STM JOURNALS

Journal of
Artificial Intelligence
Research & Advances

ISSN: 2395-6720
Volume 12, Issue 2, 2025
May–August
DOI Journal: 10.37591/JoAIRA

https://journals.stmjournals.com/joaira

Research JoAIRA

Comparative Study of Machine Learning Algorithms for Detection of Breast Cancer

Sayli Rajendra Dholam*

Abstract

Breast cancer continues to be the most commonly diagnosed cancer among women, with more than 2.3 million new cases diagnosed yearly worldwide. It is stated as the leading cause of cancer-related deaths. Therefore, this emphasizes the dire necessity for early diagnosis with a view to improving survival. Early diagnosis elevates the effectiveness of prediction and treatment. This research carries out a structured and analytical evaluation of various machine learning algorithms, namely, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and XGBoost for the purpose of breast cancer detection. This study uses the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, one of the standard criterions sets in this field, to classify tumors into benign or malignant. The performance of these algorithms is evaluated in detail using various metrics, namely, accuracy, precision, recall, F-1 score and confusion matrices, to provide scores useful for proper insight of the said groups in terms of their respective capabilities in carrying out diagnoses.

Keywords: Breast cancer detection, machine learning algorithms, logistic regression, decision tree, random forest, support vector machine (SVM), K-nearest neighbors (KNN), Naïve Bayes, XGBoost, confusion matrix, ROC-AUC analysis, accuracy, precision, recall, F1 score, Wisconsin diagnostic breast cancer (WDBC)

INTRODUCTION

Breast cancer is ranked as the major cause of cancer-related deaths among women worldwide with over 2.3 million cases diagnosed annually [1]. The World Health Organization (WHO) identifies early recognition as a critical strategy to refine the results [2]. In 2020, breast cancer was found to be the most frequently diagnosed cancer type in the world, with more than a sum total of 2.26 million breast cancer cases around the world. Breast cancer persisted to be the second most common cancer type after lung cancer in 2022, regardless of an estimated jump of new cases to more than 2.31 million [3]. It is the most common type of cancer diagnosed among women and the leading cause of cancer-related deaths. It claims the fourth spot among all the general causes of deaths caused due to cancer. In 2022, an estimated 2,296,840 females were picked with fresh breast cancer cases [4].

Machine learning is potentially revolutionary in addressing this disparity. As machines analyze large datasets to better classify and detect diseases, ML brings creative tools with the great ability to source patterns that would escape human detection. The Wisconsin Diagnostic Breast Cancer Dataset is a canonical standard dataset used in this domain [5].

***Author for Correspondence**
Sayli Rajendra Dholam
E-mail: sayli.dholam.official@gmail.com

Research Scholar, MCA, Thakur Institute of Management Studies, Career Development & Research (TIMSCDR), Mumbai, Maharashtra, India

Received Date: March 06, 2025
Accepted Date: April 22, 2025
Published Date: June 25, 2025

Citation: Sayli Rajendra Dholam. Comparative Study of Machine Learning Algorithms for Detection of Breast Cancer. Journal of Artificial Intelligence Research & Advances. 2025; 12(2): 113–129p.

This impressive collection of characteristics permits the separation of tumors into benign and malignant, thus providing a worthy foundation for the study. Such a thorough study investigates the performance of seven machine learning algorithms: Naïve Bayes, SVM, KNN, Random Forest, Decision Trees, XGBoost, and Logistic Regression, in terms of accuracy, precision, recall, F1 measure, and confusion matrices, while distilling insights into their respective strengths and weaknesses so as to contribute an insightful discourse in efforts to promote AI-based clinical diagnostics.

LITERATURE REVIEW

The implementation of machine learning in breast cancer detection marks a significant moment in Medicare where diagnostic accuracy and efficiency are expected to observe an increase. Breast cancer which is one of the most prevalent cancers globally, calls for timely and accurate recognition methods to increase the probability of survival and reduce mortality. A number of algorithms have been tested for breast cancer detection, and their unique pros and cons are well documented. Logistic Regression, viewed as an established statistical method, is acknowledged for its interpretability for binary classification. Decision Tree and Random Forest handle nonlinear relationships. Support Vector Machines are still popular and possess the ability to perform even in a high-dimensional space. K Nearest Neighbors and Naive Bayes are easy, quick, and applicable for smaller datasets. However, Ensemble techniques like XG-Boost are gaining popularity due to their ability to perform high-performance classification tasks.

In 1993, Street *et al*. presented a system intended for the enhancement of breast tumor diagnosis through nuclear feature extraction, using image processing and machine learning techniques [6]. This system uses an active contour model to detect cell nucleus boundaries and a classifier trained via a leave-one out approach with an accuracy of 97.07%.

In 2014, Vig addressed breast cancer diagnosis by employing the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which, due to Double Cross-Validation (DCV) application, differentiated itself well because it addressed the biases found in single cross-validation approaches [7]. By dividing the data into training, testing, and validation subsets, the study also assures the reliability of the validating procedure used to evaluate any classifier's performance. Random Forests reached the highest of 95.64% for overall accuracy, with excellent sensitivity (0.97) and specificity (0.94), while Naive Bayes found itself to have the worst accuracy at 65.27%. That is the explanation for why it fails when testing datasets using interdependent features. The authors conclude that advanced machine learning models, in particular, Random Forests and SVMs, can be exceptionally accurate in breast cancer diagnosis.

Bazazeh and Shubair in 2016 aimed to compare the performance of three machine learning algorithms, Support Vector Machines, Random Forest, and Bayesian Networks, in relation to breast cancer detection and diagnosis [8]. The paper evaluates these models for classification performance using Original Wisconsin Breast Cancer Data on the grounds of performance measures requiring classification accuracy, precision, recall, and the area under the OC curve. BN performed well in malignant classification cases with an 98.3%, while SVM had the overall highest mean average of 97.0%, and RF outperformed the estimates with an exceptional AUC of 99.9%. This study justifies that different ML models have their certain strengths.

Shahnaz *et al.* implemented classifiers, as of 2017, Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression, Random Forest, and neural network models such as Multilayer Perceptron (MLP) and Convolutional Neural Network (CNN) were used [9]. They achieved better results with feature selection techniques and cross-validation, where CNN gave 98.06%. The analysis provides greater insight into predictive performance levels regarding CNNs in this fashion, and provides suggestions to pursue the neural network-based classifications for more effective cancer detection.

In 2018, Tahmooresi *et al*. sought to increase early breast cancer detection with a hybrid of various ML algorithms [4]. The Wisconsin Breast Cancer Database (WBCD) provided extensive feature richness and has been used extensively in medical research, including characterizations of cell thickness, uniformity, and size. Moreover, SVM and Artificial Neural Network (ANN) models achieved the best performance, with an accuracy of 99.8%, demonstrating that they are effective in reliable early on detection.

A study by Obaid *et al*. in 2018 entitled "Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer" performed quadratic SVM and various other machine learning models on Wisconsin Breast Cancer (Diagnostic) Data Set to classify the breast cancer tumors [10]. The quadratic SVM model achieved an accuracy of 98.1%, which was higher than the accuracies obtained by the other models and a significant reduction in detecting the cancerous tumors was observed. K-NN (medium type) had the second highest accuracy of 96.7%. Decision tree (medium type-tree) achieved an accuracy of 93.7%.

Austria *et al*. in 2019, analyzed various machine learning (ML) models for breast cancer detection using clinical features from normal blood analyses, such as age, BMI, glucose, insulin, leptin, and MCP-1 [11]. The highest accuracy obtained was 74.14% for gradient boosting, while KNN had the shorter time for training (0.0006 sec) and Nonlinear SVM had zero time for testing. In emphasizing the predictors, the BMI was shown to have the best forecaster, followed by glucose.

