IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 2, April 2026, pp. 1327~1338
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i2.pp1327-1338

Attribute optimization to improve breast cancer prediction using machine learning techniques

Raghavendra Srinivasaiah[1], Santosh Kumar Jankatti[2], Niranjana Shravanabelagola Jinachandra[3], Manjunath Ramanna Lamani[4], Bellam Vijaya Lakshmi[5], Rishita Bhelwa[5]
[1]Department of Artificial Intelligence and Data Science Engineering, School of Engineering and Technology, CHRIST University, Bangalore, India
[2]Department of Computer Science and Technology, Dayananda Sagar University, Bangalore, India
[3]Department of Mechanical Engineering, CHRIST University, Bangalore, India
[4]Department of Computer Science and Engineering, Moodlakatte Institute of Technology, Kundapura, India
[5]Department of Computer Science and Engineering, CHRIST University, Bangalore, India

Article Info

Article history:

Received Jul 13, 2024
Revised Jan 8, 2026
Accepted Jan 25, 2026

Keywords:

Attribute optimization
Breast cancer prediction
Machine learning
Random forest classifier
Wisconsin

ABSTRACT

Breast cancer (BC) arises when cells grow out of control. It affects women more than men. Seeking cancer treatment can be both costly and time-consuming, with test results spanning from a few hours to several weeks. The duration of these tests depends on the number of attributes within the dataset. This research paper endeavors to optimize the dataset attributes and find the accuracy of the optimized dataset. The primary goal is to reduce features using recursive feature elimination to minimize the time taken for the test result. This work discusses the machine learning technique and the random forest (RF) algorithm, which helps determine the parameter accuracy on the Wisconsin BC diagnostic dataset. The method achieves an accuracy of 96.49% with only eighteen attributes. It has aided the healthcare industry in finding BC in less time and improving the treatment.

Corresponding Author:

Raghavendra Srinivasaiah
Department of Artificial Intelligence and Data Science Engineering, School of Engineering and Technology
CHRIST University
Kanmanike, Kumbalgudu, Mysore Road, Bangalore-74, India
Email: raghav.trg@gmail.com

1. INTRODUCTION

There are many cases in the world where most women have breast cancer (BC). It is a type of cancer caused by the overgrowth of specific cells in the human body. As the cancer is detected in a later stage, the treatment is a financial issue as its price increases, especially in advanced stages. Prompt and precise BC diagnosis is essential for efficient treatment planning and improving patient outcomes [1]. Machine learning (ML) techniques have become increasingly powerful in medical research over the past few years, particularly in disease prediction and diagnosis. Some ML techniques like artificial neural networks (ANN), support vector machine (SVM), k-nearest neighbors (KNN), and decentralized ML approaches like federated learning are also used to diagnose BC [2]–[8].

Several side effects must be addressed before treating a patient. An oncologist will be able to identify and diagnose BC with the help of magnetic resonance imaging (MRI) scan, and tissue biopsy, which is a time-consuming process [9]. Screening is a type of test that identifies most cancer cells quickly and early [10]. Image processing techniques, such as contrast-restricted adaptive histogram equalization (CLAHE), remove noise from the image and enhance image quality by intensifying pixel values and cell areas to produce a clearer picture with sharper details, facilitating the identification of objects and boundaries. Fuzzy

SVM, Bayesian classifier, and random forest (RF) approaches are used to classify these pre-processed pictures [11]. There are also several cases where BC is identified through blood reports. BC can be identified by a lump or mass in the breast tissue, and not all lumps are cancerous, but mammographic images can identify those that are [12]. BC is a complex condition that has many molecular and symptom manifestations. It is essential to identify the characteristics with high predictability among different variables to make reliable and understandable models [13].

One popular feature selection (FS) technique named recursive feature extraction (RFE) serves as a powerful tool for a systematic reduction of unnecessary data, which increases the efficiency and interpretability of models for prediction. The primary motivation for this research is exploration of the usage of RFE in the context of BC prediction. RFE's ability to assess and rank variables based on their importance creates scope for a new set of features that help predict BC most effectively [14]. The following sections would explore in detail technologies used, features of the dataset, and experiment outcome carried out using RFE for predicting BC. The aim is to highlight the benefits brought by attribute reduction in predicting BC accurately and model interpretability, which could pave the way for further efforts for early diagnosis and personalized medication.

2. LITERATURE REVIEW

To understand the different algorithms used for predicting and diagnosing BC, different research publications are studied. When it comes to the prediction of BC, some of these algorithms illustrate different levels of accuracy. The need for the usage of advanced models for the better identification of BC is described through an automated detection mechanism based on an ensemble of classifiers. The challenges faced when identifying BC at an early stage due to the small size of the cancer cells are described extensively. The time-consuming process of testing is also emphasized. The ANN and an ensemble of ML algorithms are some of the different ML techniques used. When using all of the variables within the dataset, the proposed method illustrates an accuracy rate of 98.83% [15]. The five different ML algorithms employed in the dataset are ANN, RF, KNN, logistic regression (LR), and SVM. Accuracy, specificity, sensitivity, F1-score, precision, negative predictive values, false negatives, and false positives are the metrics employed to test the performance of different techniques. As evident from the results, the ANN employs all the variables to achieve higher accuracy of 98.57% [16]. In considering maximizing accuracy and minimizing errors, the prediction of BC through ML was explored. In correcting the errors in existing methods, there was an opportunity to improve prediction models. Four different models of ML algorithms—SVM, ANN, LR, and RF—were used in the dataset through the Jupyter environment. From the experimental results, LR utilized all the variables and performed better than other models in accuracy models [17].

