

# Improving Performance of Ensemble Learners for Breast Cancer Detection Using Feature Engineering

Poonam Moral, Debjani Mustafi

**Cite as:** Poonam Moral, & Debjani Mustafi. (2023). Improving Performance of Ensemble Learners for Breast Cancer Detection Using Feature Engineering. International Journal of Microsystems and IoT, 1(3), 148-155. <https://doi.org/10.5281/zenodo.8354276>




© 2023 The Author(s). Published by Indian Society for VLSI Education, Ranchi, India



Published online: 21 August 2023.



Submit your article to this journal: 




Article views: 



View related articles: 



View Crossmark data: 

**DOI:** <https://doi.org/10.5281/zenodo.8354276>

Full Terms & Conditions of access and use can be found at <https://ijmit.org/mission.php>



## Improving Performance of Ensemble Learners for Breast Cancer Detection Using Feature Engineering

Poonam Moral<sup>1</sup>, Debjani Mustafi<sup>2</sup>

Birla Institute of Technology Mesra, Rnachi, Jharkhand(835215)

### ABSTRACT

Machine learning (ML) approaches include a variety of statistical and probabilistic methodologies that enable intelligent systems to be trained from repeated prior knowledge to find and recognize interesting patterns. Breast cancer (BC) is a form of tumour that grows in the tissues of the breast, and it is the most recurrent kind of disease across the world and one of the major reasons for fatality in women. Early identification of breast cancer may raise the chance of successful therapy and lower the mortality rate. In this study, the effectiveness of various ensemble approaches for the automatic prediction of breast cancer is compared and evaluated. The effectiveness of the learning process has been improved using Principal Component Analysis (PCA), a feature selection technique, by eliminating redundant and non-essential features. Diagnosis of Breast Cancer is achieved by utilizing the concept of Ensemble Learning (EL), an area of ML, including models like Gradient Boosting (GB), Adaptive Boosting (ADB), Extreme Gradient Boosting (XGB), and Random Forest (RF). The metrics that are utilized to analyze and evaluate the classifiers are ROC-AUC, Accuracy, Recall, Precision, and F1-score. The experimental results demonstrate that the Extreme Gradient Boosting is more accurate in predicting breast cancer, with an accuracy of 99.42%, compared to other ensemble learning algorithms.

### Keywords:

Machine Learning; Breast cancer; Random forest; Adaptive boosting; Extreme gradient boosting; Ensemble learning.

## 1. INTRODUCTION

The field of cancer research has been continuously advanced throughout the past few decades. According to a statistical report in 2018, India recorded almost 1,62,468 new cases of breast cancer and near about 87,090 reported death cases. According to the WHO [4], breast cancer diagnosis is one of the most significant challenges in the field of medical study. The patient's death is inextricably linked to some crucial phenomena in the behaviour of breast cancer patients. After lung cancer, BC is the second most prevalent aspect of death among women. India has fewer women who are newly diagnosed with breast cancer than the United States, but it has a higher annual mortality rate from this disease. Therefore, early detection of breast cancer is essential. There are various techniques that have been developed for the precise diagnosis of breast cancer. It is frequently quite challenging to predict BC in its early phases since the cancer cell is so small when viewed from the outside. It is probable to diagnose breast cancer at an early stage with mammography or breast screening [5]. Mammography checks the status of a woman's breast, which can be assessed by X-rays. During the breast screening, clinicians have to read a lot of imaging data, which reduces the accuracy of breast screening. In certain instances, this procedure misdiagnoses the problem in an effort to find it, which takes time as well. In order to avoid Inflammatory Breast Cancer and other related illnesses, an intelligent system would assist the medical professional in recognizing the many symptoms connected with breast cancer. ML techniques, for instance, are increasingly used in medical research fields

because of their great performance in forecasting outcomes, improving patients' health, enhancing the quality and value of healthcare, and enabling real-time decision-making processes to save lives. In this suggested work, we evaluated and analyzed the performance of various Ensemble Learning (EL) models, which include Gradient Boosting, Adaptive Boosting, Extreme Gradient Boosting, and Random Forest. The Wisconsin diagnosis breast cancer (WDBC) data set was retrieved from the UCI repository for this experiment. The remaining part of the research work is structured as follows: Section II comes up with a detailed explanation of the existing research on breast cancer detection using different ensemble learning models. Section III illustrates the dataset description, feature selection or feature extraction technique, and the theoretical idea behind each ensemble learning technique. A brief introduction related to the performance evaluation metrics is discussed in section IV. This section analyzes the findings or results from each experiment. The study given in this manuscript is concluded in section V, which also provides guidance for future enhancement.

