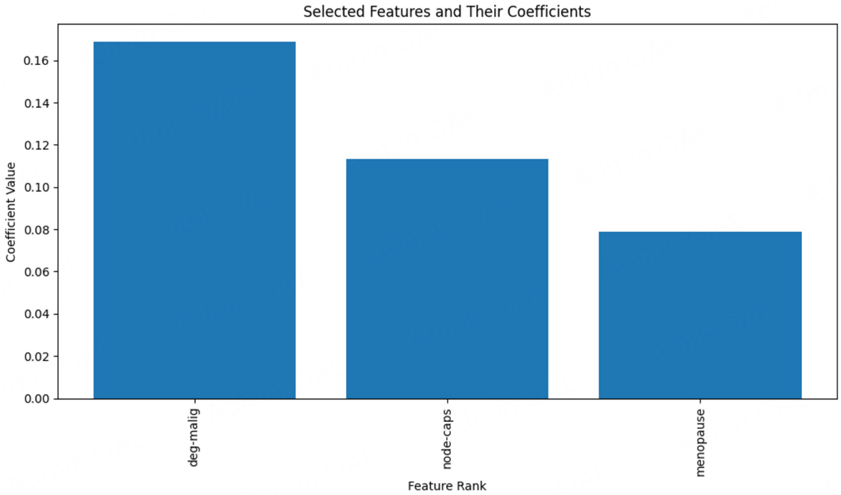


Figure 10. Selected features of breast cancer diagnosis dataset.

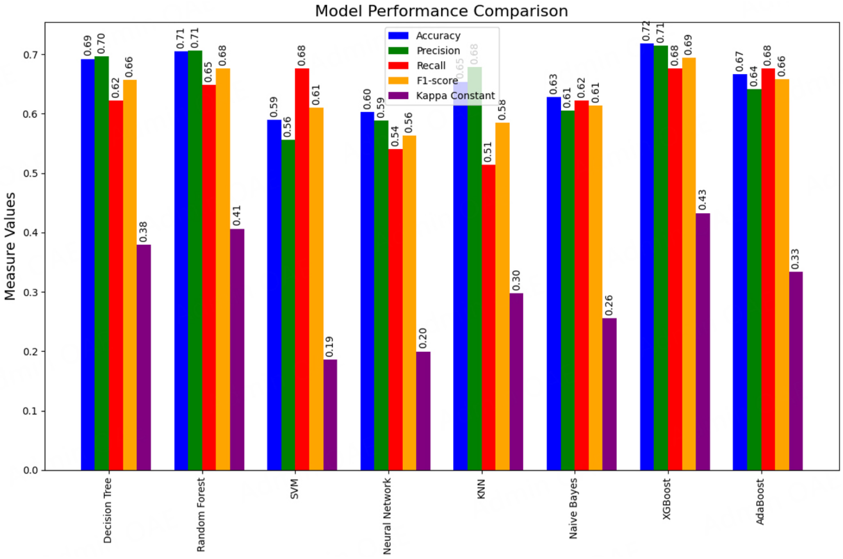


Figure 11. Performance measures.