A thorough discussion is given on segmentation-based ML and effective image processing methods for BC diagnosis. The input data for this work is mammography pictures. To improve image quality, the CLAHE method is employed, which helps reduce noise in the images and enhances image quality. Techniques like fuzzy SVM, RF, and Bayesian classifier group the preprocessed images. From the result obtained, fuzzy SVM performs better than the other methods with an accuracy of 94% [18]. A new nested ensemble technique for automated BC diagnosis was introduced, demonstrating a research gap related to the limited exploration of hyperparameter tuning and FS. In this work, Meta classes are utilized in conjunction with cross-validation techniques for model evaluation to distinguish between benign breast tumors and malignant cancers. From the result obtained it was found that SV-BayesNet-3-MetaClassifier and SV-naïve Bayes-3-MetaClassifier achieved accuracy of 98.07% with complete attributes [19].

Different ML techniques and deep learning (DL) algorithms detect benign and malignant tumors. Models such as SVM, LR, multilayer perceptron (MLP), ANN, and KNN were applied to the dataset, and the results were compared. The comparison of the results found that the ANN achieved an accuracy of 99.3% using the complete set of attributes [20]. A comparative analysis of ML algorithms for BC prediction was conducted, identifying challenges related to data size and the limitations of decision trees (DT) in specific scenarios. A dataset is subjected to several techniques, including SVM, KNN, DT, K-means, and ANN, for the early diagnosis of benign and malignant cancer. SVM was determined to have accuracy of 97.14% when all characteristics were used [21]. The necessity to find the best classification features and combine radiomics and genomes data was the main focus of the exploration of ML algorithms for BC type categorization. Triple-negative and non-triple-negative BC were classified using gene expression data and the ML technique. SVM, K-means, naïve Bayes, and DT are the four classification models that are compared. The outcome unequivocally shown that ML algorithms outperform other techniques [22]. To classify patients into groups no cancer, cancer, and non-cancerous, the researchers used ML and DL techniques for the identification of BC from the thermographic image. Three classification systems are used: RF, SVM, and convolution neural network (CNN). CNN has been found more efficient than other systems as proved in [23].

A detailed discussion on picture formation and preprocessing methods related to the detection of BC was introduced. In an effort to improve the precision of this model, the importance of the integration of artificial intelligence (AI) techniques and novel methodologies is emphasized in this paper. The significance of varying the size of the used data in deriving broader aspects was emphasized in the thorough discussion, suggesting a probable area of research for the future [24]. The most fatal and life-threatening type of cancer is BC, wherein it is first discovered when breast enlargement happens. Early diagnosis is therefore crucial. Mammography and ultrasound methods are typically employed for the detection. ML techniques like CNN can be used to detect mammograms. Each layer of the CNN identifies the features and patterns that help efficiently find anomalies. An approach based on BreaseNet-SVM is employed on the digital database for screening mammography (DDSM) datasets to automatically detect and classify BC. According to the results, the model achieved accuracy of 99.16% [25]. BC is one of the most critical worldwide health issues, and most often it affects women. The gradient boosting (GB) method is applied to the dataset to identify the vital critical factors. The GB method was employed for disease classification. To evaluate the model's performance, various criteria, including sensitivity, accuracy, specificity, F1-score, and positive and negative predictive values, were used. All methods achieved 100% accuracy [26].

FS is the process of attempting to prevent the multiplicity of features, which is the most significant problem in disease diagnosis. FS methods help in detecting the essential features that contribute effectively to the models' performance improvement. FS methods help eliminate and remove unnecessary data [27]. Three classifiers based on ANN, genetic algorithms (GA), and particle swarm optimization (PSO) with FS methods are applied on the Wisconsin dataset. It was found that the PSO classifiers achieved an accuracy of 97.2%, specificity of 95.6% and sensitivity of 98%; GA classifiers attained an accuracy of 96.6%, specificity of 93.7% and sensitivity of 97.5%; and ANN classifiers achieved an accuracy of 97.3%, specificity of 95.1% and sensitivity of 98.4% [28]. MRI is identified as a potential candidate for directing near-infrared spectral tomography, which enhances the specificity and sensitivity of BC diagnosis. However, the difficulty in light propagation in the MRI images affects the performance of spectral tomography. To overcome the problems, a 3D spectral image was developed guided by MR, which achieved an accuracy of 89.5%, specificity of 92.9%, sensitivity of 87.5%, and a receiver operating characteristic (ROC) curve of 0.98 [29]. RFE is an FS technique that continually attempts to select the most critical features, primarily focusing on classification accuracy and the learned model. RFE works by sequentially removing the worst features, which reduces the performance of the technique. RFE works by using the backward elimination technique, which involves eliminating attributes to reduce efficiency recursively [30], [31]. RFE suffers from problems such as inconsistencies with the feature ranking criterion and the maximum margin concept, as the computation of the criterion is done locally. Additionally, there is a lack of global measurement of feature importance, which is not guaranteed to be optimal, and a high risk of overfitting [32], [33].

Early diagnosis of BC is the best way to cure the disease. To solve the problems associated with errors in diagnosing the disease, a hybrid model combining principal component analysis (PCA) and SVM is proposed. PCA was used to select features in the first cycle and reduce the number of features in the second cycle. The reduced features are fed into SVM for risk assessment and diagnosis; the proposed model achieved an accuracy of 97.62%, specificity of 100%, and sensitivity of 95.24% [34]. In recent years, many computerized diagnostic systems have been developed to reduce human errors and help physicians diagnose diseases effectively. An attempt was made to create a computer-aided diagnosis system utilizing pattern recognition software. A hybrid technique was proposed by combining LR and PCA, in which PCA was used for FS and LR for classification of BC tumors. The hybrid method achieved an accuracy of 100% by outperforming many existing methods, which included reducing the quality of the attributes, the number of attributes, and response time [35]. Irrelevant and duplicated features lead to a reduction in prediction accuracy and also make the system ambiguous. The literature collection and segregation provide a better overview of current studies, allow for a deeper understanding of the research landscape, identify gaps, and provide a foundation for subsequent research work on attribute optimization for BC prediction. Through this approach, the medical costs and test time will be reduced.