## 2. RELATED WORK

Machine learning (ML) approaches are widely used in the domain of the health care system. Many studies have been conducted in the medical field to identify various diseases using machine learning (ML) algorithms. Our primary goal is to identify the most precise and appropriate model for predicting breast cancer. In order to do this, we have investigated several research works on Breast Cancer prediction algorithms. T. R. Mahesh et.al. [1] implemented an ensemble learning method consisting of six supervised ML classifiers such as Naïve Bayes, Decision Tree, K-Nearest Neighbour (KNN), Random

Forest, Linear Regression, and Support Vector Machine as base models, and a significant enhancement in AUC, recall, precision, accuracy, and F1-Score is observed in this research work. 98.14% accuracy was achieved by the ensemble model (voting). Artificial Neural Networks, Logistic Regression, K-nearest Neighbour, Support Vector Machine, and Random Forest classifiers are explored by M.Islam et.al. [3] on a WDBC data set where 98.57% accuracy was achieved by Artificial Neural Network classifier. M. Mohammad [2] investigated Ensemble models such as Random Forest and Extra Trees which were implemented on the WDBC data set for accurately predicting Breast Cancer. V. N. Gopal et.al. [8] suggested Breast cancer analysis with an IOT device by utilizing machine learning (ML) classifiers such as Logistic regression, Multi-layer Perceptron, and Random Forest and proved MLP to be the best algorithm with an accuracy of 98%. The research work [9] proposed by H. Daharir et.al. predicted the chance of getting breast cancer by using LR, LDA, QDA, KNN, SVM, GBN, RF, AB, and ET classifiers with a PCA feature selection technique. The effectiveness of the algorithms was evaluated using metrics including Accuracy, ROC, Precision, Specificity, and Sensitivity. The Adaptive Boosting classifier predicts breast cancer more accurately than other supervised machine learning algorithms, with an average accuracy of 98.23% for benign and malignant tumors. For the breast cancer dataset, the performance of the ensemble classifiers was examined by Assiri et al. [7]. Ensemble techniques were utilized in this work to reach 96.25% accuracy, which is higher than the accuracy measured by the Back Propagation Neural Network Method [4]. According to the findings, the SVM-RBF kernel outperforms other machine learning (ML) algorithms, achieving 96.84% accuracy in the Wisconsin dataset for predicting breast cancer. A screening technique was put out by H. Chougrad et al. [10] to improve survival rates for breast cancer in its early stages. For identifying breast cancer (BC), the authors used computer-aided diagnostics and the deep convolution approach. The most crucial factor discussed by E.Y. Kalafi et. al. [27] predicting breast cancer survivability was tumor size. Both deep learning and machine learning techniques yield acceptable prediction accuracy, but other elements like parameter settings and data transformations have an impact on the precision of the predictive model.

A significant issue in predicting breast cancer is developing a model that considers all known risk variables. Numerous studies have shown that analyzing and predicting breast cancer using different machine learning algorithms is a challenging task. Even though a lot of studies have been carried out utilizing ML approaches, the need for better findings still drives the researchers to investigate improved prediction strategies. Most of the recent approaches based on deep learning methods suffer from vanishing gradient descent problem and take significant time to converge. The Extreme Gradient Boosting algorithm works well on nonlinear and non-monotonic data which converges more rapidly with fewer steps minimizing computation costs.

In our research for predicting the likelihood of breast cancer (BC) recurrence, we implemented Extreme Gradient Boosting,

a cutting-edge boosting technique based on the decision tree classifier.

### 3. RESEARCH METHODOLOGY

This section explains the working methodology of the proposed work, which is divided into three phases: Preprocessing the data set, training the model, and assessing the trained model. The entire workflow of predicting Breast Cancer has been represented as a flowchart in Fig. 1.

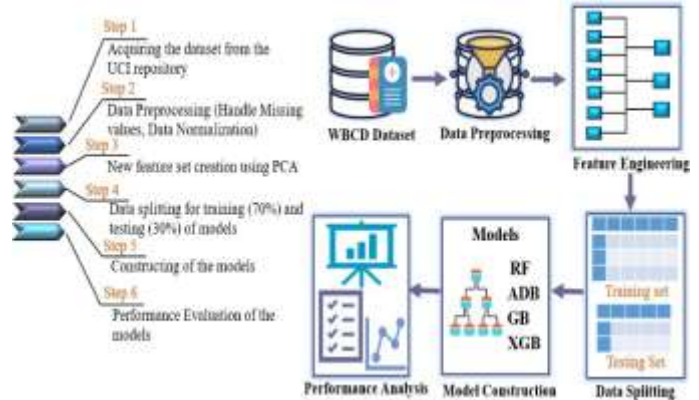


Fig. 1 Overall Framework of Breast Cancer Detection Method

#### 3.1 Dataset Description

The Wisconsin Breast Cancer Diagnostic (WBCD) data set is being used in the research to predict breast cancer. It has been sourced from the reputed UCI-Repository for Machine Learning. The WBCD dataset has a simplified size of 569\*32, where the number of observations is 569 with 32 features. The first feature is ID, and it is an identification number. The second feature is the diagnosis, and there are two diagnosis in this dataset, one of which is a malignant (cancerous) tumour and the other a benign (non-cancerous) tumour. For each cell nucleus in the dataset, 10 significant real-valued features are computed and are listed in Table 1.

For each of these ten features, the mean, standard error, and worst are computed, yielding 30 features. For instance, field 6 is Area Mean, field 16 is Area SE, and field 26 is Area Worst.

