Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 20 No. 13 (2024) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v20i13.50289

PAPER

An Efficient Breast Cancer Detection Using Machine Learning Classification Models

B. N. Ravi Kumar[1], Naveen Chandra Gowda[2], B. J. Ambika[3](✉), H. N. Veena[4], B. Ben Sujitha[5], D. Roja Ramani[6]

[1]Department of Information Science and Engineering, BMS Institute of Technology and Management, Bengaluru, Karnataka, India

[2]School of Computer Science and Engineering, REVA University, Bengaluru, Karnataka, India

[3]Department of Computer Science and Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, Karnataka, India

[4]Department of Computer Science and Engineering, SJB Institute of Technology, Bengaluru, Karnataka, India

[5]Department of Computer Science and Engineering, Noorul Islam Centre for Higher Education, Kanyakumari, Tamil Nadu, India

[6]Department of Computer Science and Engineering, New Horizon College of Engineering, Bengaluru, Karnataka, India

ambika.bj@manipal.edu

ABSTRACT

Breast cancer is still a dangerous and common disease that affects women all over the world, which highlights how crucial early identification is to better patient outcomes. In recent years, utilizing machine learning (ML) algorithms has improved accuracy and efficiency dramatically in a variety of applications, showing promising outcomes. This article provides a novel machine-learning approach to increase the accuracy of breast cancer detection. To improve diagnostic efficiency and accuracy, our suggested methodology combines sophisticated feature selection strategies, reliable classification algorithms, and enhanced model training methodologies. We investigated several ML classifiers, and after thorough hyperparameter tuning, the models were. Random forest and gradient boosting have achieved the highest performance with an accuracy of 97.90% and an ROC score of 0.99. This research highlights the effectiveness of ML, particularly the random forest algorithm, in breast cancer diagnosis and prognosis. Future work may explore deep learning techniques for determining the disorder's severity.

KEYWORDS

women health, breast cancer, machine learning (ML), classification algorithms

1 INTRODUCTION

A worldwide epidemic, cancer affects a wide range of populations. Because breast cancer affects many women, research on diagnosis and prognosis needs to be concentrated. Machine learning (ML) for early prediction is promising. It is second in female mortality after lung cancer and is frequently associated with advanced age. It is caused by abnormal proliferation of breast cells [1]. Breast cancer, a multifaceted ailment, ranks as the most prevalent cancer among women globally [2]. Roughly 30% of female cancer cases stem from it, with 1.5 million women diagnosed annually, causing 500,000 deaths worldwide [3]. Despite its rising incidence over three decades, mortality has declined, attributed 20% to mammography screening and 60% to enhanced cancer therapies [4].

Ravi Kumar, B.N., Gowda, N.C., Ambika, B.J., Veena, H.N., Sujitha, B.B., Ramani, D.R. (2024). An Efficient Breast Cancer Detection Using Machine Learning Classification Models. *International Journal of Online and Biomedical Engineering (iJOE)*, 20(13), pp. 24–40. https://doi.org/10.3991/ijoe.v20i13.50289

Article submitted 2024-05-28. Revision uploaded 2024-07-19. Final acceptance 2024-07-19.

When it comes to identifying minor abnormalities related to breast cancer, diagnostic mammography is essential. However, its effectiveness wanes when assessing suspected regions among the multitude of photos, especially when dealing with dense breast tissue, where approximately 50% of tumors escape detection. The necessity of early detection is shown by the fact that 25% of women with breast cancer obtain a negative diagnosis within two years of screening [5]. However, the traditional method of imposing set screening intervals on all women turns out to be ineffectual on a personal level, which could jeopardize screening initiatives. Including additional risk variables in addition to mammography may improve accuracy, assist with customized diagnoses, and identify individuals who are at high risk [6].

Machine learning, which is becoming more and more common in various applications [7–13], including the medical field, has the potential to predict breast cancer by utilizing a variety of data sources, such as genetic, mammographic, and demographic information. Authors in [30] and [31] have proposed ML mechanisms for lung cancer. Developing ML models that enhance the detection of minor abnormalities and improve early diagnosis accuracy, especially in cases where traditional methods may fail. However, difficulties remain in developing thorough models that include all pertinent risk factors; existing models frequently result in over-screening and psychological stress for patients [14]. By Incorporating additional risk variables beyond mammography to personalize breast cancer risk assessments. This includes leveraging genetic, mammographic, and demographic data to tailor screening strategies based on individual risk profiles. Multi-factorial models that incorporate laboratory, mammography, and demographic data are necessary for effective risk prediction of breast cancer and offer increased assessment precision. With a wide range of parameters taken into account throughout the modeling process, this study aims to predict breast cancer risk using a variety of ML techniques [15]. Optimize ML algorithms to predict breast cancer risk effectively. Focus on parameter selection and model tuning to achieve reliable predictions that can be translated into clinical practice. The present investigation aims to address these challenges by leveraging ML techniques to advance the accuracy of breast cancer diagnosis and prognosis. By focusing on enhancing early detection, personalizing risk assessment, and optimizing model complexity, the study seeks to contribute to improved outcomes and personalized care in breast cancer management.

Clinical, lifestyle, and socioeconomic factors all play a role in the incidence of breast cancer in women. The potential for ML to reveal hidden data signals to forecast this illness is promising. These developments aim to promote healthcare technology and research, lessen their worldwide impact, and enhance patient outcomes. This study integrates laboratory, mammography, and demographic data to predict breast cancer using a variety of ML techniques. Major contributions in the paper are:

- Using pre-trained models, this research uses ML approaches to extract features and create a strong predictive framework.
- It is a rigorous performance analysis that uses careful assessment metrics to confirm the effectiveness of the suggested model.
- The results of this study have the potential to completely transform patient care by providing medical professionals with data-driven insights to help them make decisions about the diagnosis and treatment of breast cancer.
- The suggested model performs well when compared to relevant past efforts.