Akbugday, in Oct 2019, monitored the performance of three machine learning algorithms; k-Nearest Neighbors (KNN), Naive Bayes (NB), Support Vector Machines (SVM), and Random Forest have been revealed as effective methods for classifying breast cancer data based on the Wisconsin Breast Cancer Dataset [2]. Of these models, Naive Bayes emerged as the best-performing method, achieving an accuracy of 97.28%. The models k-NN and C in SVM achieved the highest accuracy, with results of 96.85%. Following this, there was some gain in performance after the removal of the "Single Epithelial Cell Size" attribute, which yielded an admirable 99.01% accuracy for the NB classifier, known for its quick classification of data.

Paper by Omondiagbe *et al*. aims at using machine learning (ML) techniques integrated with dimensionality reduction methods for improving breast cancer diagnosis in 2019 [12]. The SVM-LDA approaches achieved promising results, with classification accuracy of 98.82%, sensitivity of 98.41%, specificity of 99.07%, and AUC or area under the receiver operating characteristic (ROC) curve being 0.9994.

In 2020, Sengar *et al*. conducted comparative analysis of breast cancer prediction using the data available in the Wisconsin Diagnostic Breast Cancer (WDBC) dataset [5]. The study aims to find out which algorithm works better for predicting cancer malignancy: Logistic Regression or Decision Tree. Logistic Regression achieved a high rate of accuracy, but the Decision Tree Classifier counterpart performed marginally better, thus made the final choice for its pinpoint accuracy.

Mushtaq *et al*. in 2020 evaluated the performance of K-Nearest Neighbor (KNN) Algorithm for classification of cases of breast cancer using the datasets of Wisconsin Breast Cancer (WBC) and Wisconsin Diagnostic Breast Cancer (WDBC) [13]. The study investigates the effect of eight different distance functions, including Manhattan, Canberra, and Euclidean, when varying K (from 1 to 59), on the classification accuracy. Applied in this study were feature selection techniques such as Chi-square and L1-based linear support vector classifier (LSVC), to improve the model's efficiency by selecting the most meaningful features. It was found that the Manhattan distance function combined with Chi-square feature selection yields the highest accuracy of 99.42% on the WBC dataset and 98.62% on the WDBC dataset.

Ahmed *et al*. in 2020 using five algorithms aimed at analyzing and optimizing the machine learning algorithms of breast cancer prediction: Naïve Bayes, Support Vector Machine, Multilayer Perceptron, J48(a Decision Tree), and Random Forest [1]. It has been revealed that Naïve Bayes became the best-performing method, having achieved 97.28% accuracy, followed by some gain in performance after the removal of "Single Epithelial Cell Size" attribute which has yielded an admirable 99.01% accuracy.

Naji *et al*. in 2021, aimed to identify the most effective algorithm to make accurate predictions, employing the Support Vector Machine, Random Forest, Logistic Regression, Decision Tree (C4.5), and K-Nearest Neighbors [14]. With testing accuracy of 97.2% and areas under the curve of 96.6%, SVM outperformed the rest classification algorithms according to several metrics, which included accuracy, precision, sensitivity, and so on.

In 2022, Sakib *et al*. tested how machine learning (ML) and deep learning (DL) methods compared in breast cancer detection and classification using the Breast Cancer Wisconsin Diagnostic dataset [15]. Five kinds of ML executors were evaluated in the study: Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), K Nearest Neighbors (KNN), and a Neural Network (NN) with default, tuned parameters, and 10-fold cross validated. The one with the highest performance, Random Forest, reached an accuracy of 96.66% and F1-score of 0.963 with tuned parameters, with a high grade for generalization, being able to attain the cross-validation accuracy of 96.84%, with Logistic Regression being the second in performance, while the Neural Network displayed the lowest accuracy, standing at 90.35%.

Kadhim and Kamil examined various machine learning (ML) algorithms for breast cancer classification [3]. The researchers pre evaluated 11 ML classifiers with respect to their performance: Decision Tree (DT), Quadratic Discriminant Analysis (QDA), AdaBoost (AB), Bagging Meta estimator (BME), Extra Randomized Trees (ERT), Gaussian Process Classifier (GPC), Ridge Classifier, Gaussian Naive Bayes, K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), and Support Vector Classifier (SVC). 80% of the dataset was used to train while the other 20% was used to test. The criterion of performance was measured with respect to accuracy, specificity, precision, sensitivity, and F1-score. ERT emerged as the winner in performance, recording an accuracy of 97.36% and yielding a top-level F1 score of 96.77%.

Khan *et al*. worked on data from the University of Wisconsin Hospitals. Algorithms like, XGBoost, Logistic Regression, Random Forest, Decision Tree, and Naive Bayes were evaluated on performance metrics like, accuracy, sensitivity, specificity, and F1 score [16]. The conclusion made was that the XGBoost model performed best with commendable metrics: accuracy of 94.92%, sensitivity of 98.5%, specificity of 97.5%, and an F1 score of 99%.

Kushwaha, in 2023 sought to increase the success of breast cancer detection with the help of machine learning (ML), where performance is evaluated using four supervised learning models on two datasets: the Coimbra breast cancer dataset (BCCD) and the Wisconsin diagnostic breast cancer dataset (WDBC): Logistic regression (LR), support vector machine (SVM), nearest neighbor (KNN), and decision tree (DT) [17].

In 2023, Hossin *et al*. performed their comparisons of eight systems composed of Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, Support Vector Machine, AdaBoost, and Gaussian Naive Bayes [18]. Datasets have undergone preprocessing: standardization, 80:20 split train-test, exploratory analysis, and feature selection. They further did all preprocessing steps, including data clean-up, feature conceptions applied by Univariate Feature Selection (UFS) and Recursive Feature Elimination (RFE), and cross validation. The results of the investigation indicated very good results in which the best models had an accuracy of 99.12%, sensitivity of 97.73%, specificity of 100% and an F1 score of 99%. Nearly perfect AUC scores of 98.86% for the models were also shown.

Hyperparameter tuning was achieved by Grid Search Cross Validation to further enhance model performance. Logistic regression and SVM were the best performing algorithms with accuracy of 99.42%.

METHODOLOGY

This section details the methodology adopted for the development and evaluation of machine learning models for breast cancer detection. A flowchart shows the flow of activities from data preprocessing to model evaluation. The Breast Cancer Wisconsin (Diagnostic) Dataset [14] from the UCI Machine Learning Repository served as the dataset used in this work. The machine learning methods applied were Logistic Regression, Decision Tree, Random Forest, Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Naive Bayes, and XGBoost. For model performance assessment, accuracy, precision, recall, F1-score, and ROC-AUC curve, confusion matrices were applied to check the classification efficiency and robustness. This strengthens the roadmap of standardizing various algorithms to assess their performances in the specific dataset.

Dataset

The Breast Cancer Wisconsin (Diagnostic) Dataset [8], which comes from the UCI Machine Learning Repository, is also found on Kaggle. It is a benchmark dataset for evaluating machine learning algorithms in medical diagnostics. The data consists of 569 samples expressed in 30 characteristics of real values. The features were simply derived from FNA images of breast masses. Each of these features states certain cellular properties; including radius, texture, smoothness, and compactness, giving an elaborate description of the tumor. The dataset is used to classify tumors as either benign (357 instances, approximately 63%) or malignant (212 instances, approximately 37%). Every attribute has been enumerated by three computations, those being mean, standard error (SE), and worst (maximum) value.