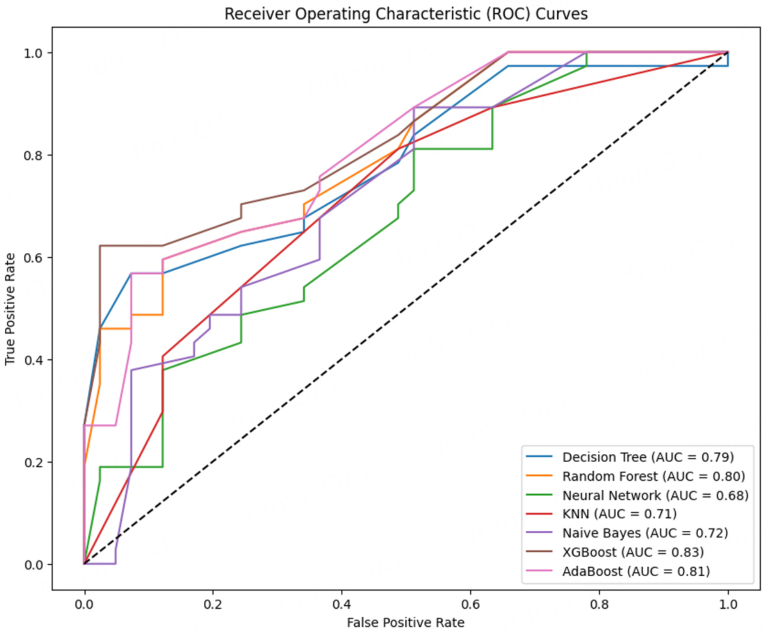


Figure 12. ROC curve.

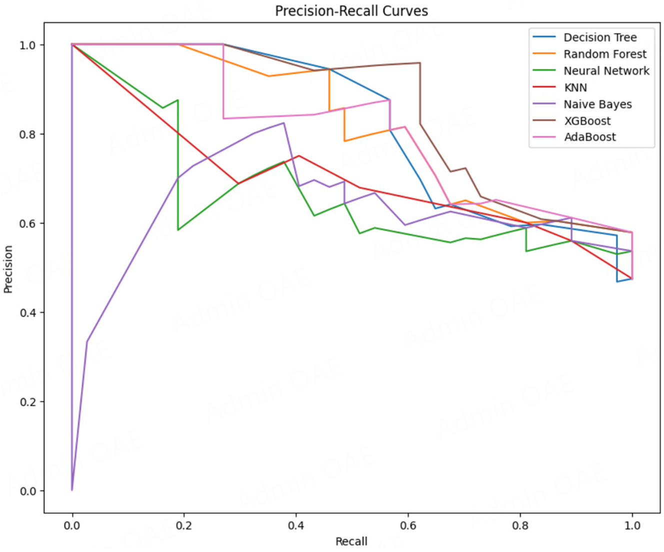


Figure 13. Precision recall curve.

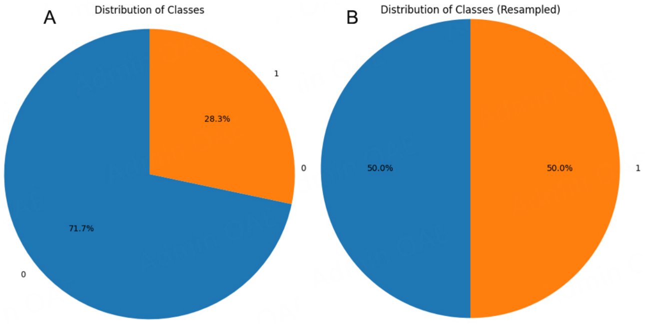


Figure 14. Dateset before and after sampling using Tomek-link SMOTE.

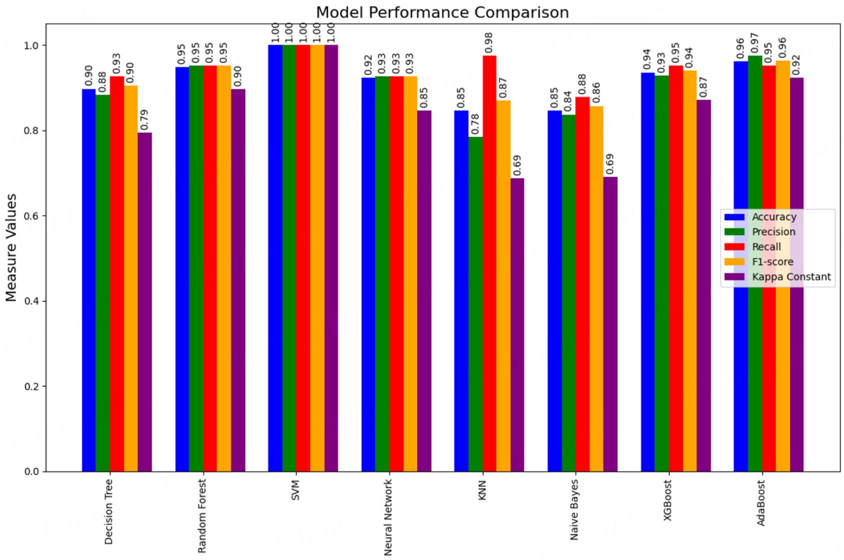


Figure 15. Performance measures.