The remainder of the study is organized as Section 2 provides an overview of the study on breast cancer detection found in the literature. The proposed work is presented in Section 3. The implementation, along with the outcomes and conversations that followed, are presented in Section 4. Section 5 contains the paper's conclusion.

2 LITERATURE REVIEW

The prediction of breast cancer has been transformed by ML algorithms; yet, choosing the best classifier remains a challenge. Many studies using different algorithms on medical datasets have shown encouraging results. Using a variety of ML approaches, researchers have worked ceaselessly to create and evaluate breast cancer detection systems [16]. The author used several classifiers and carried out a comparison study in a recent paper. Notably, support vector machine (SVM) produced a maximum accuracy of 97% when used without quick co-relation-based streamlines [17]. Maximum perimeter and texture classification were also included in the logistic regression, which was used for categorization with an impressive 95% accuracy [18]. Additionally, research has concentrated on identifying and describing cell structures, contrasting various categorization and clustering techniques, and establishing a connection between histopathological evaluation and fine needle aspiration cytology [19]. Furthermore, studies in other fields, such as the classification of thunderstorms and diabetes, have developed methods. Furthermore, cutting-edge algorithms such as adaptive resonance theory have been created expressly for the study of breast cancer, demonstrating the ongoing progress in this vital area [20].

In breast cancer detection, the use of morphological and textural features for feature extraction has been common. Significant improvements in patient outcomes throughout therapy have been observed with deep convolutional neural networks (CNNs), demonstrating their amazing potential in early-stage diagnosis. As writers in [21] investigated, the method of forecasting non-communicable diseases (NCDs) included using several algorithms. They used 10-fold cross-validation to assess different categorization algorithms on eight different NCD datasets, and the area under the curve was used as a precision parameter. Algorithms such as K-nearest neighbor (KNN), SVM, and neural networks (NN) showed resilience in the face of irrelevant features and noisy data in the datasets, with preprocessing strategies proposed to improve accuracy and reduce unnecessary attributes.

Approaches to natural inspiration computing (NIC) have shown promise as diagnostic instruments for human health issues. The authors of [22] presented diagnostic algorithms that were derived from insects and demonstrated effectiveness in the diagnosis of conditions such as diabetes and cancer, as well as tumors of the breast, lung, prostate, and ovary. Directed artificial bee colonies (ABC) combined with NN allowed for more accurate diagnosis of leukemia and diabetes in addition to breast cancer. NNs were emphasized in [23] for their promise in the categorization of cancer, especially in its early stages, even if image preprocessing requires a significant amount of processing power. Future research aims to reduce computing hurdles in medical imaging by utilizing artificial intelligence (AI) and convolutional neural networks.

A ML approach created in [24] greatly improves breast cancer diagnosis and survival prediction. In clinical settings, their technology demonstrated good accuracy and dependability, indicating its potential usefulness in supporting medical personnel. In a similar vein, [25] introduced a hybrid strategy that combined ensemble learning with deep feature extraction, improving detection rates and decreasing false positives. The author in [26] emphasized the significance of feature selection and data preprocessing while highlighting the efficacy of several approaches in the early identification and prevention of breast cancer. An adaptive voting ensemble approach was presented in [27] that enhanced classification performance by utilizing the advantages of several models, resulting in increased accuracy and robustness.

The authors [24] pointed out that the intricacy of their model and the requirement for big, varied datasets to guarantee generalizability could lead to overfitting. The computational complexity of the [25] hybrid technique was highlighted as a potential barrier to its scalability and real-time implementation in 2023. Although their methods showed promise, [26] noted that further validation in various clinical contexts and populations was necessary. In order to strike a compromise between computational cost and performance, [27] noted that their adaptive voting ensemble algorithm needs to be further optimized. To ensure practical application, they also stressed the significance of integrating their technique with clinical workflows.

3 PROPOSED WORK

The proposed method, depicted in Figure 1, describes a system that consists of concepts implemented through efficient feature selection that helps in comprehending, educating, or estimating the risk of breast cancer. The proposed model consists of six basic steps: (i) Data collection, which involves gathering data from multiple sources with varying parameters. (ii) Pre-processing the data. Eliminate the dataset's outliers as well. (iii) Data splitting for validation, testing, and training. (iv) To validate the classification findings, the classification model employs ML classifiers. (v) Model assessment using metrics for performance evaluation. (vi) The k-fold mechanism is used for model validation.

