

A comparative study to predict breast cancer using machine learning techniques

Shiva Shankar Reddy, Neelima Pilli, Priyadarshini Voosala, Swaroop Ravi Chigurupati

Department of Computer Science and Engineering, Sagi Rama Krishnam Raju Engineering College, Bhimavaram, India

Article Info

Article history:

Received Jan 23, 2022

Revised May 21, 2022

Accepted Jun 2, 2022

Keywords:

Adaboost

Artificial intelligence

Breast cancer

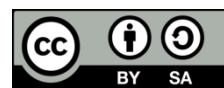
Machine learning

Multi-layer perceptron

ABSTRACT

Detection of disease at the starting stage is a very crucial problem. As the population growth increases, the risk of death incurred by breast cancer rises exponentially. Breast cancer is the most common cancer in women, and it is also the most dangerous of all cancers. Deaths because of breast cancer have been increasing in recent times. Earlier detection of the disease followed by treatment can reduce the risk and increase survival chances. There will be cases where even medical professionals can make mistakes in identifying the disease. This project deals with the detection of Breast cancer using the cell data of the tumor present in the breast. So, with the help of technologies in machine learning and artificial intelligence can substantially improve the diagnosis accuracy. The development of this project is beneficial in medical decision support systems. Several machine learning techniques, namely Adaboost, multi-layer perceptron (MLP) and stacking classifier; were used, and among all the algorithms, the stacking classifier results in the best accuracy. The accuracies 95.6%, 97.1%, and 99.2 % respectively.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Shiva Shankar Reddy

Department of Computer Science and Engineering, Sagi Rama Krishnam Raju Engineering College

Bhimavaram, Andhrapradesh, 534204, India

Email: shiva.shankar591@gmail.com

1. INTRODUCTION

Cancer is a disease that develops when the body's abnormal cells multiply uncontrollably. Instead of dying, old cells expand out of control, leading in the production of new, and abnormal cells. These extra cells may cluster together to form a tissue mass called a tumour [1]. Breast cancer (BC) occurs when malignant cells in the breast proliferate uncontrollably. BC is most typically seen in the lobules or ducts. While the milk glands are known as lobules, the milk ducts are known as ducts since they are responsible for transporting the milk from the glands to the nipple region of the breast. BC can also be found in adipose tissue and fibrous connective tissue [2]. In its early stages, BC may not cause any symptoms. Mammography may identify an anomaly even if a tumour is too small to feel. A new lump in the breast that wasn't there before is usually the first indicator of a tumour. However, not all bumps are malignant [3]. The following are symptoms of the most frequent BC: i) entire breast is covered with red, pitted skin, ii) a lump or swelling on the underneath of your arm, iii) a lump or swelling in one or both breasts, iv) a bloody discharge from your nipple, and v) a rapid, inexplicable change in the shape or size of your breast [4].

BC is divided into two types: "invasive" and "noninvasive," sometimes known as "in situ" cancer. BC is classified into five stages based on the size of the tumour(s) and the extent to which they have spread [5]:

- Stage 0: This type of BC is also known as noninvasive BC. There are no indicators that the disease has gone beyond the breast, and there are no signs that it has spread to the lymph nodes.
- Stage I: The malignancy is less than 2 cm in diameter and hasn't spread.

- Stage IIA: Tumor is in; less than 2 cm in diameter, with lymph node involvement under the arm; greater than 2 centimetres in diameter but less than 5 centimetres in diameter, with no lymph node involvement.
- Stage IIB: A tumour with a diameter of more than 5 cm and no involvement of the underarm lymph nodes. Lymph node involvement with a diameter of more than two but less than five centimetres.
- BC in Stage IIIA is also known as locally advanced breast cancer: a tumour that has progressed to lymph nodes beneath the arm or around the breastbone and is larger than 5 centimetres, any tumour with malignant lymph nodes that adhere to one another or surrounding tissue of any size.
- A tumour of any size that has migrated to the skin or chest wall is classified as Stage IIIB.
- Stage IIIC: A tumour of any size that has expanded to more lymph nodes and has spread farther.
- Stage IV BC is also known as metastatic BC. No matter what big the tumour is, it has progressed to other organs and tissues other than the breast, such as the bones, lungs, liver, brain, or distant lymph nodes [6].