3. METHOD

Here are the steps involved in the process of finding the accuracy of the optimized dataset and the original dataset using ML techniques:

i) Collecting dataset: the first process in this undertaking is the collection of the dataset that will be employed in testing and training the ML models as well as evaluating the accuracy of BC detection. This dataset can be obtained from different sources in the form of databases or spreadsheets. In this paper, the dataset being utilized has been obtained from Kaggle.

ii) Checking for missing values: missing values need to be verified after collection. There may be missing values in the data for several reasons, including mistakes in the process of data entry. The above

procedure is called data preprocessing. The row information should be deleted in case missing values exist in the data. Otherwise, the user can proceed to the next phase to split the dataset via the ML technique named random forest classifier (RFC).

iii) Delete attributes: the user needs to delete the row information for the missing value, if there are any, from this step because it will help obtain the correct positive result.

iv) Splitting dataset using RFC: if there are no missing values in the dataset, then the dataset can be split into the training data set and the testing data set by using the RFC technique.

v) Initial accuracy: the ML algorithm is trained on a training dataset after the dataset has been split. After training the algorithm, it is ideal to check how it performed using a testing dataset. The initial accuracy of the algorithm is otherwise called the accuracy of the original dataset.

vi) RFE to reduce the dataset: RFE is a technique used to reduce the number of features in the dataset. Optimizing the number of features can enhance performance and reduce the time required to predict BC. RFE removes the features that have the most negligible impact on the model's accuracy. This process is repeated until a desired number of features is reached. This paper reduces the dataset by 60% of the original dataset. The minimized dataset is reduced to 18 attributes after using the RFE technique.

As shown in Figure 1, this paper utilizes the RF algorithm and RFE to achieve the desired result. The detailed implementation steps are described in Algorithm 1. The RFC determines the result by considering the outputs of binary trees. The RFC technique used in this paper explains how to find the accuracy of the original dataset and classify it into two sets: a training set and a testing set. The testing set is used for performing tests, while the training set is used to train the system. The RFC is used once again after RFE is applied to the dataset. RFC performs the same operation on the reduced feature dataset.

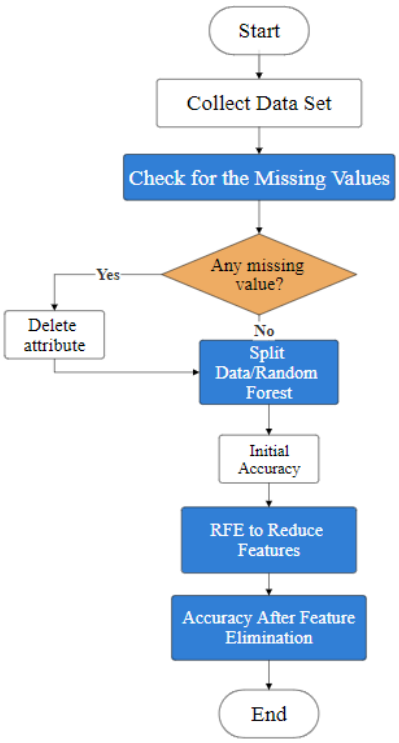

Figure 1. Implemented process

Algorithm 1. Proposed RFE–RFC procedure for feature selection and accuracy evaluation
Step 1: Begin
Step 2: Upload the dataset to colab and load the data.
Step 3: Divide the dataset into X (feature) and Y labels (target).
Step 4: Choose a ML algorithm (RFE) for FS.
Step 5: The entire data is then ranked based on the importance of the attributes using the RFE algorithm. The features are then selected based on a fixed percentage (60%), and the least important attributes are removed based on the ranking.
Step 6: Repeat step 5 until the desired number of features is selected.

Step 7: Use the desired properties to evaluate the accuracy of the RFC.
Step 8: Display the accuracy
Step 9: End
Accuracy calculation formulae using RFC:
test set:
Y_{test} = true labels
Y_{pred_rf} = predicted label (using the RF)
Accuracy is calculated as shown in (1) and (2).

$$accuracy_{rf} = \frac{Number\ of\ Coorect\ Predictions}{Total\ Number\ of\ Predictions} \times 100 \quad (1)$$

$$accuracy_{rf} = \frac{\sum_{i=1}^{N}(Y_{test}\ [i]==Y_{pred\_rf}[i])}{N} \times 100 \quad (2)$$

N=total number of samples.

RFE is an FS technique that selects a subset of relevant features from a larger dataset. It will identify and remove less critical features by working with an LR model. This approach is based on a greedy search technique. Minimization of dataset using RFE:
Feature selection:

- Let X be the original feature matrix with dimensions m × n, where, m = number of samples and n = number of features.
- Calculating the number of features to select: num_features_to_select =⌊0.60× n⌋.

Reduced feature matrix:

- Transform the original feature matrix X using RFE : X_{rfe}= rfe.fit_transform (X, Y). X_{rfe} = the reduced feature matrix with dimensions m × num_features_to_select.

Table 1 displays accuracy, attributes, and the time taken to calculate accuracy in seconds. The results shown are obtained from the current study using RFC and RFE techniques. The original dataset contains 32 attributes, and after applying the FS method, the number of attributes was reduced to 18, while maintaining the same accuracy as the original dataset. From the result obtained the most essential attributes identified are as follows: mean radius, mean texture, mean perimeter, mean smoothness, mean concavity, mean concave points, mean symmetry, texture error, area error, concavity error, worst radius, worst texture, worst perimeter, worst smoothness, worst concavity, worst concave points, worst symmetry, and worst fractal dimension.