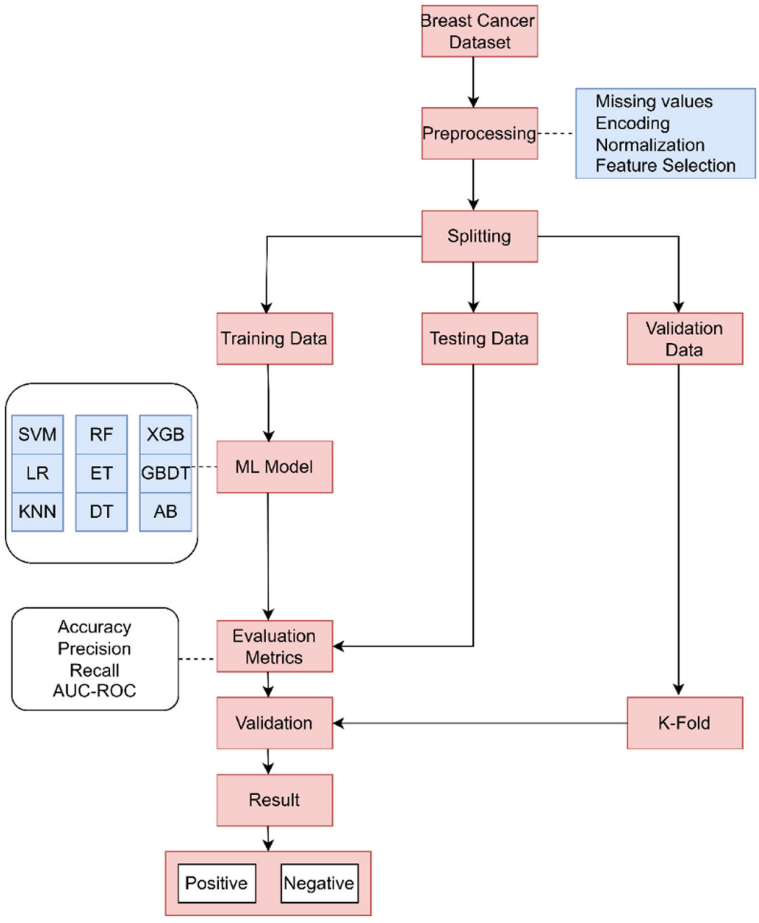

Fig. 1. Proposed system for breast cancer detection

3.1 Dataset description

The publicly accessible datasets have been used by us for research and testing [28]. The 569 × 32 dataset has 32 factors, and each feature is used to classify a person's behavior as either influenced or not. The list of features considered in the dataset is presented in Table 1.

Table 1. Features in the dataset with description

Si. No	Feature	Description
1	id	A distinct number assigned to every patient
2	diagnosis	The diagnosis outcome, usually indicated by a letter ('M' for malignant and 'B' for benign)
3	r_mean	Average of the distances between the center and the outermost points
4	t_mean	Standard deviation of values in grayscale
5	p_mean	The core tumor's average size
6	a_mean	average size of the tumor
7	s_mean	Average of the regional variance in radius lengths
8	c_mean	Average of area/perimeter^2 – 1.0
9	concavity_mean	Average degree of the contour's concave sections
10	cp_mean	The average of the contour's concave sections
11	sym_mean	Tumor's mean symmetry
12	fd_mean	The "coastline approximation" mean is –1.
13	r_se	Standard deviation of the mean radius
14	t_se	The mean texture's standard error
15	p_se	The mean perimeter's standard error
16	a_se	The mean area's standard error
17	s_se	Error standard for the average smoothness
18	c_se	The average standard error of the compactness
19	concavity_se	Value of the mean concavity standard error
20	cp_se	The standard error for the average number of the contour's concave sections
21	sym_se	Value of the mean symmetry standard error
22	fd_se	Value of the mean fractal standard error
23	r_worst	"Worst" or largest mean value for the radius;
24	t_worst	"Worst" or largest mean value for the texture;
25	p_worst	"Worst" or largest mean value for the perimeter;
26	a_worst	"Worst" or largest mean value for the area;
27	s_worst	"Worst" or largest mean value for the smoothness;
28	c_worst	"Worst" or largest mean value for the compactness;
29	concavity_worst	"Worst" or largest mean value for the concavity;
30	cp_worst	"Worst" or largest mean value for the number of concave portions of the contour;
31	sym_worst	"Worst" or largest mean value for the symmetry;
32	fd_worst	"Worst" or largest mean value for the fractal dimension

3.2 Data pre-processing

It's one technique for converting raw data into a format that can be understood and utilized. Sometimes there are a lot of nulls and errors in real-world data, which leaves it unfinished and unformatted. Handling missing values, feature selection, encoding, and data normalization have been taken into consideration as part of pre-processing.

Feature selection: Because feature selection is so important, sequential forward selection is employed. This method has been used to exclude a number of low-significance features from the dataset. We have selected only 23 features for training and validation from the set of features depicted in Table 1, excluding features such as 'id, radius mean, perimeter mean, area-mean, concavity mean, radius se, area se, radius worst, texture worst, and radius worst' because we believe they have minimal impact on classification.

Missing values: This step entails performing an exploratory data analysis in order to locate and address the abnormalities. The missing values were addressed by means of an iterative imputer. Every feature is molded depending on the other features using the iterative imputation method.

Encoding: Encoding in the pre-processing stage is a crucial step in preparing data, working with ML models, and other data analysis tasks. It involves transforming data into a format that can be efficiently processed by algorithms. Here are some common types of encoding used in pre-processing.

Data normalization: Different types of data fields may be present in the dataset. Effective categorization requires that the data values be decoded using the same object type. The parameter values are scaled between zero and one so that the numeric column values are adjusted without deleting any numbers or altering the range of possible values before being placed on a standard scale.

3.3 Dataset splitting

For testing and training purposes, the complete patient dataset is now split into two halves. The data were used in a 75:25 percent ratio for testing and training. Thus, 426 of the total dataset instances are designated for training, while 143 are designated for testing. The suggested method of validation is k-fold cross-validation, where k equals five.

3.4 Classification model

To screen and identify breast cancer, we have leveraged the capabilities of multiple well-established ML algorithms. Among these is the random forest classifier, which is renowned for its resilience and capacity to work with intricate datasets. When many models are merged to increase predictive accuracy in ensemble learning, the gradient boosting classifier, XG Boost classifier, and AdaBoost classifier perform exceptionally well. Both linear and non-linear classification tasks can be effectively handled by the SVM, while logistic regression provides simplicity and interpretability.

We have also made use of the extra trees classifier because of its strong overfitting resistance and great computational efficiency. Similarity metrics are used by the KNN method to categorize new instances according to how close they are to preexisting data points. In the meantime, the decision tree classifier uses a feature-split

structure such as a tree to provide easy decision-making. We hope to increase the precision and dependability of breast cancer screening and detection by utilizing this wide range of ML algorithms, which will ultimately lead to better patient outcomes and more effective healthcare.

3.5 Model evaluation metrics

This section explains and provides examples of the evaluation metrics. Evaluating the predictive model's performance is necessary to ascertain how well it achieves a goal. Using performance assessment metrics on the test dataset, the classification model's efficacy and performance are assessed. The true positive (TP), false positive (FP), true negative (TN), and false negative (FN) are used to define the evaluation metrics [29].