Methods Applied

The below are the methods applied:

- *Logistic regression*: Regression is a dominant supervised learning approach when it comes to binary classification tasks. Linear regression estimates continuous variables, while logistic regression gives odds that can be mapped into categories. Hence, it is intended for solution or problems with a binary outcome or those that may be non-exhaustive, for instance, a diagnosis of breast tumors being either benign malignant.
- *Decision tree*: A decision tree divides the dataset into various subsets which depend on the values of the various features thus creating a structure like a tree of decision rules. Each internode within the tree is a decision made on the basis of a particular feature while the tips of the branches are the outcomes which could either be a class label in the case of classification or a number in the case of regression. The process of building the decision tree continues until only homogeneous data is presented in the nodes or until no nodes can be further split. This structure is simple to understand since every decision can be shown by simply taking the path from a root node to a leaf node.
- *Random forest*: The Random Forest is a rich ensemble learning technique which is based on the building of several decision trees during the training, and returns the class which is chosen by the majority of the trees for classification or averages the outputs for regression. Each tree constituting the forest is grown on a different portion of the data and each tree is also different because of the feature selection as well as the data sampling (bagging), thus making the approach stable and less likely to overfit.
- *Support vector machine (SVM):* The Support Vector Machine is particularly suited for high dimensional data analysis and non-linear classification problems. SVM tries to identify the hyperplane that best divides the classes in the given feature space while maximizing the distance between the closest data points, which belong to different classes, called as support vectors.
- *K-nearest neighbors (KNN):* The K Nearest Neighbors or KNN can be used to do classification of data or regression of the data. KNN practice examines the k nearest data points (neighbors) in

the input sample's feature space, with corresponding output determined by majority vote (classification), or averaging the values (regression) of these neighbors.

- *Naïve bayes*: The Naïve Bayes represents a family of probabilistic classification algorithms grounded in Bayes' Theorem, widely applied in machine learning for classification tasks. The word "naive" means that it is assumed that all the features (or attributes) in the given dataset are completely independent of one another, that is, they do not bear any relation with each other in the real world. This algorithm finds the posterior probability of the class given some features, and uses this probability for making predictions.
- *XGBoost* (*eXtreme Gradient Boosting)*: XGBoost is one of the most effective algorithms known to date that is primarily used in machine learning. It has proven itself effective for both classification and regression tasks in multiple classes. This is achieved by constructing an ensemble of decision trees one after another, where each of the trees corrects the mistakes of previous trees leading to an enhanced performance of the overall model.

Diagram

Figure 1 shows the flow that is used to train and test the ML algorithms used in this study.

RESULTS AND INFERENCES

This section presents the methodology used for developing and evaluating machine learning models for breast cancer detection. A flowchart illustrates the sequence of steps, from data preprocessing to model evaluation. The Breast Cancer Wisconsin (Diagnostic) Dataset [14], sourced from the UCI Machine Learning Repository, was used in this study. The applied machine learning algorithms include Logistic Regression, Decision Tree, Random Forest, Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Naive Bayes, and XGBoost. To assess model performance, metrics such as accuracy, precision, recall, F1-score, ROC-AUC curve, and confusion matrices were used to evaluate classification accuracy and robustness.

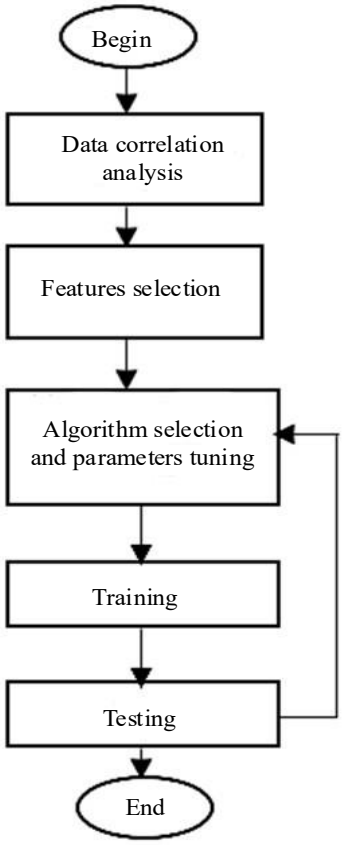

Figure 1. Flowchart for machine learning algorithms.

Table 1. Confusion matrix for binary classification.

	Predicted positive	**Predicted negative**
Actual Positive	Ture Positives (TP)	False Negatives (FN)
Actual Negative	False Positive (FP)	Ture Negatives (TN)

Confusion Matrix

A confusion matrix is an essential tool for evaluating the performance of classification models in machine learning. This is a matrix which compares the true class label for each instance with that predicted by the model, thus exhaustively illustrating the degree of correctness of the model. The matrix has four parts: True Positives (TP), or number of instances of the positive class that were correctly recognized by the model; True Negatives (TN), or number of instances of the negative class that were correctly recognized; False Positives (FP), or incorrectly classified instances of the positive class where the model erroneously predicts the positive class: these are also referred to as Type I error; and False Negatives (FN) or instances where the positive class is not detected: this is referred to as a Type II error. For binary classification, the confusion matrix can be represented as illustrated in Table 1.

ROC-AUC Curve

The Receiver Operating Characteristic curve (ROC) can be described as a visual representation of the efficiency of a classification model that makes a decision in yes or no. It shows the sensitivity also called as true positive rate (TPR) in relation to false positive rate (FPR) for different values of cut off scores. The Area Under the Curve (AUC) summarizes how well the model discriminates among the possible classes, with the AUC increasing with the ability to distinguish classes. For an ideal classifier, the area under the curve (AUC) will be 1; whereas AUC of 0.5 indicates no area under the curve thereby no power to distinguish classes from each other as in random guessing.

Accuracy

Accuracy, a key performance assessment score, is used to ascertain the count of success that classification models in machine learning attain. Specifically, it is the number of occurrences that are classified correctly (both positive and negative) divided by the total number of instances. Accuracy is a derived metric from the confusion matrix which involves four variables, these are: True Positives (TP), which entails the number of True Positives (TP) that the model forecasts accurately; True Negatives (TN) that involves the number of actual negatives correctly predicted; False Positives (FP) which explains the positives predicted but were actual negatives (Type I errors); and False Negatives (FN), which is the number of positives that were predicted negative but were actual positives (Type II errors).

Precision

Precision is one of the key performance indicators obtained from the confusion matrix and is utilized for assessing the probability of a classifier accurately predicting the positive class. As defined, precision is the ratio of the number of correctly identified positive cases which is also termed True Positives (TP) to the number of positive case predictions made which consists of True Positives (TP) and those predicted positive erroneously characterized as False Positives (FP).

Recall

Recall is a fundamental performance measure, and also sometimes referred to as sensitivity or true positive rate, derived from confusion matrix which enables assessing the ability of the classification model to capture all relevant positive cases. It is calculated as the proportion of easy or positive cases which were correctly predicted (TP) to the positions when the case was truly positive (made up of True Positives, TP and False Negatives, FN).

F1-Score

The F1 score is an essential measurement fixed in machine learning especially in classification model performance metrics in the presence of skewed datasets. More to the point, the F1 score is the weighted average of precision and recalls out of which one can use the two to define a score.

Logistic Regression

This confusion matrix displays the application of a logistic regression model within a binary classification task, presenting a high ratio in predicting "Benign" and "Malignant" classes as illustrated Figure 2. The model was able to label 70 instances of benign and 41 instances of malignant correctly, where only 1 "Benign" was predicted as "Malignant" and 2 "Malignant" instances were labeled as "Benign".