Table 1. Displays accuracy, attributes, and the time taken to calculate accuracy in seconds

Dataset	Time taken (in seconds)	Accuracy (%)
Original dataset	0.26	96.49
Reduced dataset (60%)	0.22	96.49

This study utilizes the publicly available BC Wisconsin (diagnostic) dataset from UCI/Kaggle [36], which contains 569 instances with 30 numeric predictive features derived from digitized fine-needle aspirate (FNA) images. The target variable is binary: malignant (212 cases) and benign (357 cases). There are no missing values, and the 'id' column was discarded. The dataset was stratified and split into 75% training (427 samples) and 25% testing (142 samples) sets using test_size =0.25, stratify =y, and random_state =42. All experiments employed scikit-learn's RFC with the following fixed hyperparameters: n_estimators =100, max_depth = none, min_samples_split =2, min_samples_leaf =1, class_weight = 'balanced', random_state =42. RFE was applied using the same RF as the base estimator, n_features_to_select =18, step =1, and scoring='accuracy', and reduces the feature set from 30 to 18 (40% reduction). All reported results are obtained via 5-fold stratified cross-validation on the training set (StratifiedKFold, shuffle = true, random_state =42). Final performance on the held-out test set is also reported. Metrics include accuracy, precision, recall, F1-score, specificity, and area under the curve (AUC)-ROC (mean ± standard deviation where applicable).

4. RESULTS AND DISCUSSION

Table 2 shows the comparison of the proposed method with the performance of the other existing research using the different FS methods on the dataset in terms of the number of attributes used and the accuracy achieved. It can be observed that the existing FS methods achieve better accuracy with fewer

attributes compared to the proposed method; however, the existing techniques do not focus on the time taken for prediction. Table 3 depicts the hyperparameters and experimental settings.

Table 2. Performance comparison of the proposed method with existing methods based on selected attributes

FS method	Number of features selected	Accuracy
PSO [36]	12	99.82
Modified bat algorithm [37]	10	98.70
GA [28]	9-11	97.13
REF [38]	8	97.5

Table 3. Hyperparameters and experimental settings

Component	Setting
Dataset	BC Wisconsin (diagnostic), 569×30
Train/test split	75%/25%, stratified, random_state =42
Cross-validation	5-fold stratified
RF	n_estimators =100, class_weight='balanced', random_state =42
RFE	estimator = RandomForest, n_features_to_select =18, step =1, scoring = accuracy
Software	Python 3.12, scikit-learn 1.5.0, pandas 2.2, NumPy 1.26

Figure 2 facilitates a comparison of accuracy, the number of features, and the time required to compute accuracy for both the original and optimized datasets. In Figure 2, the blue bar denotes the time taken by the original dataset and the optimized dataset, while the maroon bar shows the number of attributes considered by the original and the optimized dataset, and the green bar shows the accuracy achieved by the original and the optimized dataset. Reducing the number of attributes in the optimized dataset maintains the same accuracy as the original dataset but diminishes the time taken to calculate accuracy. Specifically, with a 40% reduction in the dataset (attributes reduced from 32 to 18), the time taken to find the accuracy decreases to 3 units compared to the original dataset.

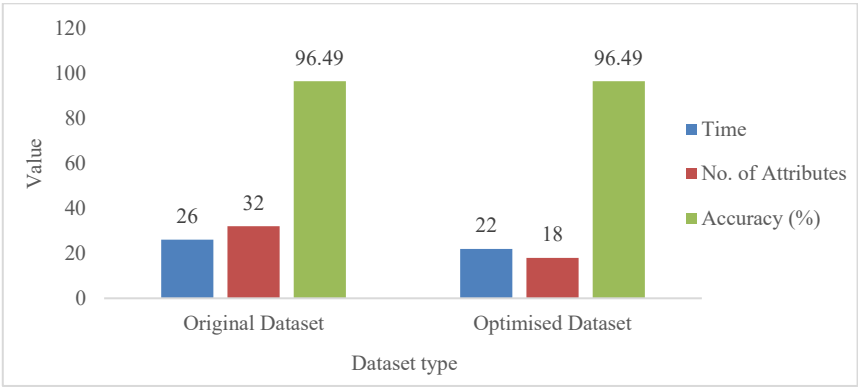

Figure 2. Comparing original and optimized dataset accuracy, attributes, and time

Thus, we may deduce that the time needed to forecast BC decreases in proportion to the number of qualities. In addition to helping patients by cutting down on test length, treatment time, expenses, and waiting times for test results, this time reduction also helps lower BC's overall death rate. In conclusion, fewer characteristics result in a more successful prediction procedure, guaranteeing prompt and efficient medical care for patients and eventually helping to manage and lower the mortality rate related to BC.

The accuracy of the suggested approach and the number of attributes employed in prediction are contrasted with those of other approaches in Figure 3. Figure 3 illustrates how some current approaches, which make use of the entire collection of features, perform better than the suggested approach. The accuracy achieved by the proposed method is nearly that of the existing methods, and it uses only 18 attributes, which is approximately 60% of the total attributes. In this paper, the prediction time is reduced, which will also decrease the time required for testing (analyzing cancer cells), as well as the time spent on test results and medical costs.

Figure 4 presents the overall evaluation of the RFE-optimized RF model, showing high scores for accuracy (96.49%), precision (95.89%), recall (98.59%), and F1-score (97.22%). These results confirm that the 18-feature model provides reliable and clinically meaningful diagnostic performance. The confusion matrix and performance metrics bar chart in Figure 5 show consistently high values for accuracy, precision, recall, and F1-score. Low misclassification rates are shown in the figure's confusion matrix, which displays only one false negative and three false positives. Overall, Figure 5 shows that for both benign and malignant instances, the model offers consistent and reliable categorization.