True positive: This refers to the number of instances in which the test accurately detects the existence of breast cancer. Stated differently, the patient does have breast cancer, and the test result is positive.

False positives, or FPs, are the number of instances in which a test results in an inaccurate diagnosis. The patient does not have breast cancer, despite the positive test result. FP results may cause patients who are truly healthy to undergo needless stress, additional testing, and potentially hazardous therapies.

The number of cases in which the test accurately determines that breast cancer is not present is known as TN. In this case, the patient does not have breast cancer, and the test result is negative.

False negative: This refers to the number of instances in which the test misidentifies the absence of breast cancer. The patient has breast cancer even though the test results are negative. Because they might cause delays in diagnosis and treatment and even impair the patient's prognosis, FN results are especially harmful.

The evaluation metrics used to determine the model's performance are accuracy, precision, recall, F1 score, and area under curve (AUC-ROC); weighted average metrics are used to quantify error.

3.6 Model validation

The validation procedure employs the K-fold validation technique. The K-fold method trains and tests on the complete dataset. This means that 75% of the dataset is utilized for training, 25% is used to test the system using the pertinent test case, and the actual, complete dataset is used to confirm and verify the results.

4 EXPERIMENTAL RESULTS

Google Co Labs was utilized for doing the experimentation. The data set is loaded using Pandas, and the Python packages are plotted using pilots. The ML processes are also implemented using Python. For the experiments to effectively operate and validate the suggested model, a Windows 10 PC with specs including a CPU speed of 2.9 GHz, core i7, RAM of 8 GB, a GPU of 620, and a 5 GB drive was utilized.

A traditional ML methodology consisting of RF, GBDT, XG Boost, LR, SVM, E Trees, ABoost, KNN, and DT was used to assess breast cancer. We have experimented with the dataset, as stated in Section 3.1, and the outcomes are documented following 25 iterations. All the algorithms are experimented with based on the hyperparameters as mentioned in Table 2.

Table 2. Hyperparameters used in the model

Classifier	Optimal Parameter Values
Random Forest Classifier	max_depth = 10, min_samples_leaf = 2, max_features = 0.5, min_samples_split = 3, criterion = 'entropy', n_estimators = 130
Gradient Boosting Classifier	learning_rate = 0.1, loss = 'exponential', n_estimators = 180
XGBoost Classifier	learning_rate = 0.01, n_estimators = 180, max_depth = 5
Logistic Regression	Default values
Support Vector Machine	C = 15, gamma = 0.01, probability = True
Extra Trees Classifier	n_estimators = 5, max_features = 2, criterion = 'entropy'
AdaBoost Classifier	n_estimators = 50
K-Nearest Neighbor	Default values
Decision Tree Classifier	max_depth = 15, min_samples_split = 5, min_samples_leaf = 4, splitter = 'random', criterion = 'entropy'

4.1 Performance assessment of machine learning models

Throughout our thorough investigation, we have meticulously measured several performance indicators, including F1 score, accuracy, precision, and AUROC. When taken as a whole, these indicators provide a thorough picture of the ML model's effectiveness across numerous areas. Nonetheless, one significant discovery is brought to light by the results: the model performs noticeably differently across datasets. This highlights how important it is to consider dataset characteristics in model evaluation, necessitating rigorous model selection and fine-tuning to get optimal outcomes across a range of circumstances. A range of performance metrics are incorporated to facilitate a comprehensive comprehension of the efficacy of the classifiers, hence promoting a data-driven methodology for model selection and enhancement. Table 3 provides an overview of the empirical performance evaluation of classifiers built using traditional ML algorithms.

Table 3. Classification reports of the considered algorithms

		Class	Precision	Recall	f1-Score
DT		0	0.92	0.96	0.94
		1	0.92	0.87	0.89
	accuracy				0.92
	macro avg		0.92	0.91	0.92
	weighted avg		0.92	0.92	0.92
KNN		0	0.95	0.98	0.96
		1	0.96	0.91	0.93
	accuracy				0.95
	macro avg		0.95	0.94	0.95
	weighted avg		0.95	0.95	0.95

(Continued)

Table 3. Classification reports of the considered algorithms *(Continued)*

		Class	Precision	Recall	f1-Score
AB		0	0.99	0.94	0.97
		1	0.91	0.98	0.95
	accuracy				0.96
	macro avg		0.95	0.96	0.96
	weighted avg		0.96	0.96	0.96
ET		0	0.95	0.99	0.97
		1	0.98	0.91	0.94
	accuracy				0.96
	macro avg		0.96	0.95	0.95
	weighted avg		0.96	0.96	0.96
SVM		0	0.96	0.98	0.97
		1	0.96	0.92	0.94
	accuracy				0.96
	macro avg		0.96	0.95	0.95
	weighted avg		0.96	0.96	0.96
LR		0	0.96	0.98	0.97
		1	0.96	0.92	0.94
	accuracy				0.96
	macro avg		0.96	0.95	0.95
	weighted avg		0.96	0.96	0.96
XGB		0	0.99	0.97	0.98
		1	0.95	0.98	0.96
	accuracy				0.97
	macro avg		0.97	0.97	0.97
	weighted avg		0.97	0.97	0.97
GB		0	0.99	0.98	0.98
		1	0.96	0.98	0.97
	accuracy				0.98
	macro avg		0.98	0.98	0.98
	weighted avg		0.98	0.98	0.98
RF		0	0.98	0.99	0.98
		1	0.98	0.96	0.97
	accuracy				0.98
	macro avg		0.98	0.98	0.98
	weighted avg		0.98	0.98	0.98

It is clear that both random forest and gradient boosting classifiers performed better in terms of accuracy during the testing phase than other ML algorithms based on the recorded training and testing accuracies after 25 iterations. The following Table 4 provides a summary of the findings. These results imply that, when

compared to other methods, random forest and gradient boosting classifiers perform better in accurately predicting the course of breast cancer. This demonstrates how effective boosting strategies and ensemble-based approaches are at raising the precision of ML models used to identify breast cancer.