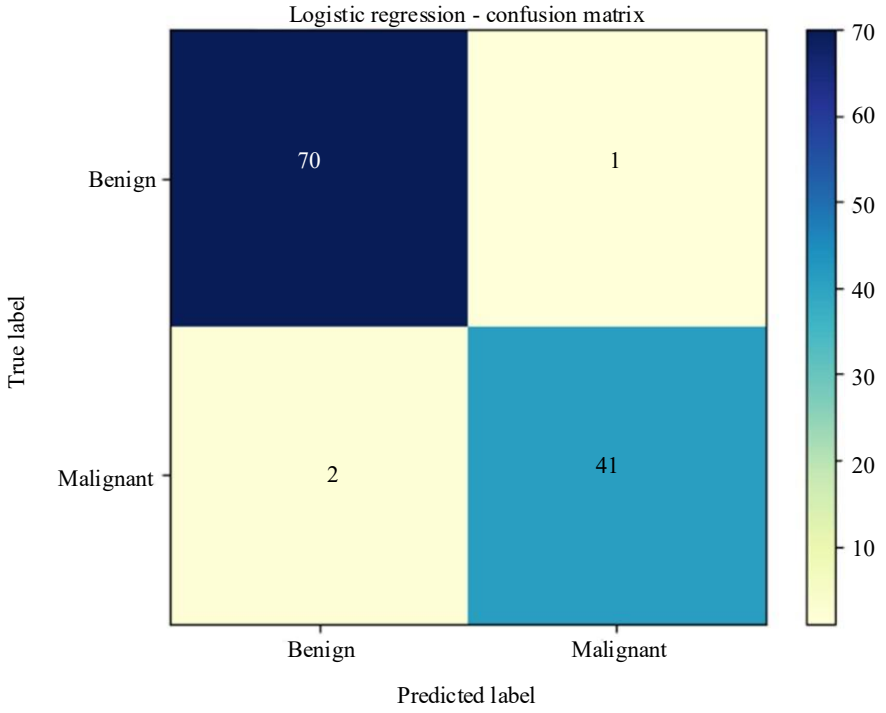

Figure 2. Logistic regression confusion matrix.

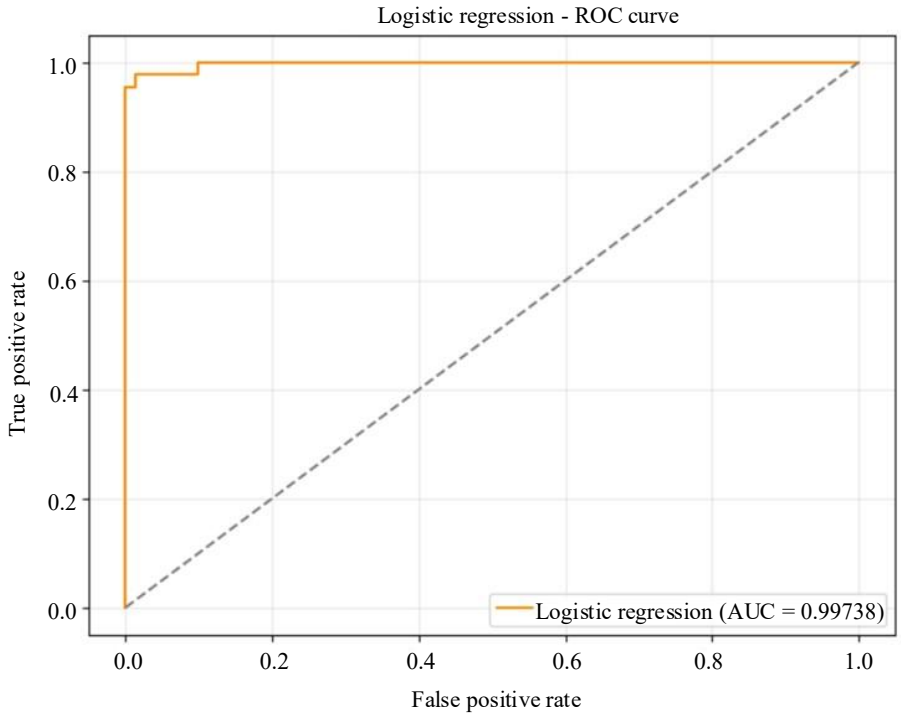

Figure 3. Logistic regression ROC curve.

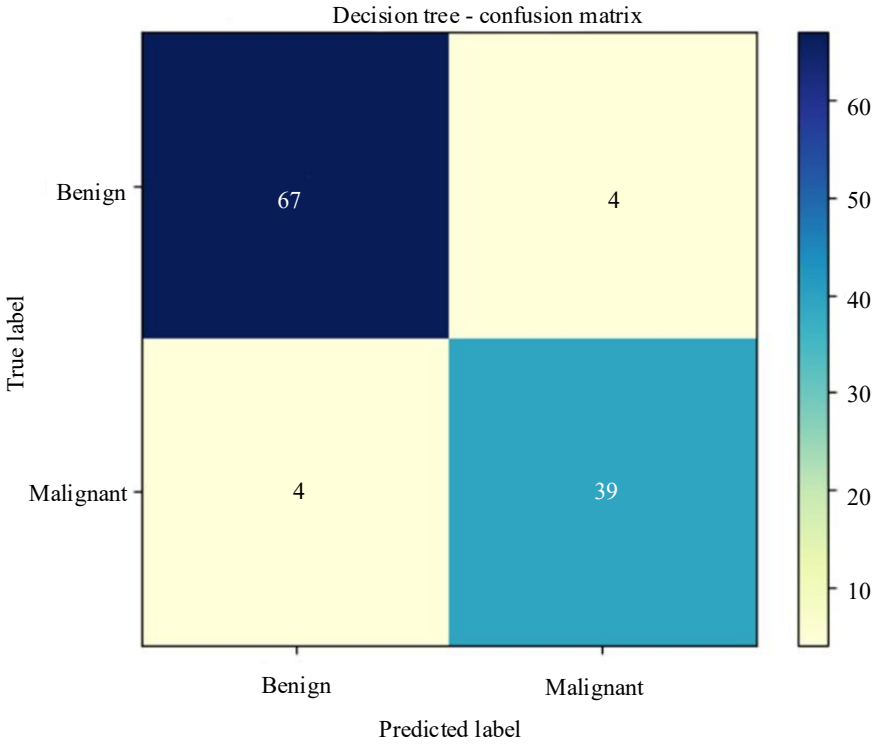

Figure 4. Decision tree confusion matrix.

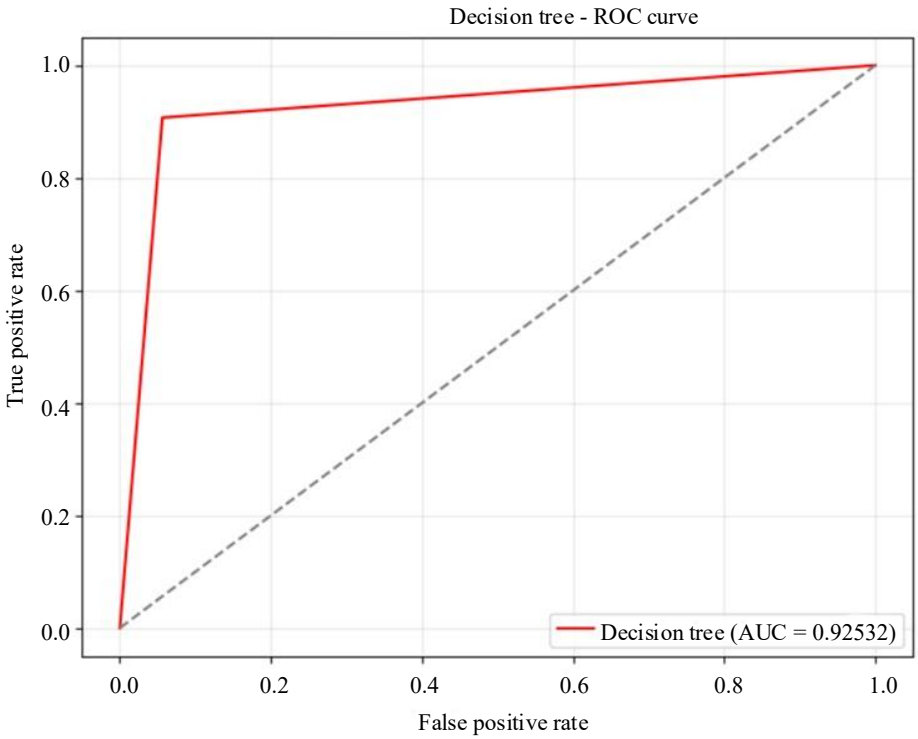

Figure 5. Decision tree ROC curve.