2. LITERATURE REVIEW

A study on evolutionary conformal prediction for breast cancer diagnosis (BCD) was proposed by Lambrou *et al.* [7]. They created a conformal prediction based on genetic algorithm (GA) in this paper, and they used it to solve the Wisconsin breast cancer diagnosis (WBCD) challenge. Saber *et al.* [8] created an Innovative deep learning (DL) approach to automated recognition and classification of BC using the transfer-learning technique (TLT). The data was pre-processed to reduce noise, enhance intensity in breast images, eliminate non-breast regions, and detect the cancerous area. Osareh and Shadgar [9] used three well-known classifiers, support vector machine (SVM), K-nearest neighbour (K-NN), and probabilistic neural network (NN) to examine the challenges of BC detection and prognostic risk appraisal of recrudescence and metastasis. Feature selection techniques for the detection of BC through clinical data were employed by Haq *et al.* [10]. Using machine learning (ML) algorithms and clinical data, they proposed a new BC detection approach. For associated feature selection from the data set, supervised (relief) and unsupervised methods were utilized. Bharat *et al.* [11] employed ML algorithms to detect and diagnose BC risk. Depending on the dataset and parameter selection, each method performs differently. The K-NN strategy has produced the best outcomes in terms of the overall methodology. A comparative study of ML algorithms for BCD and detection was provided by Bazazeh and Shadgar [12]. Here, the performance of: RF, SVM, and BN was evaluated and compared using the WBCD data set. A study on ML Classifiers in BCD was reported by Teixeira *et al.* [13]. They offered a series of classification models in their research, attempting to discover the best model to classify BC based on the data set WBCD. They chose five different ML techniques. Asri *et al.* [14] employed ML algorithms to predict and diagnose BC risk. The WBCD datasets used four key algorithms: SVM, Naïve Bayes (NB), k-NN, and decision tree (DT) C4.5. Yi and Yi [15] reported their work in the diagnosis of BC using the DT model mixed with feature selection. Various studies on the WBCD data set's various training test divisions were carried out. They eliminated certain highly relevant variables from the DT model to reduce complexity and then chose tumor as a subset of the DT model, diameter, cell morphology consistency, single epithelial cell size, and mitosis following data correlation and independence tests. Using ML, Al-sammarraie and Ibrahim [16] proposed BC in fine-needle aspiration pictures. With a 68% accuracy, the mammography lesion categorization model was effectively applied. The number of photographs accessible was the project's most significant restriction.

Nemissi *et al.* [17] used an upgraded extreme learning machine (ELM) based-NN to BCD. They employed a neural classification method to diagnose BC in this study. For the hidden neurons, they employed sigmoid activation functions with various settings. WBCD Dataset was the data set they used. The suggested classification system outperformed the traditional EL network in terms of Generalization while using fewer hidden neurons. Arora *et al.* [18] proposed a ML algorithm for BCD Predictive Analysis. To achieve these goals, they deployed supervised learning techniques. Based on the dataset they used, they saw improved outcomes in the NB, SVM, RF, K-NN, and DT algorithms. RF was the best algorithm that worked under all settings, followed by K-NN, with others doing marginally better or worse. For BCD, Khuriwal and Mishra [19] used an adaptive voting ensemble ML Algorithm. They've also developed a way for identifying BC using ensemble ML. It is demonstrated that using ANN in conjunction with a logistic algorithm is effective. Efficient BC prediction using ensemble ML models has been proposed by Naveen *et al.* [20]. Unbiased ensemble models are used to improve system performance. They compared the prediction evaluation of six ML methods, including DT, SVM, MLP, K-NN, LR, and RF, with and without ensemble techniques. A comparative study of ML algorithms for BC prediction was given by Sengar *et al.* [21]. They used two ML algorithms, DT classifier and LR, to predict BC and evaluated their accuracies to see which one was the most accurate.

BC risk prediction using XGBoost and RF algorithm was proposed by Kabiraj *et al.* [22]. The BC dataset from the UCI ML repository is analyzed using ensemble ML methods such as RF and XGBoost. The DM was classified among the DM patients whether they are having DM or not by using various ML and DL techniques [23], [24]. By using these models [25], [26] they can predict whether the patient was suffering with the exact problem or not [27], [28]. There is a chance to suffer with ailments of the DM to BC also [29], [30]. A Novel NNa real-time biopsy-based automated system for the diagnosis of BC was used by Singh *et al.* [31]. They suggested a unique ANN-based strategy for the automated identification of BC in this paper. The WDBC database was used to test supervised learning techniques of NN to diagnose BC. BC malignancy prediction using DLNN was proposed by Prakash and Visakha [32]. This paper examined the use of DL techniques in the computer-aided diagnosis of BC. For the UCI examination, the WBCD dataset was used. To avoid overfitting, the model was optimized by early halting and dropouts. Thomas *et al.* [33] used ML algorithms to do comparative analysis to predict BC. On the WBCD dataset, which is publicly available on the internet, the authors used six different ML algorithms to predict BC in an advanced stage, including DT, NB, LR, RF, SVM, and ANN. Shankar *et al.* [34] predicted whether the patient is having BC or not by using RNN. Gupta *et al.* [35] has classified the data by using five ML algorithms. Yue *et al.* [36] aimed to examine the use of ML techniques in the diagnosis and prognosis of BC. They began by outlining ML techniques such as ANNs, SVMs, DTs, and k-NNs. Then they look into how they can be used to treat BC.