Table. 1 Dataset Description

SL No.	Attributes	Description
1.	Radius	The average distance between the center and the edge points
2.	Area	The average cancer cell areas
3.	Perimeter	The core tumour's average value
4.	Texture	Grayscale value's standard deviation
5.	Compactness	$\frac{Perimeter^2}{Area - 1}$
6.	Smoothness	Local differences in radius lengths
7.	Concave Points	Number of the contour's concave allocations
8.	Concavity	Severity of the contour's concave sections
9.	Symmetry	Assessment of Breast Symmetry
10.	Fractal dimension	Coastline approximation - 1

### 3.2 Preprocessing of Data

The Breast Cancer dataset is examined in this phase for missing, duplicate, and null values because these values have a significant influence on the accuracy computation. The datasets utilized for BC prediction have no missing, duplicate, or null values, which are used during the feature selection stage. The class diagnosis in the dataset is encoded as "M" for malignant Breast Cancer (BC) tumors and "B" for benign tumors. We have transformed "M" to "1" and "B" to "0" for our analysis in this phase. To speed up the training of the classifiers, we also scaled the features using a standard scaler, which is mathematically given by,

$$S^{new} = \frac{s-F}{\sigma_F} \quad (1)$$

where  $S^{new}$  represents the standardized value of feature F with value  $s$  and  $\sigma_F$  represents the feature F's standard deviation.

### 3.3 Feature Engineering Approach

Feature selection [11] is the process of retaining the most distinctive attributes in a given dataset. In addition to reducing the dimensions, feature selection aims to increase the genericness of our approach. One of the well-known feature selection or attribute reduction methods known as Principal Component Analysis (PCA) [9] has been used in this work. This technique computes the covariance matrix and the corresponding eigenvectors. Since PCA uses linear algebra to compress the dataset, it is one of the most effective feature reduction techniques. Implementing PCA is possible with the scikit-learn PCA class of the Python library. In a result, a number of primary components are available for selection.

### 3.4 Ensemble Learning Models

The key concept of the ensemble technique is to combine several "weak learners" to produce a "strong learner." We have implemented RF, GB, ADB, and XGB as four distinct classifiers to predict Breast Cancer. Below, we briefly discuss each of these classifiers:

#### 3.4.1 Random Forest

Regression and classification are performed using a method of ensemble learning known as Random Forest [1][2][3][8]. During training, it produces a number of decision trees, and after that, it produces a class that is the average of the classes of all the decision trees. The procedure uses an ensemble of various unique decision trees to solve problems. Each decision tree assigns a class of predictions, and the class with the most votes is chosen to serve as the prediction model.

#### 3.4.2 Adaptive Boosting

The first useful boosting method is ADB [12]. It converts many weak classifiers into reliable ones. It could be applied to various learning techniques. The yield of various algorithms is combined into a weighted total that represents the yield of the boosted algorithm despite the fact that AdaBoost is sensitive to noisy information and abnormalities. When compared to other learning algorithms, it occasionally tends to be less susceptible

to the overfitting problem.

#### 3.4.3 Gradient Boosting

Gradient Boosting [12] is an ensemble forward learning approach for classification and regression tasks that creates a prediction model in the shape of an ensemble of weak classifiers in the form of decision trees.

Gradient-boosted trees are a type of weaker decision that frequently outperforms random forests. The approach creates a model in a stage-wise manner similar to the boosting method and generalizes it by allowing optimization of any differentiated loss function.

#### 3.4.4 Extreme Gradient Boosting

One of the frequently employed ensemble learning algorithms is Extreme Gradient Boosting [12]. It is applicable to supervised learning problems like ranking, classification, and regression. XGB is designed in accordance with the Gradient boosting system's criteria and aims to stretch the limits of machine calculation to produce a flexible, compact, and accurate result.

Table 2 provides an overview of the hyperparameters employed in various machine learning algorithms.

## 4. PERFORMANCE EVALUATION

Confusion matrix (CM), which can be easily visualized and contain data on predicted (columns) and actual class (rows), are one of the most often used techniques for measuring the performance of algorithms. The outcomes of the predictions are shown as a matrix. The confusion matrix provides the number of tests records the model correctly and incorrectly predicted [13-18]. The following four outcomes are possible for each prediction:

True Positive (TPos) = The algorithm that accurately predicts that the individual has breast cancer.

False Positive (FPos) = In this instance, the model assigned the breast-canceled individual as the unaffected person.

True Negative (TNeg) = The algorithm failed to identify the breast cancer patient because it believed the patient did not have breast cancer.

False Negative (FNeg) =As the patient, in this case, has breast cancer, the model misclassified her as not having it.