The AUC-ROC curve corresponding to the logistic Regression model shows a remarkable degree of accuracy, with a corresponding AUC of 0.997, which defers between the Two Categorical Classes: “Benign” and “Malignant". Likewise, the curve is positioned in the upper left-inset zone of the graph indicating a healthy true positive rate and a very low false as illustrated in Figure 3.

Decision Tree

The confusion matrix representing the decision tree model assesses the performance of correctly classifying "Benign" and "Malignant" cases. It managed to predict 67 "Benign" cases and 39 "Malignant" with no issues demonstrating its performance on both classes quite well as illustrated in Figure 4. Unfortunately, it also did not perform well on four "Benign" which were classified as false positives and four false negatives who were "Malignant" cases turned out to be "Benign".

The ROC curve's AUC of 0.925 indicates the Decision Tree Classifier's robustness as illustrated in Figure 5. As the particular model under evaluation exhibits a curve, it enhances the true positive rate without excessive costs on false positive rate, with the curve starting at the top left, the majority of the curve, is flat. Hence it can be said the Decision Tree is good at separation of classes for the quite large volume of data very efficiently.

Random Forest

In consideration of the confusion matrix provided for the Random Forest classifier, its performance appears to be satisfactory, and it is able to clearly discriminate benign cases quite easily as 70 of the 71 cases were correctly predicted, as illustrated in Figure 6. Malignant cases also show good performance as out of 43 cases, 40 were correctly identified. But there are few misclassifications to note, where 1 benign case was misclassified as malignant and three malignant cases were misclassified as benign.

This ROC curve of the Random Forest has the AUC is 0.996 as illustrated in Figure 7. The curve remains quite close to the top left corner indicating that the model is capable of performing class separability.

Support Vector Machine (SVM)

The model performs quite well on benign case predictions, yielding 71 true positives with a zero false positive count. In contrast, a total of 41 malignant cases were accurately assigned; however, there were two cases of false negative, that is, cases which were turned over as benign even though they were malignant, as illustrated in Figure 8.

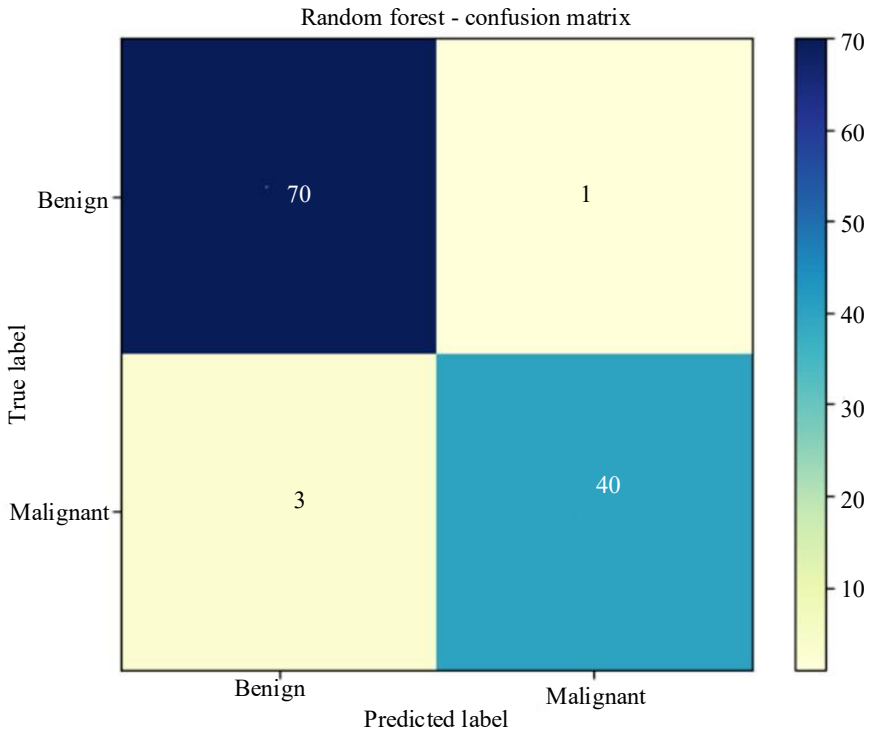

Figure 6. Random forest confusion matrix.

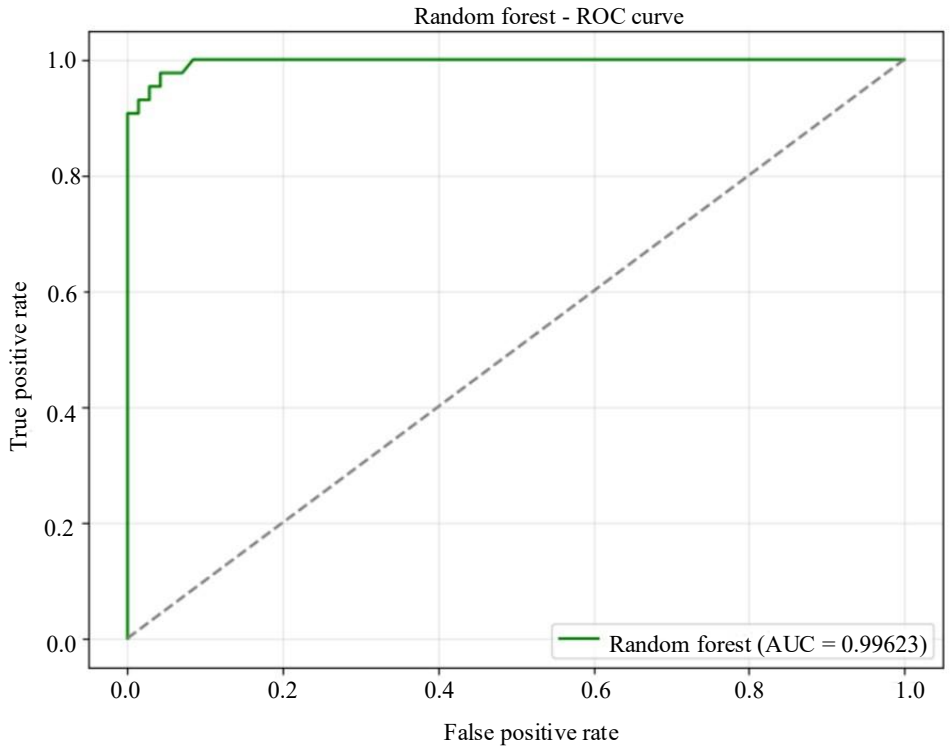

Figure 7. Random forest ROC curve.