3. METHOD

3.1. Objectives

To see which characteristics are most helpful in predicting whether a cancer is malignant or benign. To get an effective model, a suitable dataset is considered that comprises of features related to breast cancer. To predict whether the patient is effected or not by using deep learning models by considering performance of evaluation metrics.

3.2. Dataset description

The dataset used in this project is Wisconsin BC dataset has collected from the UCI repository. The dataset is available at the link: <https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/>. The dataset has 569 instances in total. It is necessary to check for 32 different characteristics in order to find any signs of breast cancer. In Table 1, all the column names in the dataset have listed. The range of values in which the corresponding feature lies has been tabulated below with feature names.

3.3. Data acquisition

The collected dataset from Kaggle has 569 rows and 32 columns. The column named diagnosis is the target value out of those column values. The rest of the columns or features are used for training the model.

3.4. Data pre-processing

The selected dataset is subjected to pre-processing techniques. The column named id is dropped as it is not providing valuable data for predicting the disease. While data pre-processing, an issue called Imbalances Classes was encountered. To resolve this issue, oversampling techniques were used. The data oversampling technique named SMOTETomek from a library named imbLearn was used. By this procedure, the tuples for both classes became almost the same.

3.5. Splitting of dataset

The dataset is downloaded from UCI repository and it has been separated into two parts. The train set is made up of about 80% of the over-sampled data, while the test set is made up of the remaining 20%. The entire dataset description was shown in Table 1.

3.6. Models

3.6.1. Adaptive boosting

Adaptive boosting, simply AdaBoost, is an ensemble boosting algorithm. In this AdaBoost classifier algorithm, each incorrectly classified instance or row is reassigned with higher weights. Here our trained dataset contains 30 features; it makes 30 decision trees. The cases incorrectly typed in our first model are given more priority and input to the second model. Steps involved in AdaBoost classifier:

Step 1: sample weights are initialized.

Step 2: one model is selected as base learner from the created decision trees with each feature.

Step 3: total error occurred for the base learner is calculated:

$$totalerror = \frac{incorrectly\ classified\ instances}{total\ number\ of\ instances}$$

Step 4: Performance of the stump base learner is calculated:

$$Significance = \frac{1}{2} \log \frac{1 - totalerror}{totalerror}$$

where Significance= Performance of the stump.

Step 5: update the sample weights so that the next tree will take the errors from the preceding tree as input. For Incorrectly classified instances:

$$Newweight = Oldweight * e^{significance}$$

for correctly classified instances:

$$Newweight = Oldweight * e^{-significance}$$

the sum of the updated weights must be equal to 1. To make the sum 1, normalized weights are calculated for each instance:

$$Normalizedweight = \frac{new\ weight}{sum\ of\ all\ the\ updated\ weights}$$

Step 6: new dataset is created with the normalized weights, and again a decision tree or stump is made based on the new dataset.

Step 7: repeat the steps from 2 to 6 until it sequentially passes through all the stumps and finds the less error.

Step 8: the test dataset will pass through all the decision trees created by the algorithm. The outcome would then be determined by a majority of votes cast between the stumps.

Table 1. Dataset description

S.No	Column Name	Description	S.No	Column Name	Description
1	id	Unique Value to identify each row	17	smoothness_se	0<=value<=0.03
2	diagnosis	Type of Cancer: Malignant and Benign-B	18	compactness_se	0<=value<=0.14
3	radius_mean	6.98<=value<=28.1	19	concavity_se	0<=value<=0.4
4	texture_mean	9.71<=value<=39.3	20	concave_points_se	0<=value<=0.05
5	perimeter_mean	43.8<=value<=189	21	symmetry_se	0.01<=value<=0.08
6	area_mean	14.4<=value<=2500	22	fractal_dimension_se	0<=value<=0.03
7	smoothness_mean	0.05<=value<=0.16	23	radius_worst	7.93<=value<=36
8	compactness_mean	0.02<=value<=0.35	24	texture_worst	12<=value<=49.5
9	concavity_mean	0<=value<=0.43	25	perimeter_worst	50.4<=value<=251
10	concave points_mean	0<=value<=0.2	26	area_worst	185<=value<=4250
11	symmetry_mean	0.11<=value<=0.3	27	smoothness_worst	0.07<=value<=0.22
12	fractal_dimension_mean	0.05<=value<=0.1	28	compactness_worst	0.03<=value<=1.06
13	radius_se	0.11<=value<=2.87	29	concavity_worst	0<=value<=1.25
14	texture_se	0.36<=value<=4.88	30	concave_points_worst	0<=value<=0.29
15	perimeter_se	0.76<=value<=22	31	symmetry_worst	0.16<=value<=0.66
16	area_se	6.8<=value<=542	32	fractal_dimension_worst	0.06<=value<=0.21