An extensive analysis of the confusion matrices for every model in the dataset provides crucial information on the particular characteristics and challenges connected with each model. It's interesting to note that Table 4 indicates that there may be an unequal distribution of classes. Classifiers may find it challenging to accurately forecast minority classes as a result of this imbalance, which could lead to biased results. Additionally, the confusion matrix demonstrates how well the accuracy predictions for the dataset align. In this instance, the balanced distribution shows that the model effectively controls class proportions, which contribute to the adolescent dataset's reliable prediction-making. These findings emphasize how important it is to consider both overall accuracy and the distribution of predictions across classes, particularly in datasets where class imbalances already exist.

Table 4. Accuracy and confusion matrix comparison of classifiers

Model	Training Accuracy	Testing Accuracy	Confusion Matrix
Random Forest Classifier	99.30	97.90	$\begin{bmatrix} 89 & 1 \\ 2 & 51 \end{bmatrix}$
Gradient Boosting Classifier	100.00	97.90	$\begin{bmatrix} 88 & 2 \\ 1 & 52 \end{bmatrix}$
XGBoost Classifier	99.30	97.20	$\begin{bmatrix} 87 & 3 \\ 1 & 52 \end{bmatrix}$
Logistic Regression	99.06	95.80	$\begin{bmatrix} 88 & 2 \\ 4 & 49 \end{bmatrix}$
Support Vector Machine	98.83	95.80	$\begin{bmatrix} 88 & 2 \\ 4 & 49 \end{bmatrix}$
Extra Trees Classifier	100.00	95.80	$\begin{bmatrix} 89 & 1 \\ 5 & 48 \end{bmatrix}$
AdaBoost Classifier	100.00	95.80	$\begin{bmatrix} 85 & 5 \\ 1 & 52 \end{bmatrix}$
K-Nearest Neighbor	96.71	95.10	$\begin{bmatrix} 88 & 2 \\ 5 & 48 \end{bmatrix}$
Decision Tree Classifier	96.71	92.31	$\begin{bmatrix} 86 & 4 \\ 7 & 46 \end{bmatrix}$

Machine learning model performance can be visually represented with the use of a comprehensive metric called Receiver Operating Characteristics curves. The accompanying Figure 2 displays the ROC curves, which provide the performance of several classifiers. The successful classifiers have Area Under the Curve values of 0.98 for gradient boosting and 0.99 for random forest regression. It achieves complete discernment among negative and positive examples, which is the ideal categorization condition as indicated by this flawless AUC score. The XG Boost and

Ada Boost classifiers, with their respectable AUC values of 0.97 and 0.96, respectively, are not far behind. While these classifiers are not as good as their counterparts, their almost perfect score suggests that they are still rather good at consistently identifying data. The ROC analysis, which emphasizes the classifiers' capacity to identify subtle patterns in the dataset, supports the classifiers' strong discriminative abilities. In the decision trees, the lowest value of 0.91 was found. Apart from emphasizing the previously noted excellent accuracy, this detailed picture provides a nuanced view of the little differences in performance across the classifiers, adding to a better understanding of their data handling capabilities.

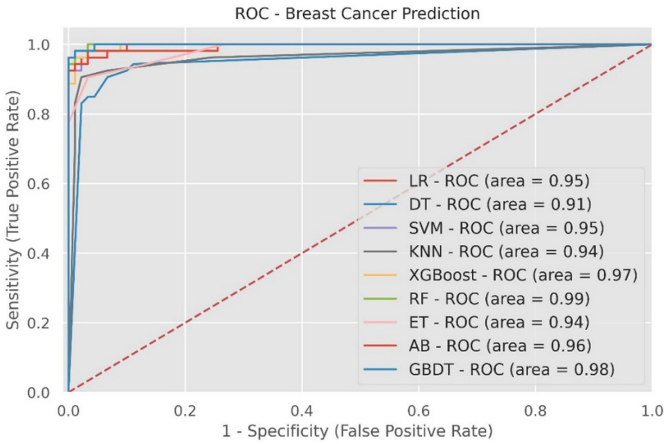

Fig. 2. Representation of ROC values of all classifiers

As can be seen in Figure 3, the amazing random forest and gradient boosting classifiers were able to attain higher accuracy, demonstrating the models' potential efficacy in recognizing intricate patterns in the data. This exceptional accuracy shows that these algorithms have mastered the classification tasks for this specific dataset, effectively recognizing the underlying structure of the features. The accompanying Figure 3 shows how well these classifiers can predict outcomes and graphically proves their high performance. Conversely, they appear to perform somewhat less accurately than their counterparts, but still attain impressive accuracy levels of 92% to 96%. More investigation into the specifics of the dataset is encouraged by this slight difference in algorithm performance, which could lead to better results.

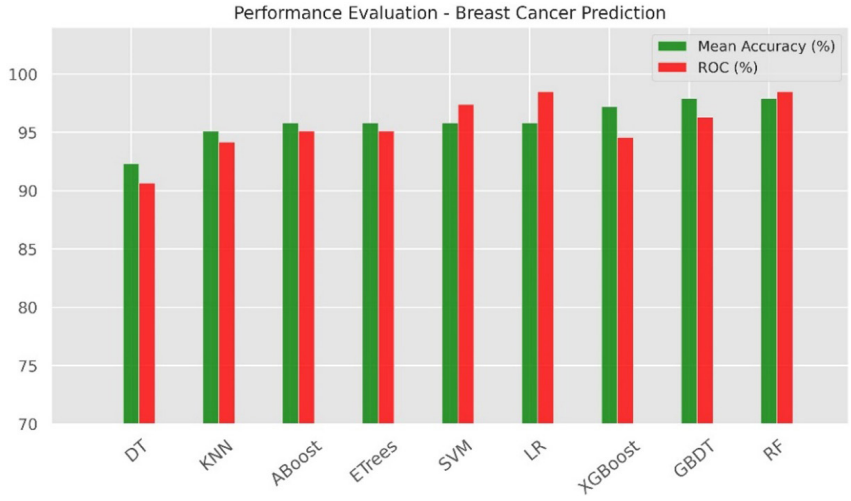

Fig. 3. Comparison of mean accuracy and ROC of all classifiers

4.2 Performance assessment of ensemble classifiers

The proposed methodology combines multiple ML models through an ensemble approach, resulting in a more accurate output overall than with a single model. The ensemble approach used in this case is soft voting, which averages the estimated probability from six different models to get a final forecast. Soft voting often outperforms hard voting and is leverages the distinct strengths of each model to increase forecast accuracy.