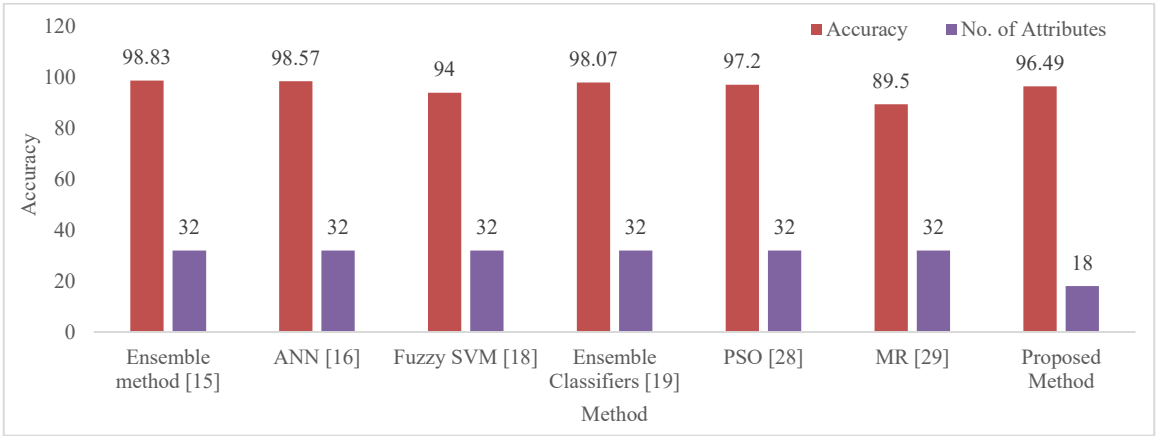

Figure 3. Comparing proposed method with existing methods

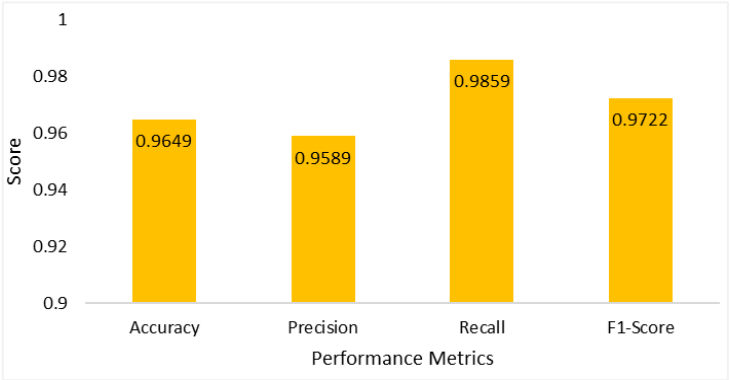

Figure 4. Proposed model-performance metrics

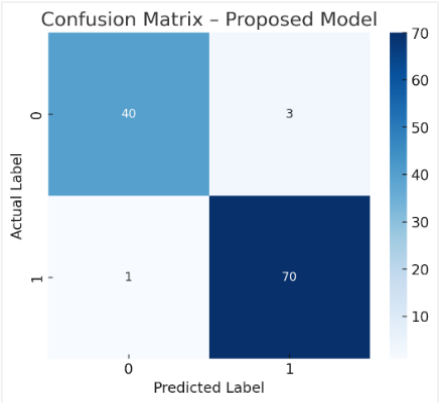

Figure 5. Proposed model-confusion matrix

Excellent sensitivity with few false positives is indicated by Figure 6 ROC curve, which rises strongly in the upper-left area. The figure's AUC score of 0.996 indicates that there is almost perfect discrimination between the two groups. Thus, Figure 6 validates that feature reduction through RFE does not compromise model performance. Figure 7 presents the precision–recall curve, which maintains high precision across nearly the entire recall range. The curve clearly shows an average precision of 0.997, signifying minimal false-positive predictions. As depicted in Figure 7, the model performs reliably even in the presence of class imbalance, making it suitable for medical diagnostic tasks.

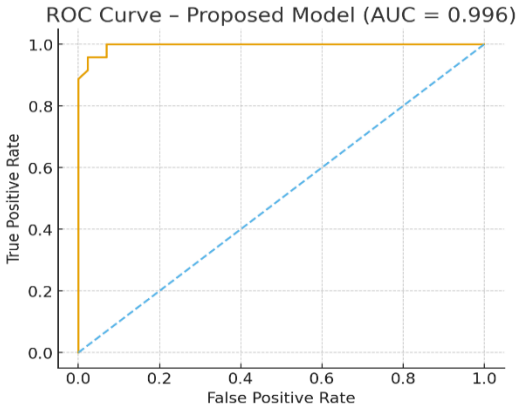

Figure 6. Proposed model-ROC curve

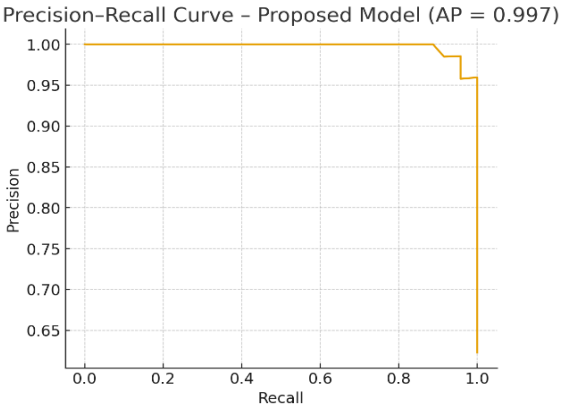

Figure 7. Proposed model-PR curve

The 18 features retained by RFE+RF are: mean radius, mean texture, mean perimeter, mean area, mean smoothness, mean concavity, mean concave points, mean symmetry, mean fractal dimension, radius error, texture error, area error, worst radius, worst texture, worst perimeter, worst area, worst smoothness, and worst concavity. These align closely with clinically interpretable morphological descriptors. The comparison of the proposed model with existing models is shown in Figure 8. The proposed model achieves accuracy almost as high as some of the existing methods, but with only 18 features. The training accuracy was 99.9%–100%, while the 5-fold CV accuracy was 96.43% (gap <0.6%), confirming minimal overfitting. The learning curves are shown in Figure 9. The importance of each feature and its contribution to the overall accuracy is shown in Figure 10.