The findings of this research emphasize the importance of integrating machine learning methodologies into the diagnostic process for breast cancer, showcasing their ability to enhance early detection and treatment outcomes. As the field continues to evolve, the approaches and insights generated from this study can serve as a roadmap for future investigations aimed at refining diagnostic tools and ultimately improving patient care.

XGBoost yielded the best results for the Breast Cancer Wisconsin and Diagnosis datasets, while SVM achieved 100% accuracy for the NKI Breast Cancer dataset, and Random Forest performed optimally for the SEER breast cancer dataset when using the selected features identified by L2 Ridge regularization. No single machine learning algorithm consistently outperformed across all datasets. The accuracy of each model is influenced by factors such as the number of features, the size of the dataset, hyperparameters, overfitting and generalization, and the feature importance.

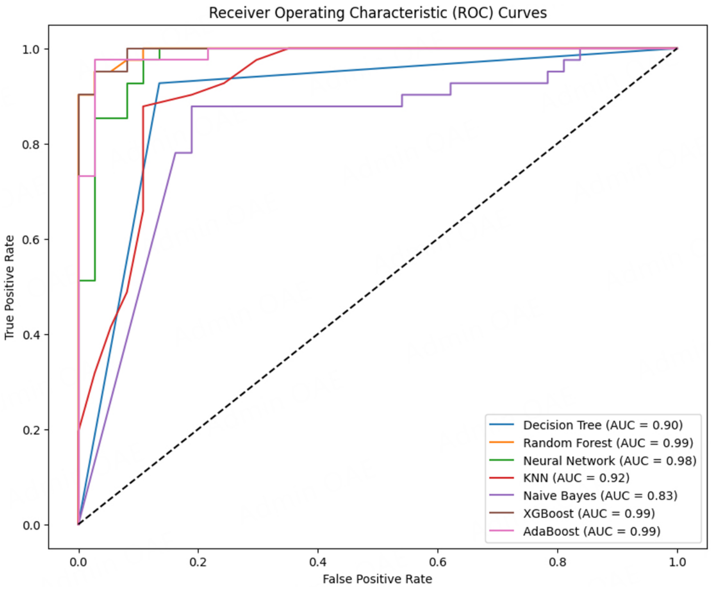


Figure 16. ROC curve.

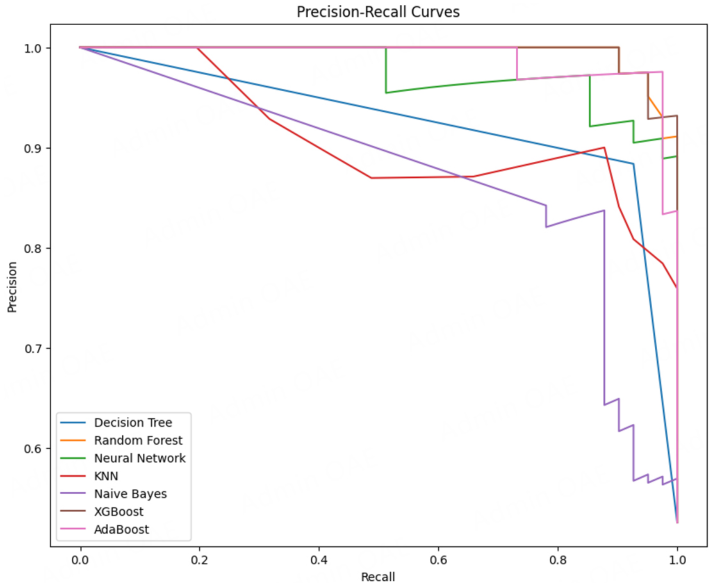


Figure 17. Precision recall curve.

