Classification of Breast Cancer Using Computational Machine Learning Algorithms

P. Gagana, Darshan Aladakatti, Ashwini Kodipalli, Trupthi Rao, and Shoaib Kamal

1 Literature Survey

In the conference paper "Breast Cancer Detection and Diagnosis Using AI" by Khalid Shaikh, we get to understand the depth of AI and its powerful techniques in medicine and healthcare. This paper emphasizes the early detection and diagnosis of breast cancer with its benefits and risks, respectively. The steps taken by Khalid Shaikh are image acquisition, and image processing where the quality of the image is improved for gaining clarity in understanding and interpretation by making use of various techniques. Processing is followed by image segmentation where the entire image is divided into segments to observe and study in detail, then the features of the segmented images are extracted, and based on these features the images of the patients are classified as either normal or abnormal (Shaikh et al. 2021; Feng et al. 2019; Das et al. 2022; Vergis et al. 2021; Vandana and Radhika 2021; Sakib et al. 2022; https://www.webmd.com/cancer/features/top-cancer-killers).

In the paper "Accurate Prediction of Neoadjuvant Chemotherapy Pathological Complete Remission (PCR) for the Four Sub-Types of Breast Cancer", Xin Feng along with his co-ordinates expresses his thoughts on different stages of breast cancer, and its subtype which is detected based on the size of real nodes. Xin Feng talks about the treatment as well which is considered to be the best

P. Gagana · S. Kamal
Electronics and Communication Engineering, Global Academy of Technology, Bangalore, India

D. Aladakatti · S. Kamal
Computer Science and Engineering, Global Academy of Technology, Bangalore, India

A. Kodipalli (✉) · T. Rao · S. Kamal
Artificial Intelligence and Data Science Engineering, Global Academy of Technology, Bangalore, India
e-mail: dr.ashwini.k@gat.ac.in

T. Rao
e-mail: trupthirao@gat.ac.in

N. R. Shetty et al. (eds.), *Advances in Communication and Applications*, Lecture Notes in Electrical Engineering 1105, https://doi.org/10.1007/978-981-99-7633-1_20

treatment option that is Neoadjuvant chemotherapy also famous known in the medical field as NAC. The results after the prediction were evaluated using performance metrics called Avc. A minimum of 0.7594 Avc was achieved for all four subtypes of breast cancer discussed in (https://analyticsindiamag.com/primer-ensemble-learning-bagging-boosting/; https://www.kaggle.com/code/jagannathrk/predicting-breast-cancer-logistic-regression/data; https://www.geeksforgeeks.org/python-seaborn-pairplot-method/; Rachana et al. 2022a; Guha et al. 2022a; Rao et al. 2022).

In the paper, "Comparison-Based Analysis and Prediction for Earlier Detection of Breast Cancer Using Different Supervised ML Approach" a comparative study was made by Soumen Das and his fellow authors on one of the deadliest diseases, breast cancer. Soumen Das opted to make use of binary classification and chose nine classification models out of which only three of those gave positive variances. Naive Bayes gave a favorable variable as 0.01433, followed by SVM-polynomial that gave 0.01931 and SVM-sigmoid that gave 0.00034 (Sagarnal et al. 2021; Bhoomika et al. 2022; Kodipalli et al. 2022a, 2022b; Ruchitha et al. 2021).

In the paper titled "Decision Support System for Breast Cancer Detection Using Biomarker Indicators" the author Spiridon Vergis develops a mobile application where the patients can enter the required prompted inputs which get tested as new data to the classification algorithms like Logistic Regression popularly known as LR, Naive Bayes, Support Vector Machine familiarly called as SVM, followed by gradient boosting classification. Among these gradient boosting classification provides the best results (Ruchitha et al. 2022; Gururaj et al. 2022; Guha et al. 2022b).

Modern-day treatments are based on the concept of molecular heterogeneity, which is why they can be influenced by various features. Early breast cancer, which is usually confined to the breast, is considered to be a curable disease. This stage is commonly known as benign, which in our dataset is represented as B.

However, when the symptoms are not detected and treated at the early stage the treatment becomes more complicated and cannot to treated using the present therapeutic options and might have an associated risk of spreading to other parts of the body. This condition is called malignant and this is represented as M in our dataset (Rachana et al. 2022b; Zacharia and Kodipalli 2022; Kodipalli and Devi 1804; Bhagwani et al. 2021; Dhanush et al. 2021; Raj et al. 2021).

2 Methodology

We have worked with two algorithms one is the simplest one and another is the most accurate one.

Machine learning has seven main steps such as

(i) Gathering the data
(ii) Preparing data
(iii) Training the model

(iv) Improving the performance

2.1 Gathering the Data

This step includes acquiring the data in the form of a dataset from either an open source or from a medical institute. We have collected our dataset from an open-source website called kaggle (https://www.kaggle.com/code/jagannathrk/predicting-breast-cancer-logistic-regression/data). We selected one such dataset which will classify the given data as either benign or malignant based on 32 features and train on 569 records.