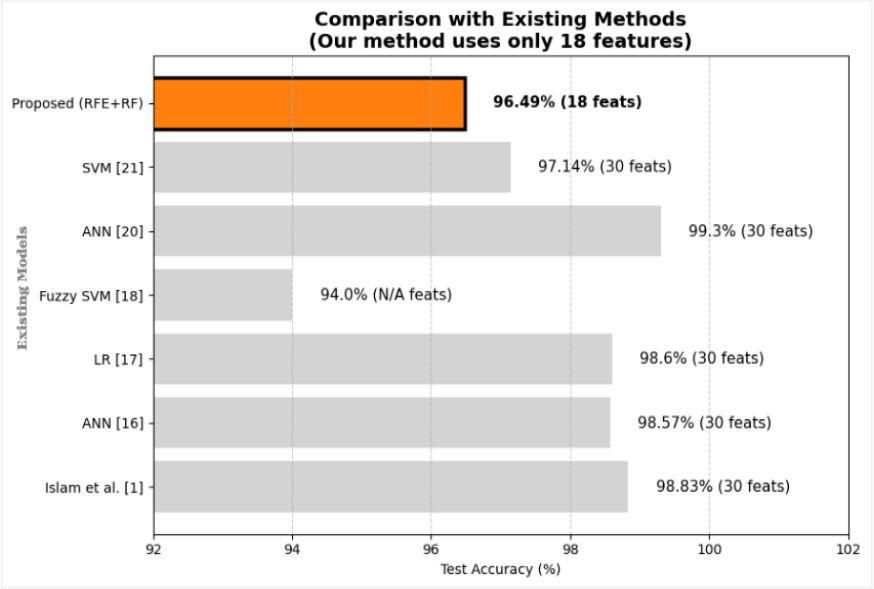

Figure 8. Comparison of the proposed model and existing models

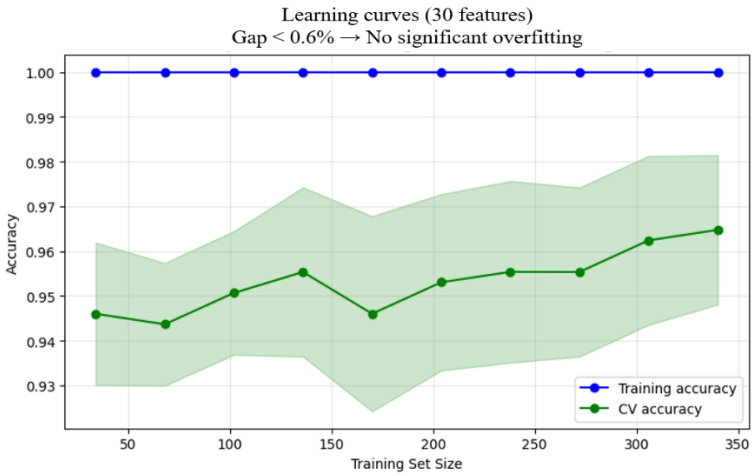

Figure 9. Learning curves

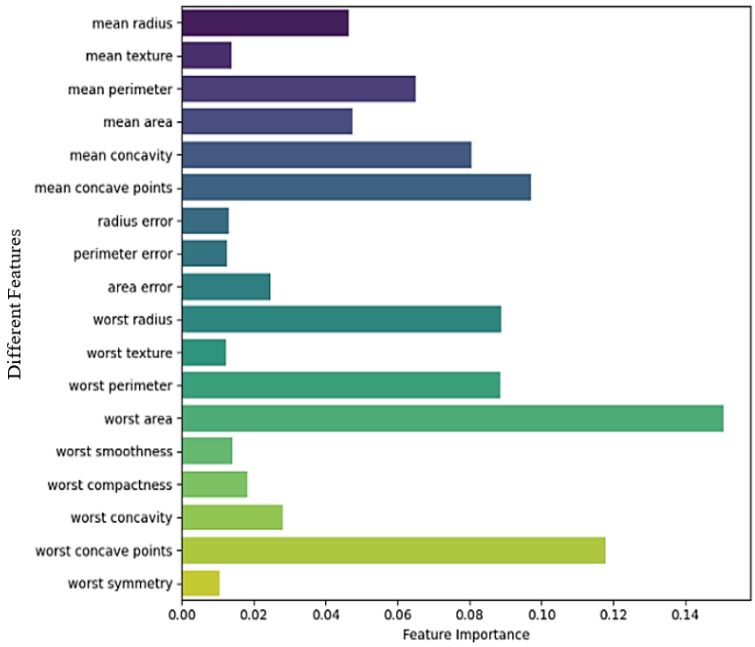

Figure 10. 18 features selected by RFE+RF

5. CONCLUSION

Conducting more accurate predictions for BC, this research underlines the importance of minimizing the characteristics of the dataset. This study helps in designing an accurate, efficient, and friendly prediction model for patients suffering from BC by applying RFE to identify the most important characteristics. Minimizing characteristics to 60% helps in increasing accuracy, accelerating test results, and providing better treatment choices for patients with BC. This approach helps patients gain access to more appropriate treatment choices and could reduce the death toll due to BC, particularly in the preliminary stages.

ACKNOWLEDGMENTS

We would like to thank CHRIST University for supporting us in completing this work.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Raghavendra Srinivasaiah	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	
Santosh Kumar Jankatti	✓	✓		✓		✓		✓	✓			✓	✓	
Niranjana Shravanabelagola Jinachandra			✓		✓	✓			✓	✓	✓	✓		
Manjunath Ramanna Lamani	✓			✓	✓		✓			✓				
Bellam Vijaya Lakshmi	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓			
Rishita Bhelwa	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study will be available in https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data.

REFERENCES

[1] S. Asif *et al.*, "Advancements and prospects of machine learning in medical diagnostics: unveiling the future of diagnostic precision," *Archives of Computational Methods in Engineering*, vol. 32, no. 2, pp. 853–883, 2025, doi: 10.1007/s11831-024-10148-w.