From the UCI repository, information or medical records of breast cancer patients have been collected. The dataset, which consists of the medical histories of BC patients, served as the input for the prediction. A feature selection or feature reduction method, PCA, is implemented to evaluate the model's efficiency and accuracy by reducing the feature from 32 to 6. A training set and a test set are created from the dataset. 70% of the original data are designated for training purposes, while 30% are used as part of a testing set. Machine Learning (ML) approaches such as RF, ADB, GB, and XGB are used on the training set data to build the classifiers. The number of the model's predictions that are accurate and inaccurate compared to the actual values in the test data of XGB Classifier for the PCA-Component values six is shown in the Confusion Matrix (CM) in Fig. 2. A comparative review of the existing literature

alongside our study is presented in Table 3 and Table 4.

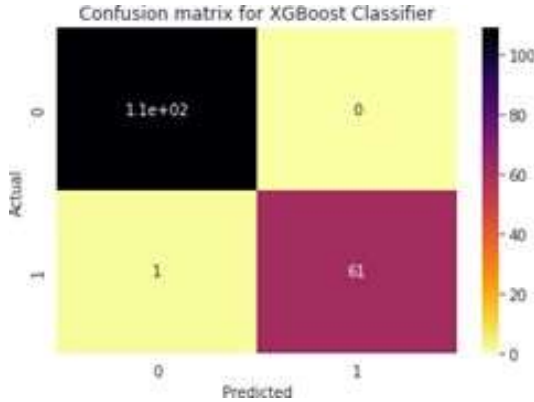


Fig. 2 CM of XGB model for PCA-Component 6

#### 4.1 Accuracy

Accuracy [6] [29] [30] is a reliable indicator of how correctly the algorithm was trained and how it would likely function in general. Fig.3 and 4 show the accuracy gained using ensemble learning approaches with various PCA Components[31-36]. In order to determine the value of accuracy, the following equation has been used:

$$Accuracy\ Acc = \frac{TPos+TNeg}{TPos+TNeg+FPpos+FNeg} \quad (2)$$

#### 4.2 Precision, Recall, and F1-Score

Precision [6] [28] is used to define the degree of accuracy in identifying the relevant results. In essence, it is the proportion of true positives to all positives. Mathematically,

$$Precision\ Pre = \frac{TPos}{TPos+FPpos} \quad (3)$$

Recall [6] [28] is defined as the proportion of correctly identified positive observations to all observations. Mathematically,

$$Recall\ Rec = \frac{TPos}{TPos+FNeg} \quad (4)$$

F1-Score [6] is the weighted mean of Precision and Recall. Therefore, both types of incorrect values are taken into account by this measurement. Mathematically,

$$F1 - Score = \frac{2*Pre*Rec}{Pre+Rec} \quad (5)$$

#### 4.3 ROC-AUC

The Receiver Operating Characteristics (ROC) [6] curve is extracted directly from the probabilistic output and is a useful technique for evaluating the performance of a classifier over all feasible decision thresholds. The most often used summary

Fig. 5, 6, and 7 represent all three metrics for PCA with component values 6.

Table. 2 Hyperparameter Tuning in Various Algorithms

Model	Hyperparameters
RF	no_estimators=200, min <sup>m</sup> _samples_leaf=4, max <sup>m</sup> _depth=70, max <sup>m</sup> _features='sqrt', min <sup>m</sup> _samples_split=10, criterion='entropy', bootstrap='True'
ADB	no_estimators=100, learning_rate=0.1
GB	learning_rate=0.1, min <sup>m</sup> _samples_split=500, subsample=0.8, min <sup>m</sup> _samples_leaf=50, max <sup>m</sup> _depth=8, max <sup>m</sup> _features='sqrt'
XGB	no_estimators=100, max <sup>m</sup> _depth=3, learning_rate=0.1, subsample=0.8, colsample_bytree=0.8

Table. 3 Comparative Analysis of XGB Algorithm Usage Across Prior Studies and This Research

Author	Dataset	FS Method	Classifier	Accuracy
Mathew [23]	WBDC	F1-Score	XGB	99.27%
Strelcenia [24]	WBDC	-	XGB	94%
Thongsuwan [25]	WBDC	-	XGB	95.9%
Likitha [26]	WBDC	F-test	XGB	98.25%
Proposed Approach	WBDC	PCA	XGB	99.42%

Table. 4 A Comparative Review of Existing Literature and Our Study

Author	Dataset	Model	Accuracy
Pati [19]	WBDC	RF	99.1%
Dey [20]	WBDC	Ensemble	97.3%
Mahendru [21]	WBDC	KNN	97.3%
Feroz [22]	WBDC	KNN & RF	97.14%
Proposed Approach	WBDC	XGB	99.42%

metric of a ROC curve is AUC. A classifier performs better when its AUC is higher. ROC-AUC Score for two PCA-Component 6 is represented in Fig. 8.

A comparison study utilizing the RF, ADB, GB, and XGB algorithms is shown in Table 5.