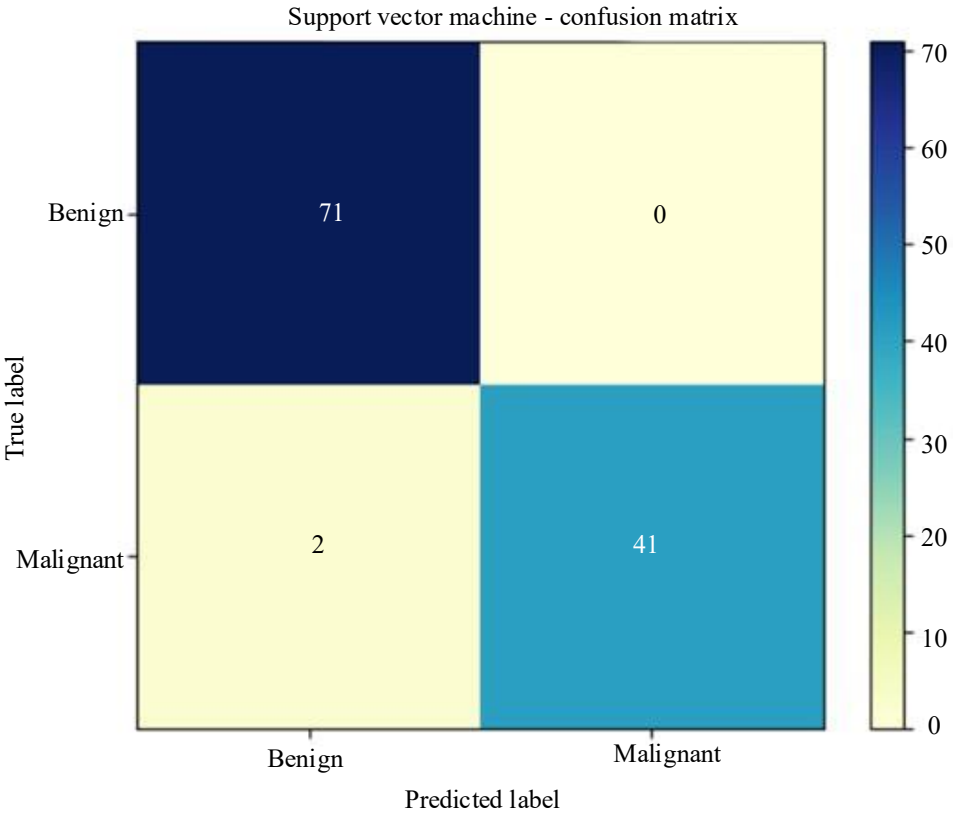

Figure 8. SVM confusion matrix.

The AUC value of 0.997 which is close to the value for perfect classification however it exhibits little overlap between benign and malignant classes, as illustrated in Figure 9.

K Nearest Neighbors (KNN)

The confusion matrix shown in Figure 10 illustrates how a (KNN) classifier performs for dual classification of breast cancer into benign and malignant cases. The predictive quality is high, being able to accurately predict 68 cases of benign cancers and 40 cases of malignant ones. Minor misclassifications do exist: three benign tumors are classified instead as malignant and three malignant tumors get identified as benign.

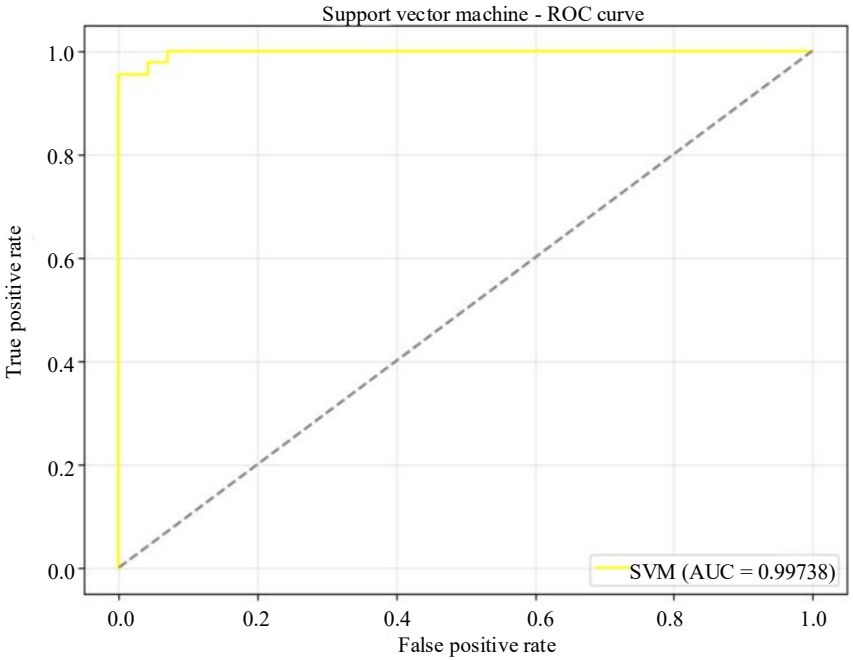

Figure 9. SVM ROC curve.

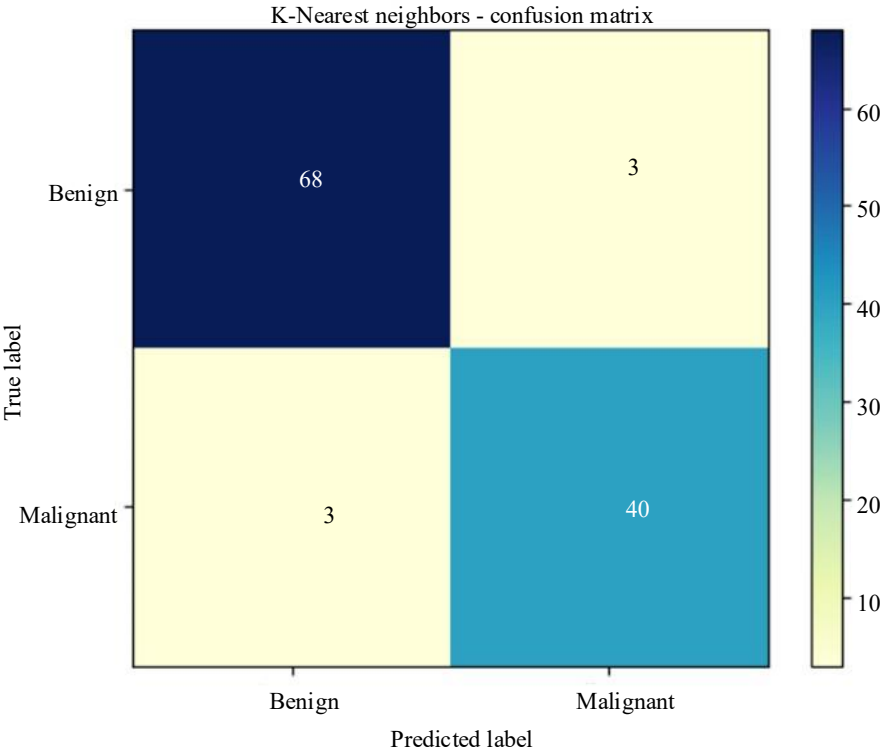

Figure 10. KNN confusion matrix.

The ROC graph for the K-Nearest Neighbors (KNN) classifier indicates that the performance of the classifier is outstanding in that the Area Under Curve (AUC) is almost equal to 0.98198 as illustrated in Figure 11. The steep rise in the curve near the y-axis implies that the true positive rate is very high at low rates of false positive which means that the sensitivity is very high.

Naïve Bayes

The Naive Bayes classifier's confusion matrix reveals that the classifier does a convincing job of predicting the breast cancer samples. The model correctly predicts 70 benign and 40 malignant samples as illustrated Figure 12. Nevertheless, the number of misclassifications is very low: one benign sample is misclassified as malignant and three malignant samples are misclassified as benign.

The AUC (Area Under the Curve) is 0.99738 which translates to the fact that the model scores high in accuracy and also the model is very class separable as illustrated in Figure 13.

XGBoost

It can be observed that the classifier performs quite well overall with 69 true negatives in case of benign finding and 40 true negatives in the case of malignant finding as illustrated in Figure 14. There are however five classification errors.

Looking at the ROC curve of the XGBoost classifier reveals remarkable results with AUC value of 0.99083 as illustrated in Figure 15.

Table 2 provides a comparison of studies that focus on the machine learning classification of breast cancer. Earlier studies which used Wisconsin datasets ranked the SVM as the leading algorithm with accuracies of 96.85 and 96.2% respectively. In the work done by us, where the Wisconsin Breast Cancer (Diagnostic) dataset was used, again SVM surpassed all the algorithms and the maximum accuracy achieved was 98.245%, proving that this algorithm is most effective for this classification task.