[2] I. Seth *et al.*, "Use of artificial intelligence in breast surgery: a narrative review," *Gland Surgery*, vol. 13, no. 3, pp. 395–411, 2024, doi: 10.21037/gs-23-414.

[3] V. Lahoura *et al.*, "Cloud computing-based framework for breast cancer diagnosis using extreme learning machine," *Diagnostics*, vol. 11, no. 2, Feb. 2021, doi: 10.3390/diagnostics11020241.

[4] T. Khater *et al.*, "An explainable artificial intelligence model for the classification of breast cancer," *IEEE Access*, vol. 13, pp. 5618–5633, 2025, doi: 10.1109/ACCESS.2023.3308446.

[5] Y. Supriya and R. Chengoden, "Breast cancer prediction using shapely and game theory in federated learning environment," *IEEE Access*, vol. 12, pp. 123018–123037, 2024, doi: 10.1109/ACCESS.2024.3424934.

[6] A. Sharma and P. K. Mishra, "Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis," *International Journal of Information Technology*, vol. 14, no. 4, pp. 1949–1960, Jun. 2022, doi: 10.1007/s41870-021-00671-5.

[7] S. Liu, H. Liu, S. Fan, L. Song, and Z. Wang, "Machine learning enables legal risk assessment in internet healthcare using HIPAA data," *Scientific Reports*, vol. 15, no. 1, Aug. 2025, doi: 10.1038/s41598-025-13720-x.

[8] T. J. Padamsee *et al.*, "Risk-management decision-making data from a community-based sample of racially diverse women at high risk of breast cancer: rationale, methods, and sample characteristics of the daughter sister mother project survey," *Breast Cancer Research*, vol. 26, no. 1, Jan. 2024, doi: 10.1186/s13058-023-01753-x.

[9] H. Aljuaid, N. Alturki, N. Alsubaie, L. Cavallaro, and A. Liotta, "Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning," *Computer Methods and Programs in Biomedicine*, vol. 223, Aug. 2022, doi: 10.1016/j.cmpb.2022.106951.

[10] M. K. D. Menon and J. Rodrigues, "Efficient ultra wideband radar based non invasive early breast cancer detection," *IEEE Access*, vol. 11, pp. 84214–84227, 2023, doi: 10.1109/ACCESS.2023.3303333.

[11] S. Chaudhury *et al.*, "Effective image processing and segmentation-based machine learning techniques for diagnosis of breast cancer," *Computational and Mathematical Methods in Medicine*, vol. 2022, pp. 1–6, Apr. 2022, doi: 10.1155/2022/6841334.

[12] M. L. Marinovich *et al.*, "Artificial intelligence (AI) for breast cancer screening: breastscreen population-based cohort study of cancer detection," *eBioMedicine*, vol. 90, Apr. 2023, doi: 10.1016/j.ebiom.2023.104498.

[13] U. Naseem *et al.*, "An automatic detection of breast cancer diagnosis and prognosis based on machine learning using ensemble of classifiers," *IEEE Access*, vol. 10, pp. 78242–78252, 2022, doi: 10.1109/ACCESS.2022.3174599.

[14] L. Sang, Z. Liu, C. Huang, J. Xu, and H. Wang, "Multiparametric MRI-based radiomics nomogram for predicting the hormone receptor status of her2-positive breast cancer," *Clinical Radiology*, vol. 79, no. 1, pp. 60–66, 2024, doi: 10.1016/j.crad.2023.09.013.

[15] B. O. Macaulay, B. S. Aribisala, S. A. Akande, B. A. Akinnuwesi, and O. A. Olabanjo, "Breast cancer risk prediction in African women using random forest classifier," *Cancer Treatment and Research Communications*, vol. 28, 2021, doi: 10.1016/j.ctarc.2021.100396.