The development of customized methodologies and the exploration of unique features not only enhance diagnostic accuracy in breast cancer but also hold the promise of transforming clinical practice, paving the way for more effective and personalized healthcare solutions.

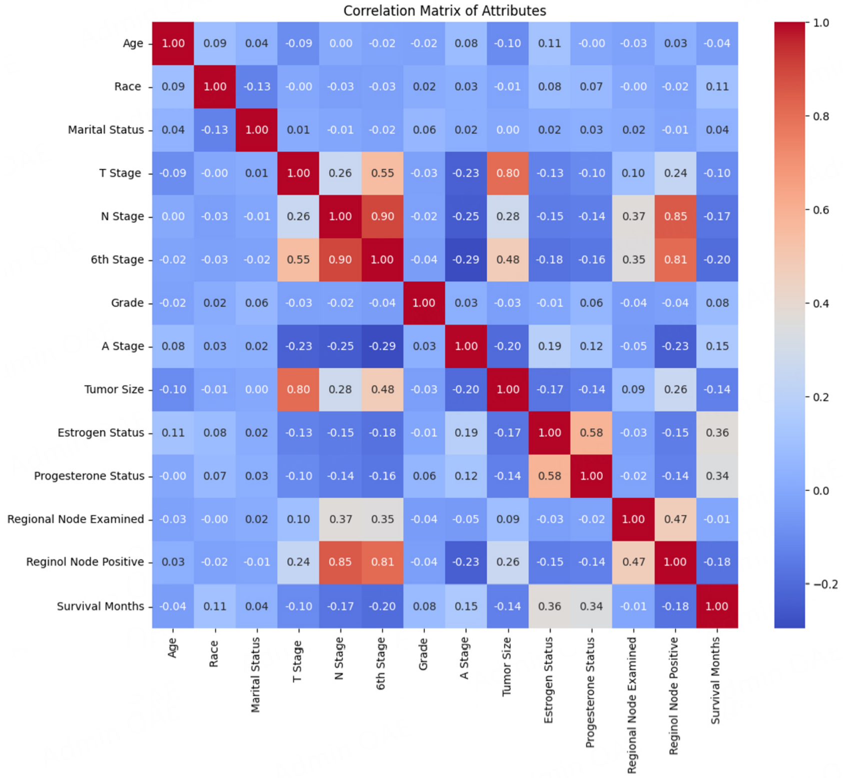


Figure 18. Correlation matrix.

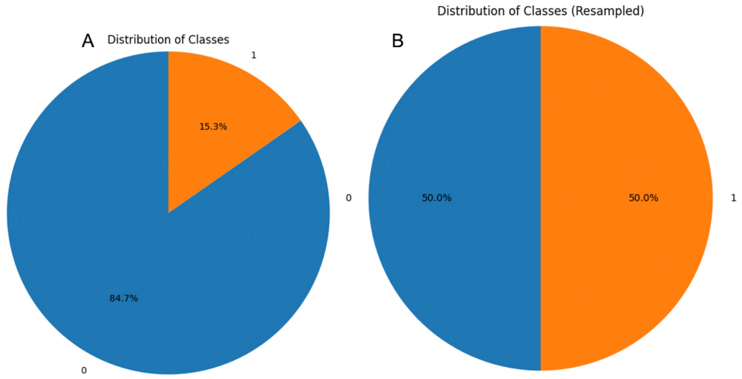


Figure 19. The dataset before and after Tomek-link SMOTE sampling.