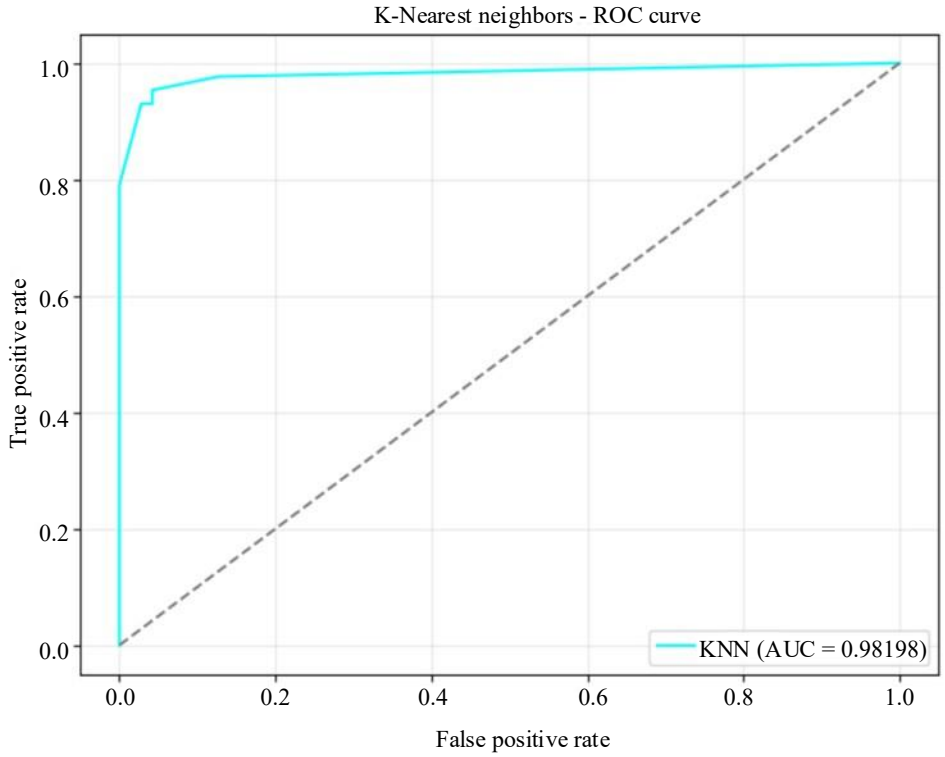

Figure 11. KNN ROC curve.

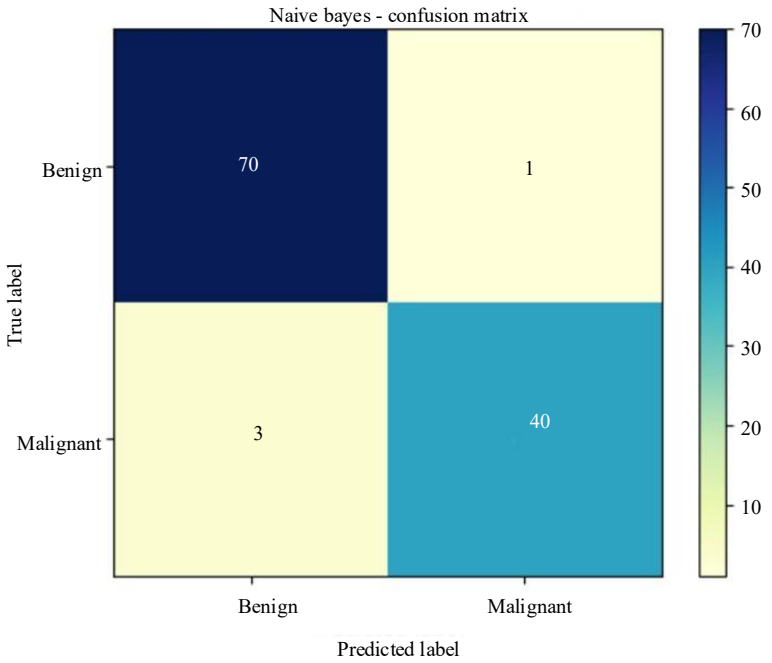

Figure 12. Naïve Bayes confusion matrix.

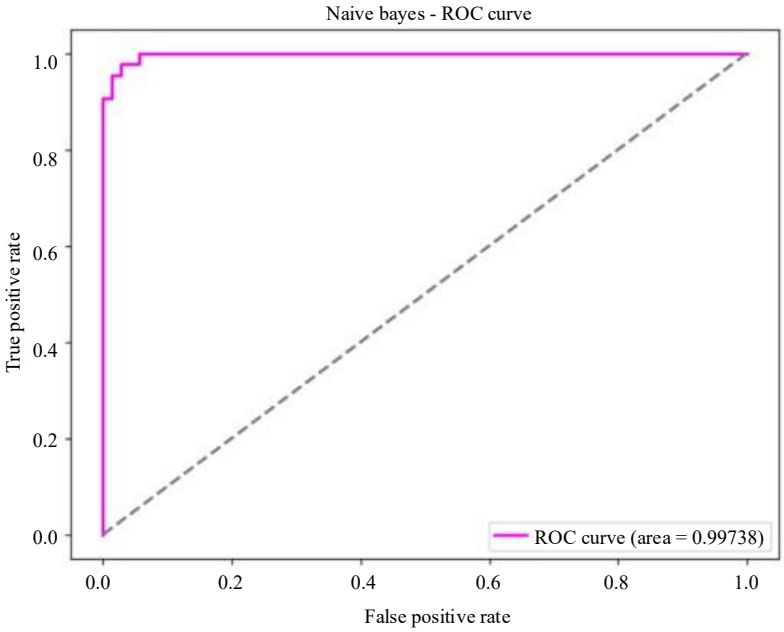

Figure 13. Naïve Bayes ROC curve.

Table 2. Performance Metrics Summary.

ML Algo	Accuracy	Precision	F1Score	Recall
Logistic Regression	97.368%	97.619%	96.470%	95.348%
Decision Tree	92.982%	90.697%	90.697%	90.697%
Random Forest	96.491%	97.560%	95.238%	93.023%
Support Vector Machine	98.245%	100.000%	97.619%	95.348%
K-Nearest Neighbors	94.736%	93.023%	93.023%	93.023%
Naive Bayes	96.491%	97.560%	95.238%	93.023%
XGBoost	95.614%	95.238%	94.117%	93.023%

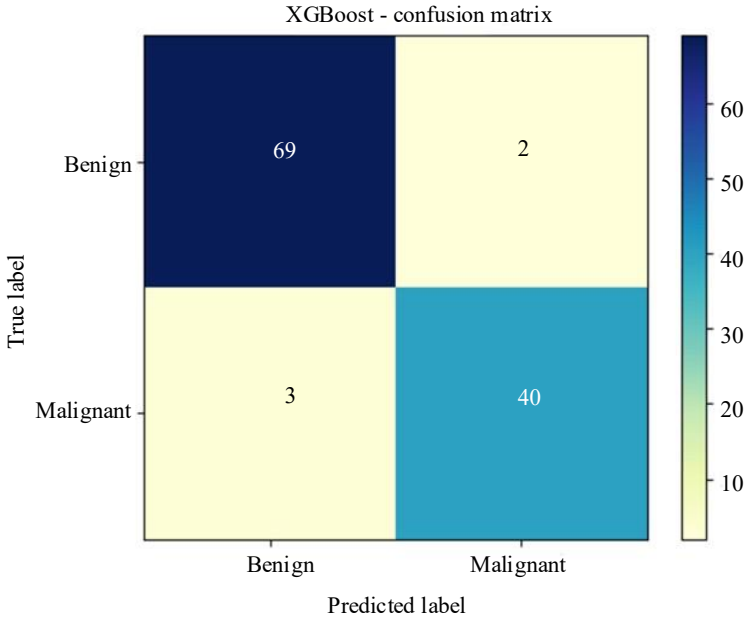

Figure 14. XGBoost confusion matrix.