The performance of ensemble classifiers on the two datasets under consideration is shown in Table 5. With an equally impressive precision of 0.98 and an exceptionally low rate, the ensemble classifier achieved an impressive accuracy of 0.99. The model demonstrates its ability to identify TP with a recall of 1.00. The F1 score, which accounts for accuracy and recall, was 0.99. The outstanding capacity to differentiate across classes was indicated by the AUC, which attained a perfect 1.00. The ROC-AUC overall is displayed in Figure 4.

Table 5. Performance of ensemble classifiers

Metric	Achieved Percentage
Accuracy	0.99
Precision	0.98
Recall	1.00
F1 Score	0.99
AUC	1.00

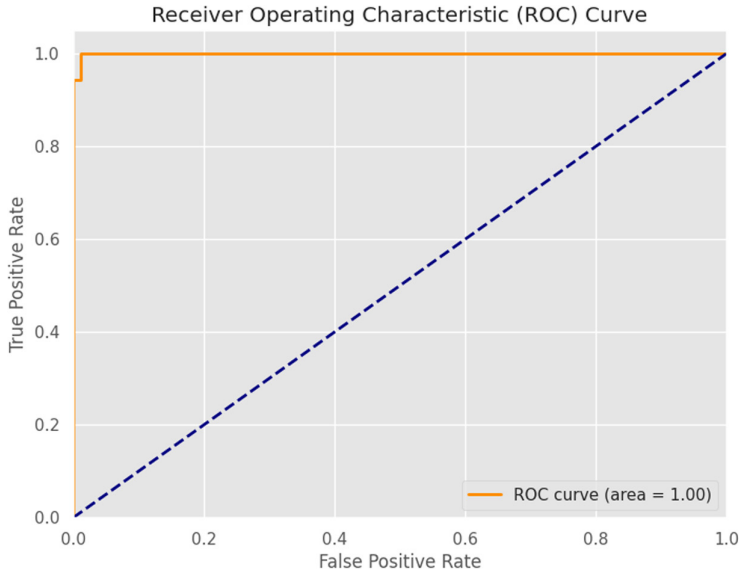

Fig. 4. ROC-AUC curve at of ensemble model

4.3 Discussions and comparison

This work aims to prove that RF and GBDT models outperform other ML methods such as XG Boost, LR, SVM, E Trees, ABoost, KNN, and DT in detecting breast cancer.

It is specifically predicted that RF and GBDT models will show improved AUC-ROC values, increased TP rates, and higher accuracy. An accuracy rate of 97.90%, the highest TP count of 89, and an AUC-ROC value of 0.99 suggest that RF and GBDT perform noticeably better in this situation. By decreasing the amount of and FN, these metrics demonstrate how well the models can distinguish between cases of benign and malignant breast cancer. The underlying advantages of the RF and GBDT algorithms form the basis of the theory. RF uses an ensemble of decision trees with the bagging idea to lessen overfitting and enhance generalization. GBDT, on the other hand, produces solid predictive performance by optimizing models sequentially through the minimization of a loss function. It is thought that these traits have a part in the higher performance measures that the study found.

As part of the study project, the datasets were collected and examined in order to increase the precision of breast cancer detection. As seen in Table 6, the accuracy of the proposed model is compared with other works [24–27]. Most importantly, the model was effective in identifying breast cancer in patients, and the outcomes suggest that the developed model may have additional applications.

Table 6. Comparison of performance results

	[24]	[25]	[26]	[27]	Proposed Work
Random Forest Classifier	93	94.5	96.49	96.12	97.90
Gradient Boosting Classifier	95	94	–	–	97.90
XGBoost Classifier	–	94.5	–	95.60	97.20
Logistic Regression	–	–	92.98	–	95.80
Support Vector Machine	87	–	89.41	–	95.80
Extra Trees Classifier	–	94.5	–	94.20	95.80
AdaBoost Classifier	–	90.5	–	–	95.80
K-Nearest Neighbor	–	–	92.10	–	95.10
Decision Tree Classifier	–	–	93.85	90.36	92.31

5 CONCLUSION

The main goal of this study is to employ ML classification algorithms to build a system that will help doctors forecast patient survival and the diagnosis of breast cancer tumors. We trained using nine different algorithms and leveraged the publicly available breast cancer dataset for our training. To maximize performance, each algorithm underwent a thorough hyperparameter tuning procedure. The models were evaluated based on key performance indicators, and the random forest algorithm achieved an accuracy rate of 97.90% and came out with the highest ROC score of 0.99. On the other hand, decision trees produced 92.31% accuracy and an ROC of 0.91. This thorough investigation highlights the effectiveness of ML methods in improving the diagnosis and prognosis of breast cancer, with random forest demonstrating the best prediction performance in this regard. Furthermore, the work will be extended for an image-based dataset using deep learning techniques for determining the disorder's severity.