Table. 5 Predictive Performance of classifiers under different PCA-Components

PCA Components	Classifiers	Accuracy	Precision	Recall	F1-Score	AUC
	RF	92.40%	91.80%	87.50%	89.60%	97%

3	ADB	94.15%	90.90%	93.75%	92.30%	98%
	GB	93.57%	93.44%	89.06%	91.20%	97%
	XGB	95.90%	93.84%	95.31%	94.57%	98%
	RF	93.57%	92.54%	91.17%	91.85%	99%
4	ADB	94.15%	92.65%	92.64%	92.64%	98%
	GB	94.15%	95.31%	86.71%	92.42%	99%
	XGB	97.08%	97.01%	85.59%	96.30%	100%
	RF	97.67%	96.77%	96.78%	96.77%	100%
6	ADB	98.25%	98.36%	96.78%	97.56%	100%
	GB	98.83%	98.39%	98.39%	98.39%	100%
	XGB	<b>99.42%</b>	<b>100%</b>	<b>98.39%</b>	<b>99.19%</b>	<b>100%</b>
	RF	97.66%	95.08%	93.30%	96.6%	99%
8	ADB	96.49%	93.44%	96.61%	95.00%	99%
	GB	98.24%	95.16%	100%	97.52%	100%
	XGB	98.83%	98.30%	98.30%	98.30%	100%
	RF	95.32%	95.16%	92.18%	93.65%	98%
10	ADB	94.74%	95.08%	90.62%	92.80%	97%
	GB	94.15%	92.18%	92.18%	92.18%	97%
	XGB	96.49%	93.94%	96.88%	95.38%	98%