3.6.2. Multi-layer perceptron

The MLP classifier is used to tackle this classification task, which requires matching a given input tuple to one of two target classes. To improve the model's accuracy, multiple combinations of parameters, such as the number of hidden layers, hidden nodes, and epochs, are employed to train it. artificial neurons (NN or ANN) are made up of a collection of connected units or nodes MLP is a multi-layered and completely integrated FFN. Backpropagation, a supervised learning approach, is used to train the network in Figure 1. There are three levels in the basic model: one input, one (or more) hidden, and one output layers. Each of the

30 predictor or independent variables (x_i) in the training set is linked to a random weight (W_{ij}) that runs from the input layer's i th node to the hidden layer's j th node, where I and j are the neuron indices.

The weighted sum (u_i) is computed by the summation function of (1) and fed to the activation function which calculates the hidden layer output as shown in (2).

$$\text{WeightedSum} = u_i = \sum_{i=1}^n (x_i W_{ij}^h) \tag{1}$$

$$\text{OutputofHiddenLayer} = h_i = f(u_i) = \frac{1}{1+e^{-u_i}} \tag{2}$$

The nonlinear sigmoid function ($f()$) is used in the hidden layer. The output of the hidden layer is fed into the output layer's nodes as input. The output layer likewise computes the weighted sum average using the summing function in (3), which is indicated by v_i , and feeds it to the activation function to generate the layer's output (y_i) in (4).

$$\text{WeightedSum} = v_i = \sum_{i=1}^n (h_i W_{ij}^y) \tag{3}$$

$$\text{OutputofOutputLayer} = y_i = f(v_i) = \frac{1}{1+e^{-v_i}} \tag{4}$$

Where h_i denotes the hidden layer's output, and W_{ij} indicates the weight from the i th hidden layer node to the j th output layer node (i and j are indices of neurons). The backpropagation learning approach is based on reducing the error between the network's actual output (Y_i) and the desired output ($Y_i(d)$) as much as possible. The following is how the error E in (5) is determined using the Euclidean function:

$$E = \frac{1}{2} (y_i(d) - y_i)^2 \tag{5}$$

here, E shows the error of the i^{th} node of the output layer. To normalize the error, the weight of the neuron (6) must be updated as follows:

$$W_{ij}^{K+1} = W_{ij}^K - \eta \frac{\partial E}{\partial W_{ij}} \tag{6}$$

where, W_{ij}^{K+1} is updated weight, W_{ij}^K is old weight and $\frac{\partial E}{\partial W_{ij}}$ is the rate of change of error concerning weight.

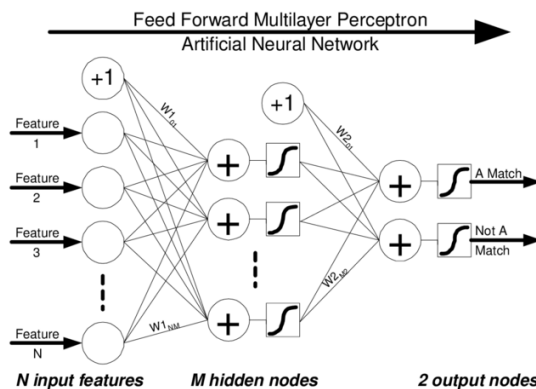


Figure 1. Architecture of MLP

Here $N=30$ input features. Two Output Nodes are the "B" class labels encoded as 0 and "M" encoded as 1. The following is a description of the Multilayer Perceptron Neural Network's learning algorithm:

- Step 1: It is to read the breast cancer input split.
- Step 2: Use random values to set the weight of the inputs.
- Step 3: Evaluate the summation function.
- Step 4: Evaluate the function of activation.

Step 5: Subtract the desired output to get the prediction error E.

Step 6: Take the error averaged over all of the training cases.

Step 7: Send the error back through the network and determine the error gradient as a function of weight changes.

Step 8: Adjust or update the weight to reduce the mistake.

3.6.2. Stacking classifier

Stacked generalization, called Stacking, is an ensemble technique. Here, this algorithm combines the multiple classifications by using the meta-classifier. In stacking, there are two levels of models called level-0 models (also called base-level models) and level-1 models (also called meta-models). Unlike bagging and boosting in the Stacking architecture, the base-level models are all different learning algorithms, i.e., heterogeneous models are used and shown in Figure 2. In level-0, all the models are applied on the same training dataset, and the predictions made by those models are given as input (features) to the meta-model. Here, one important thing is that the training data set is split into the k parts, then the level-0 models have trained on the (k-1) parts of the training dataset, and then the last part is used to make the predictions. Then the actual data is used to train the base models to calculate the performance on the test set. The predictions from base classifiers and the expected outputs make the input and output pairs train the level-1 model. The level-0 or base-level models used for the classification of BC are SVM, RF, Naive-Bayes, and the meta-classifier used for the stacking classifier is logistic regression.