[16] K. Guleria, A. Sharma, U. K. Lilhore, and D. Prasad, "Breast cancer prediction and classification using supervised learning techniques," *Journal of Computational and Theoretical Nanoscience*, vol. 17, no. 6, pp. 2519–2522, Jun. 2020, doi: 10.1166/jctn.2020.8924.
[17] M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon, "Predicting factors for survival of breast cancer patients using machine learning techniques," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, 2019, doi: 10.1186/s12911-019-0801-4.
[18] V. Kumar, B. K. Mishra, M. Mazzara, D. N. H. Thanh, and A. Verma, "Prediction of malignant and benign breast cancer: a data mining approach in healthcare applications," in *Advances in Data Science and Management*, 2020, pp. 435–442, doi: 10.1007/978-981-15-0978-0_43.
[19] M. Abdar *et al.*, "A new nested ensemble technique for automated diagnosis of breast cancer," *Pattern Recognition Letters*, vol. 132, pp. 123–131, Apr. 2020, doi: 10.1016/j.patrec.2018.11.004.
[20] G. G. N. Geweid and M. A. Abdallah, "A novel approach for breast cancer investigation and recognition using m-level set-based optimization functions," *IEEE Access*, vol. 7, pp. 136343–136357, 2019, doi: 10.1109/ACCESS.2019.2941990.
[21] A. Gupta, D. Kaushik, M. Garg, and A. Verma, "Machine learning model for breast cancer prediction," in *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, IEEE, Oct. 2020, pp. 472–477, doi: 10.1109/I-SMAC49090.2020.9243323.
[22] J. A. B. Hurtado, I. A. C. Albarran, M. T. Ayala, M. A. I. Manzano, L. A. M. Hernandez, and C. A. P. Ramirez, "Diagnostic strategies for breast cancer detection: from image generation to classification strategies using artificial intelligence algorithms," *Cancers*, vol. 14, no. 14, Jul. 2022, doi: 10.3390/cancers14143442.
[23] J. N. Semin, D. Palm, L. M. Smith, and S. Ruttle, "Understanding breast cancer survivors' financial burden and distress after financial assistance," *Supportive Care in Cancer*, vol. 28, no. 9, pp. 4241–4248, Sep. 2020, doi: 10.1007/s00520-019-05271-5.
[24] H. Chougrad, H. Zouaki, and O. Alheyane, "Multi-label transfer learning for the early diagnosis of breast cancer," *Neurocomputing*, vol. 392, pp. 168–180, Jun. 2020, doi: 10.1016/j.neucom.2019.01.112.
[25] J. Ahmad, S. Akram, A. Jaffar, M. Rashid, and S. M. Bhatti, "Breast cancer detection using deep learning: an investigation using the DDSM dataset and a customized alexnet and support vector machine," *IEEE Access*, vol. 11, pp. 108386–108397, 2023, doi: 10.1109/ACCESS.2023.3311892.
[26] M. Kivrak, "Breast cancer risk prediction with stochastic gradient boosting," *Clinical Cancer Investigation Journal*, vol. 11, no. 2, pp. 26–31, 2022, doi: 10.51847/21qrrkLo4Y.
[27] A. Al-Qerem *et al.*, "Feature selection in socio-economic analysis: a multi-method approach for accurate predictive outcomes," *International Journal of Crowd Science*, vol. 9, no. 1, pp. 64–78, Jan. 2025, doi: 10.26599/IJCS.2023.9100035.
[28] S. Aalaei, H. Shahraki, A. Rowhanimanesh, and S. Eslami, "Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets," *Iranian Journal of Basic Medical Sciences*, vol. 19, no. 5, 2016, pp. 476-482.
[29] J. Feng *et al.*, "Deep learning enables fast and accurate quantification of MRI-guided near-infrared spectral tomography for breast cancer diagnosis," *IEEE Transactions on Medical Imaging*, vol. 44, no. 11, pp. 4390–4403, Nov. 2025, doi: 10.1109/TMI.2025.3574727.
[30] H. Jeon and S. Oh, "Hybrid-recursive feature elimination for efficient feature selection," *Applied Sciences*, vol. 10, no. 9, 2020, doi: 10.3390/app10093211.
[31] M. L. D. Castro, A. G. Galindo, and R. Armañanzas, "Conformal recursive feature elimination," *arXiv: 2405.19429*, May 2024.
[32] X. Ding, Y. Li, and S. Chen, "Maximum margin and global criterion based-recursive feature selection," *Neural Networks*, vol. 169, pp. 597–606, Jan. 2024, doi: 10.1016/j.neunet.2023.10.037.
[33] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: a review," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 4, pp. 1060–1073, Apr. 2022, doi: 10.1016/j.jksuci.2019.06.012.
[34] B. A. Akinnuwesi, B. O. Macaulay, and B. S. Aribisala, "Breast cancer risk assessment and early diagnosis using principal component analysis and support vector machine techniques," *Informatics in Medicine Unlocked*, vol. 21, 2020, doi: 10.1016/j.imu.2020.100459.
[35] D. Houfani *et al.*, "An improved model for breast cancer diagnosis by combining PCA and logistic regression techniques," *International Journal of Computing and Digital Systems*, vol. 13, no. 1, pp. 701–716, Apr. 2023, doi: 10.12785/ijcds/130156.
[36] R. Kazerani, "Improving breast cancer diagnosis accuracy by particle swarm optimization feature selection," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, Mar. 2024, doi: 10.1007/s44196-024-00428-5.
[37] S. Jeyasingh and M. Veluchamy, "Modified bat algorithm for feature selection with the Wisconsin diagnosis breast cancer (WDBC) dataset," *Asian Pacific Journal of Cancer Prevention*, vol. 18, no. 5, pp. 1257–1264, 2017, doi: 10.22034/APJCP.2017.18.5.1257.
[38] T. E. Mathew, "A logistic regression with recursive feature elimination model for breast cancer diagnosis," *International Journal on Emerging Technologies*, vol. 10, no. 3, pp. 55–63, 2019.

BIOGRAPHIES OF AUTHORS

Raghavendra Srinivasaiah is currently working as associate professor in the Department of Computer Science and Engineering at CHRIST University, Bangalore. He completed his Ph.D. degree in Computer Science and Engineering from VTU, Belgaum, India in 2017 and has more than 21+ years of teaching experience. His interests include data mining, artificial intelligence, and big data. He can be contacted at email: raghav.trg@gmail.com.

Santosh Kumar Jankatti is currently working as associate professor in the Department of Computer Science and Technology at Dayananda Sagar University, Bangalore. He completed his Ph.D. degree in Computer Science and Engineering from VTU, Belgaum, India in 2022 and has more than 14 years of teaching experience and 3 years of IT Industry experience. His interests include data mining, artificial intelligence, and big data. He can be contacted at email: sjankatti@gmail.com.

Niranjana Shravanabelagola Jinachandra completed his Ph.D. from VTU, Belagavi in 2022. He has done his masters in machine design from VTU, Belagavi. His areas of interest are image processing, machine learning, and fluid dynamics. He is currently working as assistant professor in the Department of Mechanical Engineering at CHRIST University. He can be contacted at email: sjniranjan86@gmail.com.

Manjunath Ramanna Lamani holds a Ph.D. in Computer Science and Engineering from CHRIST University. He us currently working as associate professor in the Department of Computer Science and Engineering, Moodlakatte Institute of Technology, Kundapura, Udupi, Karnataka, India. His academic interests span deep learning, AI, IoT, and programming. He can be contacted at email: manjunathlamani01@gmail.com.

Bellam Vijaya Lakshmi completed Bachelor of Technology student in Computer Science and Engineering at Christ University, India. She has completed her studies with an academic background in technology and computing. She can be contacted at email: bellam.vijaya@btech.christuniversity.in or bellamvijaya789@gmail.com.

Rishita Bhelwa completed Bachelor of Technology in Computer Science and Engineering at Christ University, India. She has completed her studies with an academic background in computing and technology. She can be contacted at email: rishita.bhelwa@btech.christuniversity.in or rishitabhelwa2001@gmail.com.