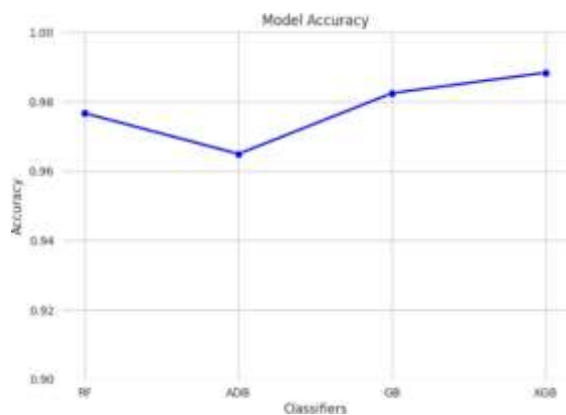


Fig. 3 Accuracy Graph for PCA-Component-8

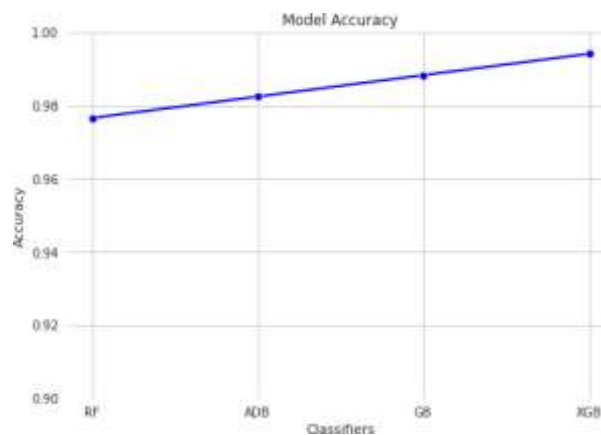


Fig. 4 Accuracy Graph for PCA-Component-6

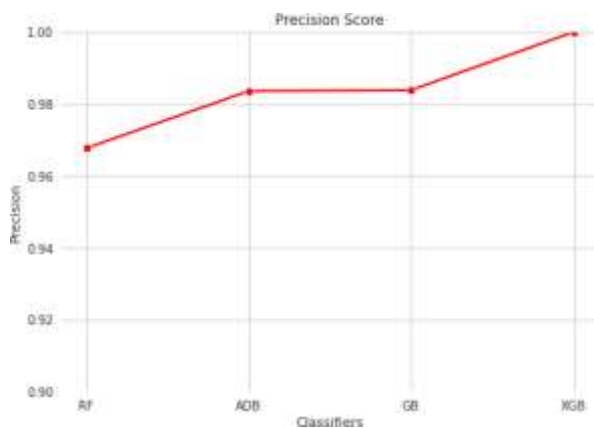


Fig.5 Precision Graph for PCA-Component-6

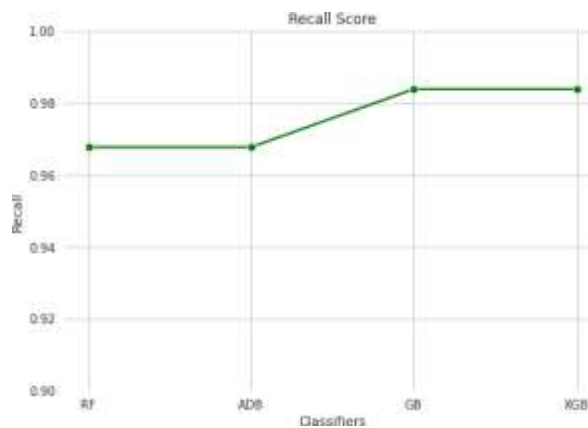


Fig. 6 Recall Graph for PCA-Component-6

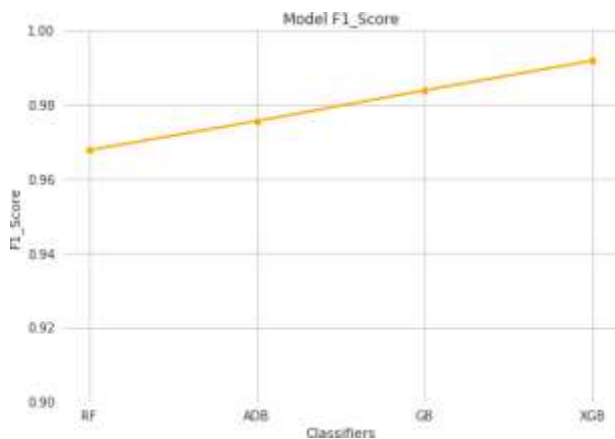


Fig. 7 F1-Score Graph for PCA-Component-6

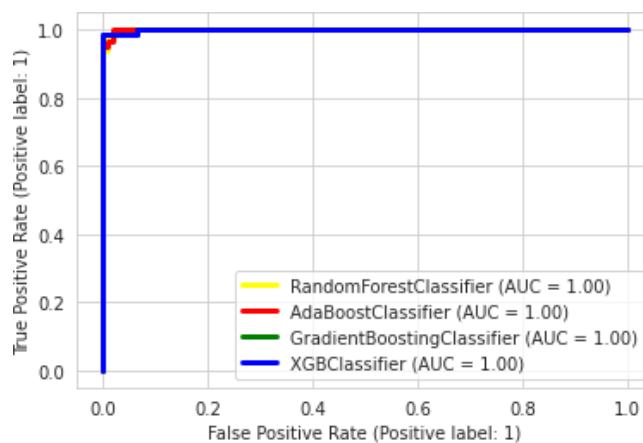


Fig. 8 AUC Graph for PCA-Component-6

## 5. CONCLUSION

In the field of medical informatics, an important problem is the automated prediction of illnesses with high predictive performance. Breast cancer is the type of cancer that occurs most frequently among women. A woman chosen at random has a 12% probability of being diagnosed with the disease. Therefore, early diagnosis of such disease can save many precious lives. This research proposes a model that compares various ensemble learning methods for identifying breast cancer on the WDBC dataset. We also attempted to compare the effectiveness of these classifiers in respect of ROC- AUC, Accuracy, Recall, Precision and F1- Score. Throughout the investigation, we observed that the Extreme Gradient Boosting (XGB) model has a high accuracy percentage for diagnosing breast cancer. The efficiency of the algorithm could be enhanced by implementing evolutionary algorithms such as differential evolution, multi-objective genetic algorithm etc. for the selection of features which is our next plan of investigation. In the near future, novel ensemble learning techniques might be developed and evaluated. Also, we still need to investigate other crucial factors that affect the likelihood of breast cancer reappearance, taking into consideration how a disease develops and the interaction between patients and their caretakers

## Acknowledgment

I would like to express my gratitude to DST-PURSE for their invaluable support and contributions to this work.

## REFERENCES

1. Mahesh, T.R., Vinoth Kumar, V., Vivek, V., Karthick Raghunath, K.M. and Sindhu Madhuri, G (2022), Early predictive model for breast cancer classification using blended ensemble learning. *International Journal of System Assurance Engineering and Management*, pp.1-10. <https://doi.org/10.1007/s13198-022-01696-0>
2. Ghiasi, M.M. and Zendejboudi, S (2021), Application of decision tree- based ensemble learning in the classification of breast cancer. *Computers in Biology and Medicine*, 128, p.104089. <https://doi.org/10.1016/j.compbimed.2020.104089>
3. Islam, Islam, M., Haque, M., Iqbal, H., Hasan, M., Hasan, M. and Kabir, M.N (2020). Breast cancer prediction: a comparative study using machine learning techniques. *SN Computer Science*, 1(5), pp.1-14. <https://doi.org/10.1007/s42979-020-00305-w>
4. Asri, H., Mousannif, H., Al Moatassime, H. and Noel, T (2016), Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, 83, pp.1064-1069. <https://doi.org/10.1016/j.procs.2016.04.224>
5. Mori, M., Akashi-Tanaka, S., Suzuki, S., Daniels, M.I., Watanabe, C., Hirose, M. and Nakamura, S (2017), Diagnostic accuracy of contrast- enhanced spectral mammography in comparison to conventional full- field digital mammography in a population of women with dense breasts. *Breast Cancer*, 24(1), pp.104-110. <https://doi.org/10.1007/s12282-016-0681-8>
6. Sokolova, M. and Lapalme, G (2009), A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4), pp.427-437. <https://doi.org/10.1016/j.ipm.2009.03.002>
7. Assiri, A.S., Nazir, S. and Velastin, S.A (2020), Breast tumor classification using an ensemble machine learning method. *Journal of Imaging*, 6(6), p.39. <https://doi.org/10.3390/jimaging6060039>
8. Gopal, V.N., Al-Turjman, F., Kumar, R., Anand, L. and Rajesh, M (2021), Feature selection and