6 REFERENCES

[1] E. A. Bayrak, P. Kirci, and T. Ensari, "Comparison of machine learning methods for breast cancer diagnosis," in *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, 2019, pp. 1–3. https://doi.org/10.1109/EBBT.2019.8741990

[2] S.-I. Chen, H.-T. Tseng, and C.-C. Hsieh, "Evaluating the impact of soy compounds on breast cancer using the data mining approach," *Food & Function*, vol. 11, no. 5, pp. 4561–4570, 2020. https://doi.org/10.1039/C9FO00976K

[3] V. Chaurasia, S. Pal, and B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," *Journal of Algorithms & Computational Technology*, vol. 12, no. 2, pp. 119–126, 2018. https://doi.org/10.1177/1748301818756225

[4] Y. Guan *et al.*, "Willingness to decrease mammogram frequency among women at low risk for hereditary breast cancer," *Scientific Reports*, vol. 9, 2019. https://doi.org/10.1038/s41598-019-45967-6

[5] S. Blandin Knight, P. A. Crosbie, H. Balata, J. Chudziak, T. Hussell, and C. Dive, "Progress and prospects of early detection in lung cancer," *Open Biology*, vol. 7, no. 9, pp. 1–12, 2017. https://doi.org/10.1098/rsob.170070

[6] N. Jothi, N. A. Rashid, and W. Husain, "Data mining in healthcare – A review," *Procedia Computer Science*, vol. 72, pp. 306–313, 2015. https://doi.org/10.1016/j.procs.2015.12.145

[7] R. K. B and N. Chandra Gowda, "A framework for sentiment analysis in customer product reviews using machine learning," in *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, 2020, pp. 267–271. https://doi.org/10.1109/ICSTCEE49637.2020.9276877

[8] N. C. Gowda and B. M. A, "A trust prediction mechanism in edge communications using optimized support vector regression," in *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, 2023, pp. 784–789. https://doi.org/10.1109/ICCMC56507.2023.10083686

[9] L. Shalini, S. S. Manvi, B. Gardiner, and N. C. Gowda, "Image based classification of COVID-19 infection using ensemble of machine learning classifiers and deep learning techniques," in *2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, 2022, pp. 1–6. https://doi.org/10.1109/ICDSAAI55433.2022.10028859

[10] M. Rahul, P. Ravichandra, M. M. Yakoobi, and N. C. Gowda, "Deep learning-based solution for differently-abled persons in the society," in *2023 4th International Conference for Emerging Technology (INCET)*, 2023, pp. 1–6. https://doi.org/10.1109/INCET57972.2023.10170230

[11] H. N. Veena, K. K. Patil, P. Vanajakshi, A. Ambore, and N. C. Gowda, "An enhanced RNN-LSTM model for fundus image classification to diagnose glaucoma," *SN Computer Science*, vol. 5, 2024. https://doi.org/10.1007/s42979-024-02867-5

[12] N. C. Gowda, V. H. N., and D. R. Ramani, "Efficient identification of deepfake images using CNN," in *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, 2024, pp. 1542–1547. https://doi.org/10.1109/IDCIoT59759.2024.10467555

[13] V. H. N., Rajani, N. C. Gowda, and D. R. Ramani, "Two stage video classification approach using convolution neural network," in *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, 2024, pp. 1548–1554. https://doi.org/10.1109/IDCIoT59759.2024.10467591

[14] K. N. Maxwell and K. L. Nathanson, "Common breast cancer risk variants in the post-COGS era: A comprehensive review," *Breast Cancer Research*, vol. 15, 2023. https://doi.org/10.1186/bcr3591

[15] C. Hou *et al.,* "Predicting breast cancer in Chinese women using machine learning techniques: Algorithm development," *JMIR Medical Informatics*, vol. 8, no. 6, 2020. https://doi.org/10.2196/17364

[16] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns,* vol. 4, no. 9, 2023. https://doi.org/10.1016/j.patter.2023.100804

[17] B. He *et al.,* "A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation," *Biochimica et BiophysicaActa (BBA) – Molecular Basis of Disease*, vol. 1866, no. 11, 2020. https://doi.org/10.1016/j.bbadis.2020.165916

[18] V. Amudha, R. G. Babu, K. Arunkumar, and A. Karunakaran, "Machine learning-based performance comparison of breast cancer detection using support vector machine," *AIP Conference Proceeding,* vol. 2519, no. 1, 2022. https://doi.org/10.1063/5.0110848

[19] A. Bhardwaj, H. Bhardwaj, A. Sakalle, Z. Uddin, M. Sakalle, and W. Ibrahim, "Tree-based and machine learning algorithm analysis for breast cancer classification," in *Computational Intelligence and Neuroscience*, A. R. Javed, Ed., vol. 2022, 2022, no. 1, pp. 1–6. https://doi.org/10.1155/2022/6715406

[20] O. J. Egwom, M. Hassan, J. J. Tanimu, M. Hamada, and O. M. Ogar, "An LDA–SVM machine learning model for breast cancer classification," *BioMedInformatics*, vol. 2, no. 3, pp. 345–358, 2022. https://doi.org/10.3390/biomedinformatics2030022

[21] N. Fatima, L. Liu, S. Hong, and H. Ahmed, "Prediction of breast cancer, comparative review of machine learning techniques, and their analysis" in *IEEE Access*, vol. 8, pp. 150360–150376, 2020. https://doi.org/10.1109/ACCESS.2020.3016715

[22] F. Khan *et al.,* "Cloud-based breast cancer prediction empowered with soft computing approaches," *Journal of Healthcare Engineering,* vol. 2020, no. 1, 2020. https://doi.org/10.1155/2020/8017496

[23] U. N. Wisesty, T. R. Mengko, and A. Purwarianti, "Gene mutation detection for breast cancer disease: A review," *IOP Conference Series: Materials Science and Engineering*, vol. 830, no. 3, p. 032051, 2020. https://doi.org/10.1088/1757-899X/830/3/032051

[24] A. Gago, J. M. Aguirre, and L. Wong, "Machine learning system for the effective diagnosis and survival prediction of breast cancer patients," *International Journal of Online and Biomedical Engineering (iJOE),* vol. 20, no. 2, pp. 95–113, 2024. https://doi.org/10.3991/ijoe.v20i02.42883