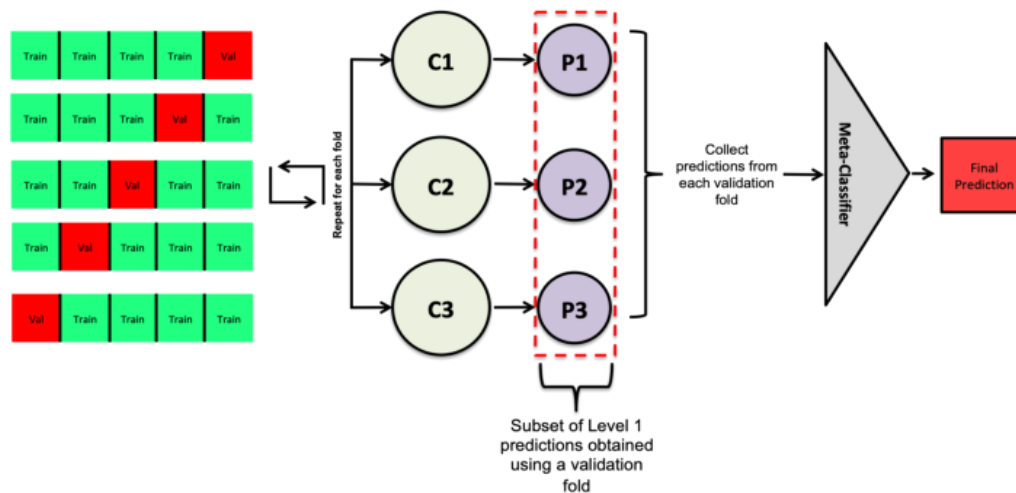


Figure 2. The architecture of stacking classifier

Steps involved in the stacking classifier:

Step 1: divide the data into two sets: training and testing. The Training set is then separated into K-folds, similar to how k-fold cross-validation works.

Step 2: the base model is trained using the (K-1) train sets, and then the validation set is used for the predictions.

Step 3: the process continues until all the folds have been predicted.

Step 4: the base model is then trained using the whole train dataset for calculating the performance on test data.

Step 5: repeat the steps from 2 to 4 for all the base models.

Step 6: predictions made by the Base models are used as features for the meta classifier.

Step 7: the meta model is then used to make predictions on the test dataset.

4. RESULTS ANALYSIS

The model's performance is evaluated based on the efficiency and error that occurred using evaluation metrics such as accuracy, precision, recall, f1-score, and specificity. These evaluation metrics can be calculated by using the confusion matrix. A confusion matrix summarizes prediction results on the classification problem, as shown in Figure 3.

		Predicted 0	Predicted 1
Actual 0	TN	FP	
Actual 1	FN	TP	

Figure 3. Confusion matrix

By using the confusion matrix, we have evaluated the following metrics:

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

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

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

$$Specificity = \frac{TN}{TN + FP}$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

in Table 2 all the Results obtained for the three algorithms are provided. Figure 4 shows the comparison graph for all the three algorithms of evaluation metrics. By observing the graph, the value of accuracy obtained for Stacking Classifier 0.9927 is more than the other algorithms.

Table 2. Comparison of metrics for all the applied models

Model/ Metric	Accuracy	Precision	Recall	Specificity	F1-Score
AdaBoost classifier	0.9565	0.9863	0.9350	0.9836	0.9599
Multi-layer perceptron	0.9710	0.9861	0.9594	0.9843	0.9725
Stacking classifier	0.9927	0.9866	1.00	0.9843	0.9932