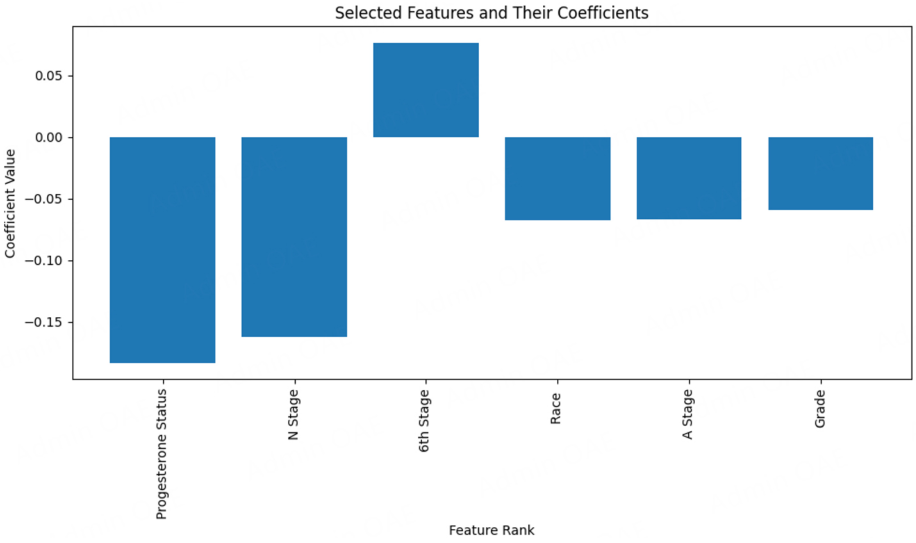


Figure 20. Selected features.

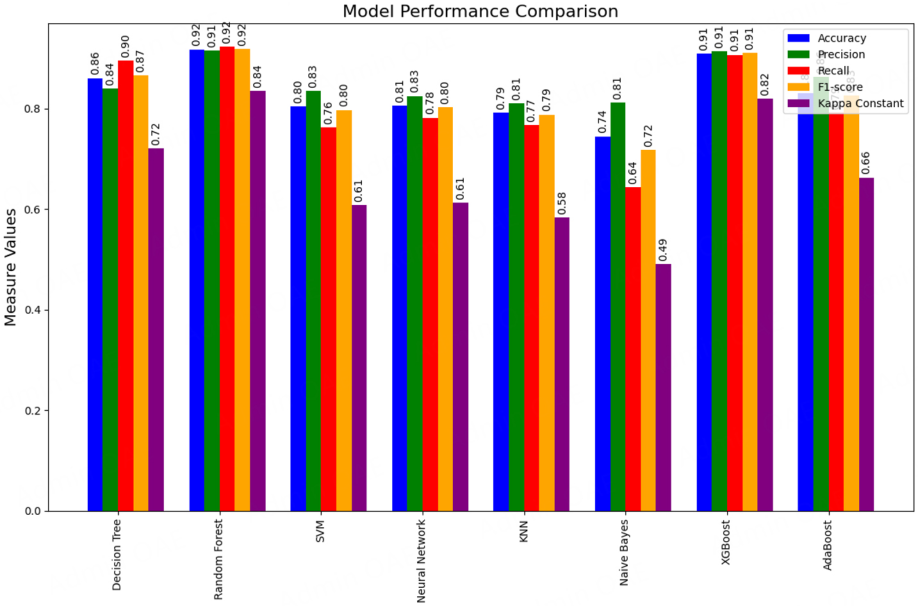


Figure 21. Confusion matrix for (a) Decision Tree, (b) Random forest, (c) SVM, (d) Neural Network, (e) K-nearest neighbor, (f) Naïve Bayes, (g) XGBoost, and (h) AdaBoost.

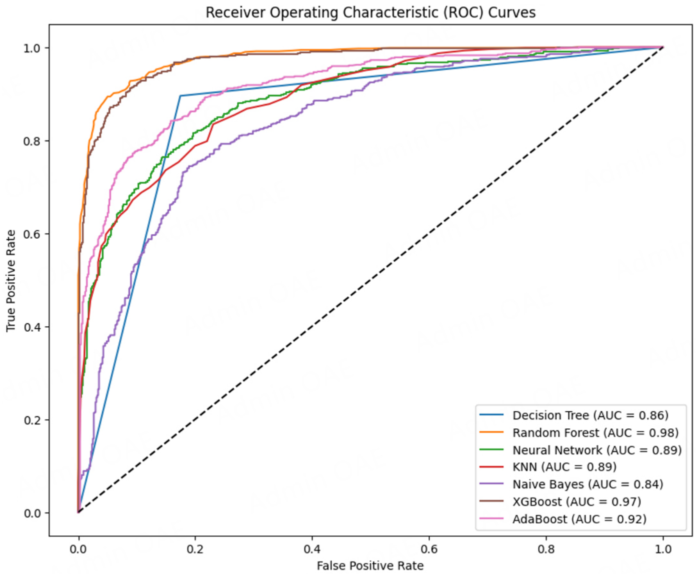


Figure 22. ROC curve.