- classification in breast cancer prediction using IoT and machine learning. *Measurement*, 178, p.109442.  
<https://doi.org/10.1016/j.measurement.2021.109442>
9. Dhahri, H., Al Maghayreh, E., Mahmood, A., Elkilani, W. and Faisal Nagi, M (2019), Automated breast cancer diagnosis based on machine learning algorithms. *Journal of healthcare engineering*.  
<https://doi.org/10.1155/2019/4253641>
  10. Chougrad, H., Zouaki, H. and Alheyane, O (2018), Deep convolutional neural networks for breast cancer screening. *Computer methods and programs in biomedicine*, 157, pp.19-30.  
<https://doi.org/10.1016/j.cmpb.2018.01.011>
  11. Li, Y., Li, T. and Liu, H (2017), Recent advances in feature selection and its applications. *Knowledge and Information Systems*, 53(3), pp. 551- 577.  
<https://doi.org/10.1007/s10115-017-1059-8>
  12. Habib, A.Z.S.B., Tasnim, T. and Billah, M.M (2019), A study on coronary disease prediction using boosting-based ensemble machine learning approaches. 2nd International Conference on Innovation in Engineering and Technology (ICIET) (pp. 1-6). IEEE.  
<https://doi.org/10.1109/ICIET48527.2019.9290600>
  13. Abdar, M., Zomorodi-Moghadam, M., Zhou, X., Gururajan, R., Tao, X., Barua, P.D. and Gururajan, R (2020), A new nested ensemble technique for automated diagnosis of breast cancer. *Pattern Recognition Letters*, 132, pp.123-131.  
<https://doi.org/10.1016/j.patrec.2018.11.004>
  14. Hajiabadi, H., Babaiyan, V., Zabihzadeh, D. and Hajiabadi, M (2020), Combination of loss functions for robust breast cancer prediction. *Computers & Electrical Engineering*, 84, p.106624.  
<https://doi.org/10.1016/j.compeleceng.2020.106624>
  15. Fatima, N., Liu, L., Hong, S. and Ahmed, H (2020), Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access*, 8, pp.150360-150376.  
<https://doi.org/10.1109/ACCESS.2020.3016715>
  16. Gu, D., Su, K. and Zhao, H (2020), A case-based ensemble learning system for explainable breast cancer recurrence prediction. *Artificial Intelligence in Medicine*, 107, p.101858.  
<https://doi.org/10.1016/j.artmed.2020.101858>
  17. Onan A. On the performance of ensemble learning for automated diagnosis of breast cancer (2015), In *Artificial intelligence perspectives and applications* (pp. 119-129). Springer.  
[https://doi.org/10.1007/978-3-319-18476-0\\_13](https://doi.org/10.1007/978-3-319-18476-0_13)
  18. Ganggayah, M.D., Taib, N.A., Har, Y.C., Lio, P. and Dhillon (2019), S.K. Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC medical informatics and decision making*, 19(1), pp.1-17.  
<https://doi.org/10.1186/s12911-019-0801-4>
  19. Pati N, Panigrahi M, Patra KC (2022), Evaluation of Different Paradigms of Machine Learning Classification for Detection of Breast Carcinoma. *In Smart and Sustainable Technologies: Rural and Tribal Development Using IoT and Cloud Computing: Proceedings of ICSST 2021 2022 Jul 28* (pp. 349-356). Singapore: Springer Nature Singapore.  
[https://doi.org/10.1007/978-981-19-2277-0\\_3](https://doi.org/10.1007/978-981-19-2277-0_3)
  20. Dey D, Jana R, Samanta PK (2023), Improved Breast Cancer Detection Using Ensembled Machine Learning Models. In *Computational Vision and Bio-Inspired Computing: Proceedings of ICCVBIC 2022 2023 Apr 8* (pp. 579-588). Singapore: Springer Nature Singapore.  
[https://doi.org/10.1007/978-981-19-9819-5\\_41](https://doi.org/10.1007/978-981-19-9819-5_41)
  21. Mahendru S, Agarwal S (2019), Feature selection using metaheuristic algorithms on medical datasets. In *Harmony Search and Nature Inspired Optimization Algorithms: Theory and Applications, ICHSA 2018 2019* (pp. 923-937). Springer Singapore.  
[https://doi.org/10.1007/978-981-13-0761-4\\_87](https://doi.org/10.1007/978-981-13-0761-4_87)
  22. Feroz N, Ahad MA, Doja F (2021), Machine learning techniques for improved breast cancer detection and prognosis—A comparative analysis. In *Applications of Artificial Intelligence and Machine Learning: Select Proceedings of ICAAIML 2020 2021* (pp. 441-455). Springer Singapore.  
[https://doi.org/10.1007/978-981-16-3067-5\\_33](https://doi.org/10.1007/978-981-16-3067-5_33)
  23. Mathew TE (2023), Breast Cancer Classification Using an Extreme Gradient Boosting Model with F-Score Feature Selection Technique. *Journal of Advances in Information Technology* ;14(2).  
<https://doi.org/10.12720/jait.14.2.363-372>
  24. Strelcenia E, Prakoowit S (2023), Effective Feature Engineering and Classification of Breast Cancer Diagnosis: A Comparative Study. *BioMed Informatics*. 2;3(3):616-31.  
<https://doi.org/10.3390/biomedinformatics3030042>
  25. Thongsuwan S, Jaiyen S, Padcharoen A, Agarwal P (2021), ConvXGB: A new deep learning model for classification problems based on CNN and XGBoost. *Nuclear Engineering and Technology*. 1;53(2):522-31.  
<https://doi.org/10.1016/j.net.2020.04.008>
  26. Likitha B, Nakka J, Verma J, Naik NS (2021), Prediction of breast cancer analysis using machine learning algorithms and xgboost technique. In *Data*