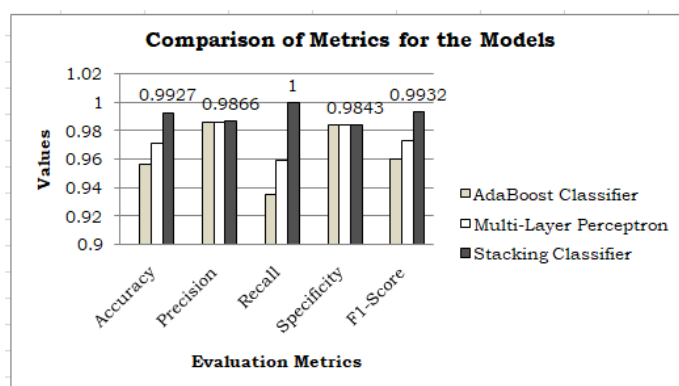
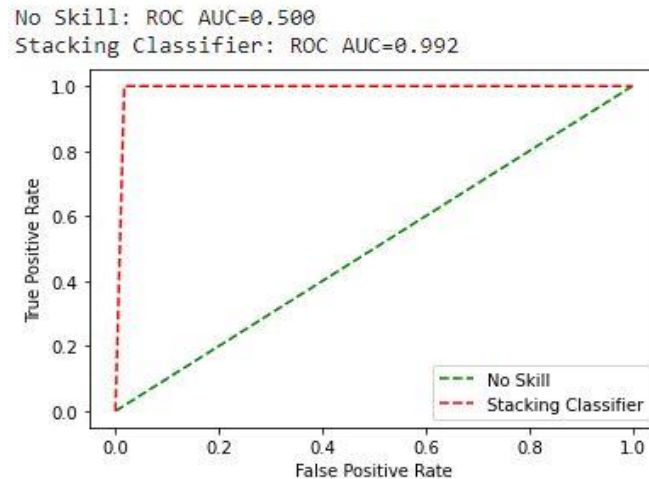
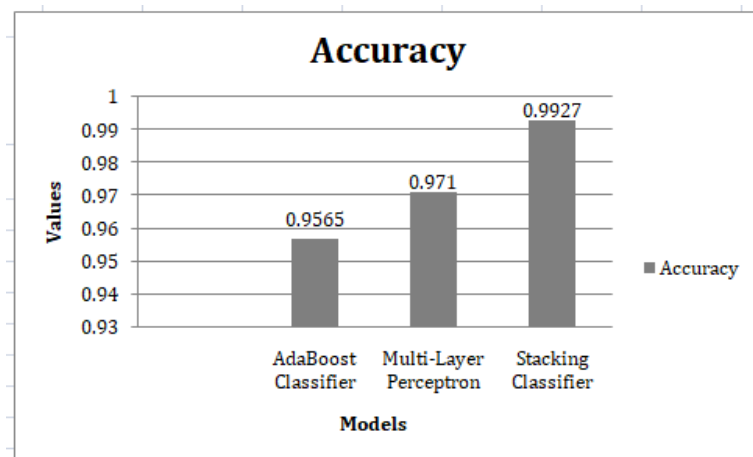


Figure4. Comparison of metrics for the various models

From Figure 5, we can observe the graphs obtained for stacking classifier and other models; (a) shows the AUCROC curve for the stacking classifier and (b) shows the graph for accuracy of all the models. By observing the results the stacking classifier is highest compared with other models.



(a)



(b)

Figure 5. Curve for Stacking classifier and accuracy models for the graph, (a) AUCROC curve for stacking classifier and (b) Comparison graph for accuracy with the models

5. CONCLUSION





The type of Breast Cancer is predicted by concerning the cell data of the Breast cells of a patient. Breast cancer is detected using the Adaboost Classifier, Multi-Layer Perceptron and Stacking classifier. The stacking classifier achieved the best efficiency with an accuracy of 99.20%. We can expand this effort in the future by working on large datasets with more real-time attributes and integrating the model into the real-time website.

REFERENCES





- [1] M. Kapalczyńska *et al.*, "2D and 3D cell cultures—a comparison of different types of cancer cell cultures," *Archives of medical science: AMS*, vol. 14, no. 4, p. 910, 2018, doi: 10.5114/aoms.2016.63743.
- [2] W. E. Barlow *et al.*, "Performance of diagnostic mammography for women with signs or symptoms of breast cancer," *Journal of the National Cancer Institute*, vol. 94, no. 15, pp. 1151-9, Aug. 2002, doi: 10.1093/jnci/94.15.1151.
- [3] M. R. Narasingarao *et al.*, "Prediction of the breast cancer disease using machine learning techniques—a comparison," *Harbin Gongye Daxue Xuebao/Journal of Harbin Institute of Technology*, vol. 54, no. 1, pp. 126-33, 2002.
- [4] G. H. Rauscher, C. E. Ferrans, K. Kaiser, R. T. Campbell, E. E. Calhoun, and R. B. Warnecke, "Misconceptions about breast lumps and delayed medical presentation in urban breast cancer patients," *Cancer Epidemiology and Prevention Biomarkers*, vol. 19, no. 3, pp. 640-647, 2010.
- [5] B. A. Oppong and T. A. King, "Recommendations for women with lobular carcinoma in situ (LCIS)," *Oncology-Melville*, vol. 25, no. 11, p. 1051, 2011.
- [6] Y. S. Sun, "Risk factors and preventions of breast cancer," *International journal of biological sciences*, vol. 13, no. 11, p. 1387, 2017.