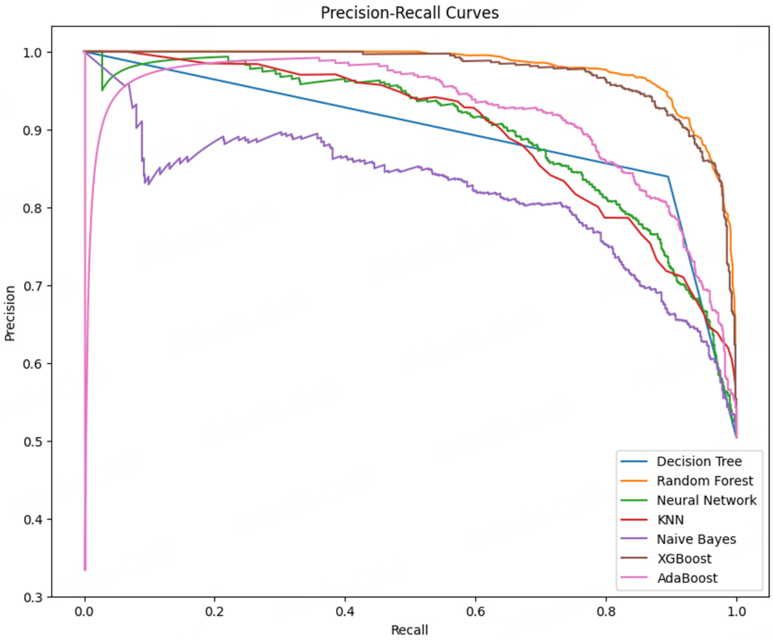


Figure 23. Precision recall curve.

DECLARATIONS

Author's contributions

Conceptualization of the study, dataset collection, and preparation of the methodology, Conducted feature selection using ridge regularization and supervised the machine learning experiments, Contributed to writing and reviewing the manuscript: Kandhasamy P

Implemented the machine learning models and carried out the data analysis, Performed evaluation and comparison of results across multiple datasets, Assisted in preparing the manuscript and contributed to editing and formatting: Devi DP

Conducted literature review and provided critical insights on the relevance of feature selection techniques, Contributed to data preprocessing and visualization, Participated in reviewing the manuscript and provided final approval for submission: Kandhasamy S

Availability of data and materials

1. Breast cancer Wisconsin available in Kaggle dataset https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data; 2. Breast cancer diagnosis available in UCI machine learning repository https://archive.ics.uci.edu/dataset/14/breast+cancer; 3. NKI breast cancer dataset available in data world https://data.world/deviramanan2016/nki-breast-cancer-data; 4. SEER breast cancer dataset available in Kaggle dataset https://www.kaggle.com/datasets/mansigambhir13/seer-breast-cancer-dataset.

Financial support and sponsorship

None.