- Science and Computational Intelligence: Sixteenth International Conference on Information Processing, ICInPro 2021, Bengaluru, India, October 22–24, 2021, Proceedings 16 (pp. 298-313). Springer International Publishing.  
[https://doi.org/10.1007/978-3-030-91244-4\\_24](https://doi.org/10.1007/978-3-030-91244-4_24)
27. Kalafi EY, Nor NA, Taib NA, Ganggayah MD, Town C, Dhillon SK (2019). Machine learning and deep learning approaches in breast cancer survival prediction using clinical data. *Folia biologica*. 1;65(5/6):212-20.
  28. Ali M.M., Paul B.K., Ahmed K., Bui F.M., Quinn J.M. and Moni M.A. (2021), Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, 136, pp.104672.  
<https://doi.org/10.1016/j.combiomed.2021.104672>
  29. Sanni R.R. and Guruprasad, H.S (2021). Analysis of performance metrics of heart failed patients using python and machine learning algorithms. *Global transitions proceedings*, 2(2), pp.233-237.  
<https://doi.org/10.1016/j.gltp.2021.08.028>
  30. Rajendran R. and Karthi A (2022). Heart disease prediction using entropy based feature engineering and ensembling of machine learning classifiers. *Expert Systems with Applications*, 207, p.117882.  
<https://doi.org/10.1016/j.eswa.2022.117882>
  31. T. S Reddy, K.A. M Junaid, Y. Sukhi and Y. Jeyashree and P. Kavitha and V. Nath (2023), Analysis and design of wind energy conversion with storage system. *e-Prime - Advances in Electrical Engineering, Electronics and Energy* 100206(Vol. 17). <https://doi.org/10.1016/j.prime.2023.100206>
  32. D. Sharma, A. Rai, S. Debbarma, O. Prakash, M K Ojha and V. Nath (2023), Design and Optimization of 4-Bit Array Multiplier with Adiabatic Logic Using 65 nm CMOS Technologies, *IETE Journal of Research*, 1-14. <https://doi.org/10.1080/03772063.2023.2204857>
  33. J. Tirkey, S. Dwivedi, S. K. Surshetty, T. S. Reddy, M. Kumar, and V. Nath. (2023), An Ultra Low Power CMOS Sigma Delta ADC Modulator for System-On-Chip (SoC) Micro-Electromechanical Systems (MEMS) Sensors for Aerospace Applications. *International Journal of Microsystems and Iot*, 26–34(Vol.1). <https://doi.org/10.5281/zenodo.8186894>
  34. D. Sharma, N. Shylashree, R. Prasad, and V. Nath. (2023), Analysis of Programmable Gain Instrumentation Amplifier. *International Journal of Microsystems and Iot*, 41–47(Vol. 1). <https://doi.org/10.5281/zenodo.8191366>
  35. N. Anjum, V. K. Singh Yadav, and V. Nath. (2023). Design and Analysis of a Low Power Current Starved VCO for ISM band Application. *International Journal of Microsystems and IoT*, 82–98. (Vol. 1) <https://doi.org/10.5281/zenodo.8288193>
  36. K A Mohamed Junaid, Y Sukhi , N Anjum et.al., (2023). PV-based DC-DC buck-boost converter for LED driver. *ePrime - Advances in Electrical Engineering, Electronics and Energy*, 100271. (Vol. 5) <https://doi.org/10.1016/j.prime.2023.100271>

## AUTHORS



**Poonam Moral** received her BCA and MCA degrees from Birla Institute of Technology Mesra, Ranchi, India. She is currently pursuing PhD at the Department of Computer Science and Engineering, Birla Institute of

Technology Mesra, Ranchi, India. Her research focuses on the interdisciplinary application of machine learning to solve real-world problems in the field of medical data analysis.

**Corresponding Author Email:** [phdcs10051.21@bitmesra.ac.in](mailto:phdcs10051.21@bitmesra.ac.in)



**Debjani Mustafi** is affiliated to Birla Institute of Technology, Mesra, India. She is currently working as an Assistant Professor in the Department of Computer Science and Engineering. She has been actively associated with

academics and research. She has authored and co-authored multiple peer-reviewed journals and conferences. Her research interests include Machine Learning, Text mining, Data analysis and visualization, and Evolutionary Computing.

**E-mail:** [debjani.mustafi@bitmesra.ac.in](mailto:debjani.mustafi@bitmesra.ac.in)