2.2 Preparing Data

The data which we acquire may or maybe be in an ideal form so we need to process it according to our requirements. With our dataset, we first do Exploratory Data Analysis (EDA) look for the presence of null values, and remove them as they will not contribute in any way to classify the tumor as benign or malignant. We then checked for outliers which renders the accuracy, following which we count the number of values for both benign and malignant to make sure we have nearly the same values. If there is a huge difference in the values of these classes then our classification might get biased toward the class with the higher count. We have value counts of benign and malignant as 357 and 212, respectively. Noting that we can say that the difference is minimal and in an acceptable range. Now, we identified the categorical data and replaced them with numerical values for easy and uniform processing by making use of Label encoder. For more reliable and easy understanding we need to scale the values within the range of 0 to 1 which was done using normalization.

To represent our dataset pictorially we plotted a bivariate distribution using a pairplot present in the seaborn library (https://www.geeksforgeeks.org/python-seaborn-pairplot-method/). Pairplot gives us multiple plots of two classes, in which the diagonal plots are univariate. The pairplot was constructed for five columns seen in Fig. 1.

2.3 Training the Model

We have worked on one of the simplest algorithms which are Logistic Regression algorithm, which is tuned using a suitable technique called SVC and the most accurate one Random Forest algorithm which is one bagging technique algorithm and we have also worked on Boosting algorithms like Adaptive boosting algorithm popularly known as Adaboost, Gradient boosting algorithm, and eXtreme gradient boosting algorithm popularly known as XG-boost algorithm. We shall see each one in detail.

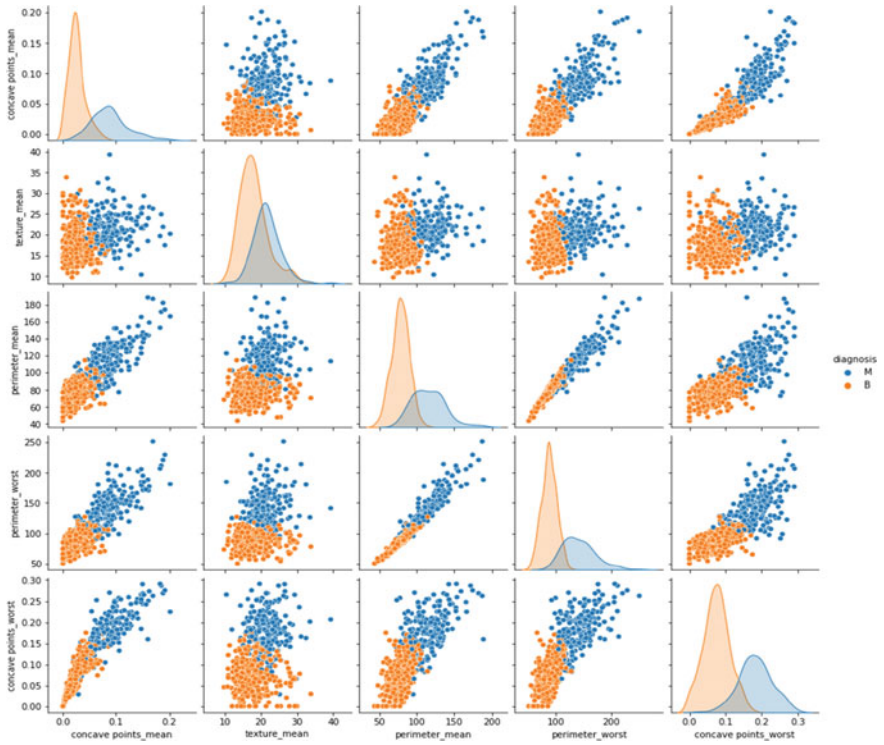

Fig. 1 Pairplot of concave points_mean, texture_mean, perimeter_mean, perimeter_worst, concave points_worst

2.3.1 Logistic Regression

Logistic Regression is a supervised machine learning technique, it is mainly used for classification problems. The value of logistics is a regression between 0 and 1 so it forms the S-like curve, this s-form curve is known as a sigmoid function or the logistic function.

$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

We had two classes to bifurcate, to do the same we used Support Vector Classification popularly known as SVC which is present in Sci-kit learn library of python. Our dataset 1 represents malignant which states that breast cancer can be dangerous and spreads to other parts of the patient's body. Whereas 0 represents benign which states that the cancer is at its early stage and can be treated completely.