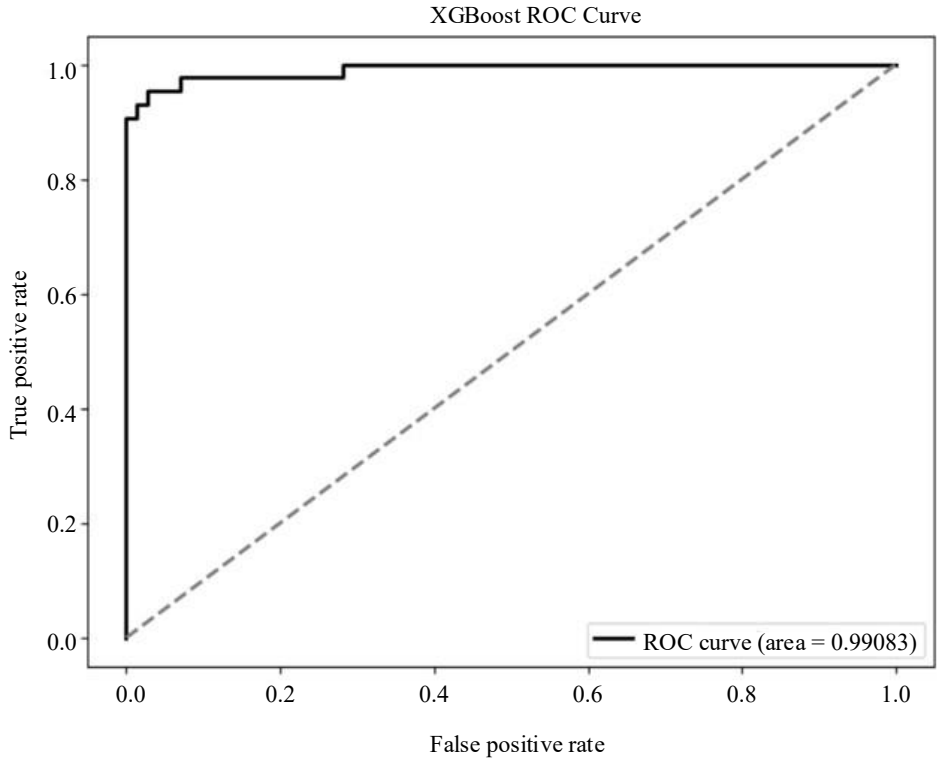

Figure 15. XGBoost ROC curve.

CONCLUSION

This research presents an extensive review of the performance of various machine learning algorithms on breast cancer detection using the Wisconsin Diagnostic Dataset. The Support Vector Machine (SVM) was the most effective algorithm in this set of evaluations since it was able to achieve the highest accuracy of 98.245% thus proving its great discriminating power when it comes to malignant and benign cases. This is because SVM can create the best decision surfaces in wide dimension feature spaces, thus making it ideal for the given diagnosis.

Logistic regression employed similar metrics and was equally interpretable, thus providing a realistic approach to less complicated diagnosis situations. It is well known that Decision Tree is simple and easy to understand, however its accuracy and precision proved to be low and this challenge was solved by Random Forest, XGBoost and other similar approaches. K-Nearest Neighbors (KNN) was a viable option but due to its inefficiency in dealing with large data sets, scalability and computation problems, it performed poorly when compared with other models. On the other hand, SVM had a better performance than these algorithms since it was capable of dealing with both linear and nonlinear relations with equal efficiency, which was crucial in cases with many features. Naive Bayes, despite its simplicity and computational efficiency, was constrained by its assumptions of feature independence, which limited its performance in handling correlated attributes within the dataset.

Looking ahead, there is scope for improvement by combining advanced feature selection methods and hybrid modeling in order to tackle the issues of diagnostic accuracy, class imbalance, real-world clinical interpretation of the results, as well as dataset overfitting.

REFERENCES

1. Ahmed MT, Imtiaz MN, Karmakar A. Analysis of Wisconsin Breast Cancer original dataset using data mining and machine learning algorithms for breast cancer prediction. J Sci Technol Environ Inform. 2020; 9(2): 665–72.
2. Akbugday B. Classification of breast cancer data using machine learning algorithms. In 2019 IEEE Medical technologies congress (TIPTEKNO). 2019 Oct 3; 1–4.
3. Kadhim RR, Kamil MY. Comparison of breast cancer classification models on Wisconsin dataset. Int J Reconfigurable & Embedded Syst. 2022; 11(2): 166–174.
4. Tahmooresi M, Afshar A, Rad BB, Nowshath KB, Bamiah MA. Early detection of breast cancer using machine learning techniques. Journal of Telecommunication, Electronic and Computer Engineering (JTEC). 2018 Sep 26; 10(3–2): 21–7.
5. Sengar PP, Gaikwad MJ, Nagdive AS. Comparative study of machine learning algorithms for breast cancer prediction. In 2020 IEEE Third International Conference on Smart Systems and Inventive Technology (ICSSIT). 2020 Aug 20; 796–801.
6. Street WN, Wolberg WH, Mangasarian OL. Nuclear feature extraction for breast tumor diagnosis. In SPIE Biomedical image processing and biomedical visualization. 1993 Jul 29; 1905: 861–870.
7. Vig L. Comparative analysis of different classifiers for the Wisconsin breast cancer dataset. Open Access Library Journal (OALib Journal). 2014 Sep 1; 1(6): 1–7.
8. Bazazeh D, Shubair R. Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In 2016 IEEE 5th international conference on electronic devices, systems and applications (ICEDSA). 2016 Dec 6; 1–4.
9. Shahnaz C, Hossain J, Fattah SA, Ghosh S, Khan AI. Efficient approaches for accuracy improvement of breast cancer classification using wisconsin database. In 2017 IEEE region 10 humanitarian technology conference (R10-HTC). 2017 Dec 21; 792–797.
10. Obaid OI, Mohammed MA, Ghani MK, Mostafa A, Taha F. Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer. Int J Eng Technol. 2018 Dec; 7(4.36): 160–6.
11. Austria YD, Jay-ar PL, Maria Jr LB, Goh JE, Goh ML, Vicente HN. Comparison of machine learning algorithms in breast cancer prediction using the coimbra dataset. Int. J. Simulat. Syst. Sci. Tech. 2019; 7(10): 23.1–23.8.
12. Omondiagbe DA, Veeramani S, Sidhu AS. Machine learning classification techniques for breast cancer diagnosis. In IOP Conf Ser: Mater Sci Eng. IOP Publishing. 2019 Jun 7; 495: 012033.
13. Mushtaq Z, Yaqub A, Sani S, Khalid A. Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets. J Chin Inst Eng. 2020 Jan 2; 43(1): 80–92.
14. Naji MA, El Filali S, Aarika K, Benlahmar EH, Ait Abdelouhahid R, Debauche O. Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Comput Sci. 2021 Jan 1; 191: 487–92.

15. Sakib S, Yasmin N, Tanzeem AK, Shorna F, Md. Hasib K, Alam SB. Breast cancer detection and classification: A comparative analysis using machine learning algorithms. In Proceedings of Third International Conference on Communication, Computing and Electronics Systems: ICCCES 2021. Singapore: Springer Singapore; 2022 Mar 20; 703–717.
16. Khan RH, Miah J, Rahman MM, Tayaba M. A comparative study of machine learning algorithms for detecting breast cancer. In 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC). 2023 Mar 8; 647–652.
17. Kushwaha V. Breast cancer diagnostic using machine learning: applying supervised learning techniques to Coimbra and Wisconsin datasets. Finland: Lappeenranta–Lahti University of Technology LUT; 2023.
18. Hossin MM, Shamrat FJ, Bhuiyan MR, Hira RA, Khan T, Molla S. Breast cancer detection: an effective comparison of different machine learning algorithms on the Wisconsin dataset. Bull Electr Eng Info. 2023 Aug 1; 12(4): 2446–56.