[25] S. Sharmin, T. Ahammad, Md. A. Talukder, and P. Ghose, "A hybrid dependable deep feature extraction and ensemble-based machine learning approach for breast cancer detection," in *IEEE Access*, vol. 11, pp. 87694–87708, 2023. https://doi.org/10.1109/ACCESS.2023.3304628

[26] A. Khalid *et al.,* "Breast cancer detection and prevention using machine learning," *Diagnostics,* vol. 13, no. 19, p. 3113, 2023. https://doi.org/10.3390/diagnostics13193113

[27] A. Batool and Y.-C. Byun, "Toward improving breast cancer classification using an adaptive voting ensemble learning algorithm," *IEEE Access*, vol. 12, pp. 12869–12882, 2024. https://doi.org/10.1109/ACCESS.2024.3356602

[28] M. H. Yasser, "Breast cancer dataset," *Kaggle*, 2021. https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset

[29] N. Chandrasekhar and S. Peddakrishna, "Enhancing heart disease prediction accuracy through machine learning techniques and optimization," *Processes*, vol. 11, no. 4, p. 1219, 2023. https://doi.org/10.3390/pr11041210

[30] H. H. Muljo, A. S. Perbangsa, Y. Yulius, and B. Pardamean, "Mobile learning for early detection of cancer," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 12, no. 2, pp. 39–53, 2018. https://doi.org/10.3991/ijim.v12i2.7814

[31] P. Rajesh, A. Murugan, B. Murugamantham, and S. Ganesh Kumar, "Lung cancer diagnosis and treatment using AI and mobile applications," *International Journal of Interactive Mobile Technologies (iJIM),* vol. 14, no. 17, pp. 189–203, 2020. https://doi.org/10.3991/ijim.v14i17.16607

7 AUTHORS

B. N. Ravi Kumar is an Assistant Professor in the Dept. of ISE at BMS Institute of Technology & Management, Bangalore. He completed Ph.D. in Computer and Information Sciences at VTU in 2023, M.Tech in CSE in 2013, and B.E. in CSE in 2008. He has teaching experience of 14 years, research experience of 6 years, and 2 years industry experience. He has 15 publications in international conferences and reputed journals. His research interests are in the areas of software engineering, artificial intelligence, and machine learning (E-mail: ravikumarbn@bmsit.in).

Naveen Chandra Gowda received his Doctorate in Computer and Information Sciences from Vishveswaraya Technological University, Belgaum, and Karnataka. He is currently working as an Assistant Professor in the School of CSE, REVA University, Bengaluru. He has more than 14 years of experience in teaching and research. He has published around 30 papers in national and international journals and presented at national conferences. He is also the reviewer for six international journals and national and international conferences. He has presented various keynote talks in workshops organized around India. He is a Life Member of ISTE, CSI, and ACM (E-mail: ncgowdru@gmail.com).

B. J. Ambika is an esteemed Assistant Professor in the Senior Scale at the Department of Computer Science and Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India. With a commendable 13 years of teaching experience across various institutions and universities in Karnataka, India, she has established herself as a dedicated educator and researcher in the field of computer science. Her research interests span a broad spectrum of contemporary topics, including computer networks, big data, cloud computing, artificial intelligence (AI), and machine learning (ML). Her passion for advancing knowledge in these areas is reflected in her prolific academic output, having published more than 16 articles in reputed journals and conferences. She is a Life Member of ISTE, CSI (E-mail: ambika.bj@manipal.edu).

H. N. Veena received his Doctorate in Computer and Information Sciences from Vishveswaraya Technological University, Belgaum, Karnataka. She is currently working as an Associate Professor in the Department of CSE, SJB Institute of Technology Bengaluru. She has more than 10 years of experience in teaching and research. She has published around 15 papers in national and international journals and presented at national conferences. She is also the reviewer for International Journals and national and international conferences (E-mail: hn.veenagowda@gmail.com).

B. Ben Sujitha is serving as a Professor at the Department of Computer Science and Engineering, Noorul Islam Centre for Higher Education, Thuckalay, Kanyakumari, Tamil Nadu. She has completed B.E., and M Tech, Ph.D. She has received a doctorate degree from Anna University, Chennai. She has worked in the best-reputed institutions. She has published more than 25 national and international journals. She has delivered motivational lectures to the faculty development program and

student community. Received "BEST FACULTY ADVISOR AWARD" for outstanding achievements and remarkable role in the field of education. She has published books and patents. She received certificates from IIT and NPTEL. Motivating, dedicated, and passionate lecturer fostering a student-centered learning environment with excellent administrative and interpersonal skills. My area of expertise is intrusion detection, machine learning, edge computing and health informatics. Spanning over areas of data science with deep learning, machine learning and artificial intelligence. Professional membership in ACM, ISTE, and CSI (E-mail: bensujitha@gmail.com).

D. Roja Ramani is an Associate Professor in the Department of Computer Science and Engineering at New Horizon College of Engineering, Bangalore, Karnataka. She completed her B.Tech, M.Tech, and Ph.D. from Anna University. With over 40 publications, including books, journal articles, conference proceedings, and patents, she has delivered motivational lectures to faculty and students. She received the "BHARAT SHIKSHA RATAN AWARD" and "Best Young Scientist Award 2021." She was honored with the "Competent Communicator" Award by Toastmasters International and was the 2nd Runner Up in the AICTE Lilavati Award 2021–2022, as well as receiving the Kalam's Digital Award. Additionally, she is recognized as a Silver and Bronze partner faculty under the Inspire-Infosys Campus Connect Faculty Partnership Model. Dr. Ramani's expertise lies in computerized image processing, machine learning, biomedical and health informatics, data science, deep learning, and artificial intelligence. She is a member of ISTE and CSI (E-mail: rosevsroja@gmail.com).