Conflict of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

REFERENCES

1. Sun YS, Zhao Z, Yang ZN, et al. Risk factors and preventions of breast cancer. *Int J Biol Sci.* 2017;13:1387-97. DOI PubMed PMC
2. Giaquinto AN, Sung H, Newman LA, et al. Breast cancer statistics 2024. *CA Cancer J Clin.* 2024;74:477-95. DOI
3. Chaudhury AR, Iyer R, Iychettira KK, Sreedevi A. Diagnosis of invasive ductal carcinoma using image processing techniques. In Proceedings of the 2011 International Conference on Image Information Processing, 3-5 November 2011, Shimla, India; pp. 1-6. DOI
4. Graydon J, Galloway S, Palmer-Wickham S, et al. Information needs of women during early treatment for breast cancer. *J Adv Nurs.* 1997;26:59-64. DOI
5. Gamble P, Jaroensri R, Wang H, et al. Determining breast cancer biomarker status and associated morphological features using deep learning. *Commun Med.* 2021;1:14. DOI PubMed PMC
6. Foo CT, Langton D, Thompson BR, Thien F. Functional lung imaging using novel and emerging MRI techniques. *Front Med.* 2023;10:1060940. DOI PubMed PMC
7. Yen C, Lin CL, Chiang MC. Exploring the frontiers of neuroimaging: a review of recent advances in understanding brain functioning and disorders. *Life.* 2023;13:1472. DOI PubMed PMC
8. Nagpal P, Prakash A, Pradhan G, et al. MDCT imaging of the stomach: advances and applications. *Br J Radiol.* 2017;90:20160412. DOI PubMed PMC
9. Khalid A, Mehmood A, Alabrah A, et al. Breast cancer detection and prevention using machine learning. *Diagnostics.* 2023;13:3113. DOI PubMed PMC
10. Khan F, Khan MA, Abbas S, et al. Cloud-based breast cancer prediction empowered with soft computing approaches. *J Healthc Eng.* 2020;2020:8017496. DOI PubMed PMC
11. Al-Antari MA, Al-Masni MA, Choi MT, Han SM, Kim TS. A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. *Int J Med Inf.* 2018;117:44-54. DOI PubMed
12. Naji MA, Filali SE, Aarika K, Benlahmar EH, Abdelouhahid RA, Debauche O. Machine learning algorithms for breast cancer prediction and diagnosis. *Proc Comput Sci.* 2021;191:487-92. DOI
13. Nakagawa T, Hayashi K, Ogawa A, et al. Bone marrow carcinomatosis in a stage IV breast cancer patient treated by letrozole as first-line endocrine therapy. *Case Rep Oncol.* 2022;15:436-41. DOI PubMed PMC
14. Purrahman D, Mahmoudian-Sani MR, Saki N, Wojdasiewicz P, Kurkowska-Jastrzębska I, Poniatowski ŁA. Involvement of progranulin (PGRN) in the pathogenesis and prognosis of breast cancer. *Cytokine.* 2022;151:155803. DOI PubMed

15. Ogundokun RO, Misra S, Douglas M, Damaševičius R, Maskeliūnas R. Medical internet-of-things based breast cancer diagnosis using hyperparameter-optimized neural networks. *Future Int.* 2022;14:153. DOI
16. Sharmin S, Ahammad T, Talukder MA, Ghose P. A hybrid dependable deep feature extraction and ensemble-based machine learning approach for breast cancer detection. *IEEE Access.* 2023;11:87694-708. DOI
17. Manikandan P, Durga U, Ponnuraja C. An integrative machine learning framework for classifying SEER breast cancer. *Sci Rep.* 2023;13:5362. DOI PubMed PMC
18. Little RJA, Rubin DB. Statistical analysis with missing data. Hoboken, NJ: John Wiley & Sons; 2002. DOI
19. Raschka S, Mirjalili V. Python machine learning. Birmingham, UK: Packt Publishing Ltd.; 2017. Available from: http://radio.eng.niigata-u.ac.jp/wp/wp-content/uploads/2020/06/python-machine-learning-2nd.pdf [Last accessed on 22 Jan 2025].
20. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res.* 2002;16:321-57. DOI
21. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics.* 1970;12:55-67. DOI
22. Quinlan JR. Induction of decision trees. *Mach Learn.* 1995;1:81-106. DOI
23. Breiman L. Random forests. *Mach Learn.* 2001;45:5-32. DOI
24. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* 1995;20:273-97. DOI
25. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature.* 1986;323:533-6. DOI
26. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inform Theory.* 1967;13:21-7. DOI
27. John GH, Langley P. Estimating continuous distributions in bayesian classifiers. In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI). 1995; pp. 338-45. DOI
28. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). 2016; pp. 785-94. DOI
29. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. In: Lecture notes in computer science. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995; pp. 23-37. DOI