- [7] A. Lambrou, H. Papadopoulos and A. Gammerman, "Evolutionary conformal prediction for breast cancer diagnosis," in *2009 9th International Conference on Information Technology and Applications in Biomedicine*, 2009, pp. 1-4, doi: 10.1109/ITAB.2009.5394447.
- [8] A. Saber, M. Sakr, O. M. Abo-Seida, A. Keshk, and H. Chen, "A novel deep-learning model for automatic detection and classification of breast cancer using the transfer-learning technique," in *IEEE Access*, vol. 9, pp. 71194-71209, 2021, doi: 10.1109/ACCESS.2021.3079204.
- [9] A. Osareh and B. Shadgar, "Machine learning techniques to diagnose breast cancer," in *2010 5th International Symposium on Health Informatics and Bioinformatics*, 2010, pp. 114-120, doi: 10.1109/HIBIT.2010.5478895.
- [10] A. U. Haq *et al.*, "Detection of breast cancer through clinical data using supervised and unsupervised feature selection techniques," in *IEEE Access*, vol. 9, pp. 22090-22105, 2021, doi: 10.1109/ACCESS.2021.3055806.
- [11] A. Bharat, N. Pooja, and R. A. Reddy, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," in *2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C)*, 2018, pp. 1-4, doi: 10.1109/CIMCA.2018.8739696.
- [12] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," in *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, 2016, pp. 1-4, doi: 10.1109/ICEDSA.2016.7818560.
- [13] F. Teixeira, J. L. Z. Montenegro, C. A. da Costa, and R. R. Righi, "An analysis of machine learning classifiers in breast cancer diagnosis," in *2019 XLV Latin American Computing Conference (CLEI)*, 2019, pp. 1-10, doi: 10.1109/CLEI47609.2019.235094.
- [14] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Computer Science*, vol. 83, pp. 1064-1069, 2016, doi:10.1016/j.procs.2016.04.224.
- [15] L. Yi and W. Yi, "Decision tree model in the diagnosis of breast cancer," in *2017 International Conference on Computer Technology, Electronics and Communication (ICCTEC)*, 2017, pp. 176-179, doi: 10.1109/ICCTEC.2017.00046.
- [16] L. H. A. al-sammarrife and A. A. Ibrahim, "Predicting breast cancer in fine needle aspiration images using machine learning," in *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2020, pp. 1-4, doi: 10.1109/ISMSIT50672.2020.9254891.
- [17] M. Nemissi, H. Salah, and H. Seridi, "Breast cancer diagnosis using an enhanced extreme learning machine based-neural network," in *2018 International Conference on Signal, Image, Vision and their Applications (SIVA)*, 2018, pp. 1-4, doi: 10.1109/SIVA.2018.8661149.
- [18] M. Arora, S. Som, and A. Rana, "Predictive analysis of machine learning algorithms for breast cancer diagnosis," in *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2020, pp. 835-839, doi: 10.1109/ICRITO48877.2020.9197945.
- [19] N. Khuriwal and N. Mishra, "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm," in *2018 IEEMA Engineer Infinite Conference (eTechNXT)*, 2018, pp. 1-5, doi: 10.1109/ETECHNXT.2018.8385355.
- [20] Naveen, R. K. Sharma, and A. R. Nair, "Efficient breast cancer prediction using ensemble machine learning models," in *2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*, 2019, pp. 100-104, doi: 10.1109/RTEICT46194.2019.9016968.
- [21] P. P. Sengar, M. J. Gaikwad and A. S. Nagdive, "Comparative study of machine learning algorithms for breast cancer prediction," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2020, pp. 796-801, doi: 10.1109/ICSSIT48917.2020.9214267.
- [22] S. Kabiraj *et al.*, "Breast cancer risk prediction using xgboost and random forest algorithm," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, pp. 1-4, doi: 10.1109/ICCCNT49239.2020.9225451.
- [23] M. C. Cheang *et al.* "Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype," *Clinical cancer research*, vol. 14, no. 5, pp. 1368-1376, 2008.
- [24] B. G. Haffty *et al.* "Locoregional relapse and distant metastasis in conservatively managed triple negative early-stage breast cancer," *Journal of clinical oncology*, vol. 24, no. 36, pp. 5652-5657, 2006.
- [25] R. G. Dumitrescu and I. Cotarla, "Understanding breast cancer risk-where do we stand in 2005?," *Journal of cellular and molecular medicine*, vol. 9, no. 1, pp. 208-221, 2005.
- [26] Y. M. Coyle, "The effect of environment on breast cancer risk," *Breast cancer research and treatment*, vol. 84, no. 3, pp. 273-88, 2004.
- [27] C. M. Friedenreich and A. E. Cust, "Physical activity and breast cancer risk: impact of timing, type and dose of activity and population subgroup effects," *British journal of sports medicine*, vol. 42, no. 8, pp. 636-647, 2008.
- [28] J. Lin, J. E. Manson, I. M. Lee, N. R. Cook, J. E. Buring, and S. M. Zhang, "Intakes of calcium and vitamin D and breast cancer risk in women," *Archives of Internal Medicine*, vol. 167, no. 10, pp. 1050-1059, 2007.
- [29] S. S. Reddy, N. Sethi, and R. Rajender, "Safe prediction of diabetes mellitus using weighted conglomeration of mining schemes," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2020, pp. 1213-1220.
- [30] S. S. Reddy, R. Rajender, and N. Sethi, "A data mining scheme for detection and classification of diabetes mellitus using voting expert strategy," *International Journal of Knowledge-Based and Intelligent Engineering Systems*, vol. 23, no. 2, pp. 103-8, 2019.
- [31] S. Singh, J. Harini, and B. R. Surabhi, "A novel neural network based automated system for diagnosis of breast cancer from real time biopsy slides," *International Conference on Circuits, Communication, Control and Computing*, 2014, doi:10.1109/cimca.2014.7057755.
- [32] S. S. Prakash and K. Visakha, "Breast cancer malignancy prediction using deep learning neural networks," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2020, pp. 88-92, doi: 10.1109/ICIRCA48905.2020.9183378.
- [33] T. Thomas, N. Pradhan, and V. S. Dhaka, "Comparative analysis to predict breast cancer using machine learning algorithms: a survey," in *2020 International Conference on Inventive Computation Technologies (ICICT)*, 2020, doi:10.1109/iciict48043.2020.9112464.
- [34] R. S. Shankar, K. V. Murthy, and C. S. Rao, "Breast cancer disease prediction with recurrent neural networks (RNN)," *International Journal of Industrial Engineering & Production Research*, vol. 31, no. 3, pp. 379-86, 2020.
- [35] R. S. Shankar, V. M. Gupta, K. V. Murthy, and C. S. Rao, "Breast cancer data classification using machine learning mechanisms," *Indian Journal of Public Health Research & Development*, vol. 10, no. 5, 2019.
- [36] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, "Machine learning with applications in breast cancer diagnosis and prognosis," *Designs*, vol. 2, no. 2, doi: 10.3390/designs2020013.





BIOGRAPHIES OF AUTHORS

Shiva Shankar Reddy     is Assistant Professor at Department of Computer Science and Engineering in Sagi RamaKrishnam Raju Engineering College, Bhimavaram, Andhrapradesh, India. He is pursuing PhD degree in Computer Science and Engineering with specialization in Medical Mining, Machine Learning. His research areas are Image Processing, Medical Mining, Machine Learning, Deep Learning and Pattern Recognition. He was published 30+ papers in International Journals and Conferences. S.S. Reddy has filed 04 patents. His research interests include Image Processing, Medical Mining, Machine Learning, Deep Learning and Pattern Recognition. He can be contacted at email: shiva.shankar591@gmail.com.







Neelima Pilli     is Assistant Professor at Sagi Ramakrishnam Raju Engineering College, Department of Computer Science and Engineering, India. She Received B.Tech degree in Sagi Ramakrishnam Raju Engineering College, Department of Computer Science and Engineering in 2006, She Holds a M. Tech degree in Sagi Ramakrishnam Raju Engineering College, Department of Computer Science and Engineering in 2010. Her research areas are Cloud Computing, Fog Computing, Edge Computing, Machine Learning and IoT. She can be contacted at email: neelima.p47@gmail.com.



Priyadarshini Voosala     is Assistant Professor at Sagi Ramakrishnam Raju Engineering College, Department of Computer Science and Engineering, India. She Received B. Tech degree in Sri Vishnu Engineering College for Women, Department of Computer Science and Engineering in 2005. She Holds a M. Tech degree in Sagi Ramakrishnam Raju Engineering College, Department of Computer Science and Engineering in 2010. Her research areas are Cloud Computing, Fog Computing, Edge Computing, Machine Learning and Image processing. She can be contacted at email: priyavoosala@gmail.com.



Swaroop Ravi Chigurupati     is Assistant Professor at Sagi Ramakrishnam Raju Engineering College, Department of Computer Science and Engineering, India. He Received B.Tech degree in Sagi Ramakrishnam Raju Engineering College, Department of Information Technology in 2012. He Holds a M.Tech degree in Sagi Ramakrishnam Raju Engineering College, Department of Information Technology in 2018. His research areas are Image Processing, Bioinformatics, Machine Learning, Deep Learning, and Data Mining. He can be contacted at email: raviswaroop.chigurupati@gmail.com