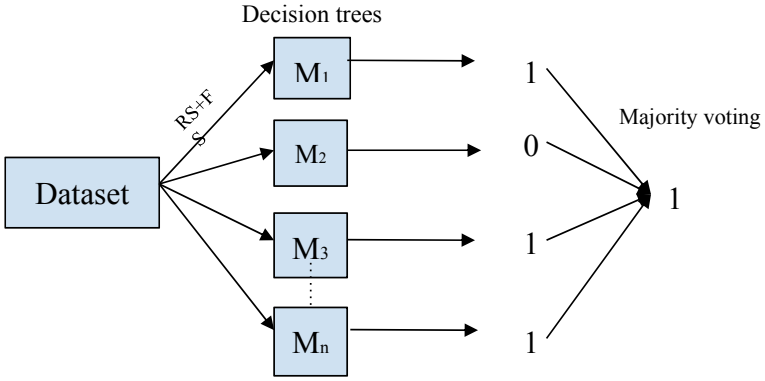

Fig. 2 Representation of bagging technique in Random Forest algorithm

The accuracy obtained from Logistic Regression was 93.57%. We needed to prove the accuracy, for which we used a tuning technique after which our accuracy improved by 4.09% and gave us 97.66%.

2.3.2 Random Forest

Random Forest algorithm is the most widely used algorithm, it is a combination of many decision tree models each of the models will be provided with a sample dataset.

In Fig. 2, we can observe that every decision tree model is provided with a sample dataset, and every new sample dataset is obtained by Row sampling and feature sampling methods. Every decision tree model gives an output and then a voting classifier is applied, the majority of the votes given by the models will be considered.

2.3.3 Adaptive Boosting Algorithm

Ada boosting algorithm combines many weak learners and makes it a strong learner as shown in Fig. 3. Here the base learners are decision trees which are called stomps, for every feature in the dataset sample one stomp will be created. In the first step, every record will be assigned a sample weight, sample weight is calculated using the formula

$$w = \frac{1}{n}$$

where n is the number of records

Initially all the records are assigned the same weights.

In step 2 the model with less entropy or Gini value will be selected as the first base learning model, the total error of the misclassified record in stomp is calculated

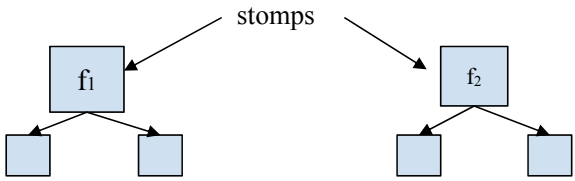

Fig. 3 Block diagram of weights distribution in adaptive boosting algorithm

as

$$\text{Total error} = \frac{1}{\text{sum of all the weights}}$$

In step 3 the performance of the stomp that is how the model has been classified. To calculate that we use

$$\text{performance of stomp} = \frac{1}{2}\log_e\left(\frac{1-\text{total error}}{\text{total error}}\right)$$

The wrong classified data is only sent to the next sequential model, the weights are updated to every stomp, the weight is increased for the wrong classified record, and the weight is increased in case of the correct classified record.

To update the weight of an incorrectly classified record:

$$\text{new sample weight} = \text{weight} * e^{\text{performance value}}$$

To update the weight of correctly classified record

$$\text{new sample weight} = \text{weight} * e^{(-\text{performance value})}$$

For the second base learning model the records from the updated normalized weight are selected as the new dataset and again the same process will be done. Whenever the new data is given it is classified based on the majority voting technique.

2.3.4 Gradient Boosting Algorithm

In gradient boosting algorithm ensembling happens by optimizing the loss, hence it is classification problem for building the first tree it considers the log of odds ratio, log of odds ratio can be calculated as:

$$\text{odds ratio} = \frac{\text{probability of success}}{\text{probability of failure}}$$

Then probability will be calculated using:

$$\text{probability} = \frac{e^{\text{odds ratio}}}{1 + e^{\text{odds ratio}}}$$

Extra column will be added to the dataset as predicted value with one of either classes based on the threshold value. Then residuals are calculated for the first base tree with respect to the predicted value column. The second tree will be built on independent values and residuals as the target column. Considering the learning rate specified, the model will again update the odds ratio and the whole will be repeated again.

2.3.5 Extreme Gradient Boosting Algorithm [XG-Boost]

XG-boost is a decision tree-based machine learning algorithm that uses a gradient boosting framebook. As there are two classes namely malignant and benign in the dataset the average probability will be 0.5 and this probability is used to calculate the residuals. The decision tree is constructed as a binary classifier, leaf nodes will always be two. Similarity score will be calculated using:

$$\text{similarity score} = \frac{\sum(\text{residuals})^2}{\sum \text{probability} * (1 - \text{probability}) + \lambda}$$

gain is calculated by:

$$\text{gain} = \text{left similarity} + \text{right similarity} - \text{root similarity}$$

Post pruning method is to build the final decision tree, the post pruning is done based on cover value:

$$\text{cover value} = \sum[p(1 - p)]$$

New probability will be calculated based on a new dataset using the sigmoid function on the base model. This process is continued until the value of residual becomes a very small or maximum number.

(iv) Improving the performance:

To improve the performance of the models we have applied hyperparameter tuning for each and every algorithm based on the observations performance of every model was increased. We have used the GridSearchCV approach for hyperparameter tuning, this approach searches for the best hyperparameters from the grid of given parameter values.

3 Results

Results before tuning the model as shown in Fig. 4 and in Table 1.

Results after tuning the models as shown in Fig. 5 and in Table 2.

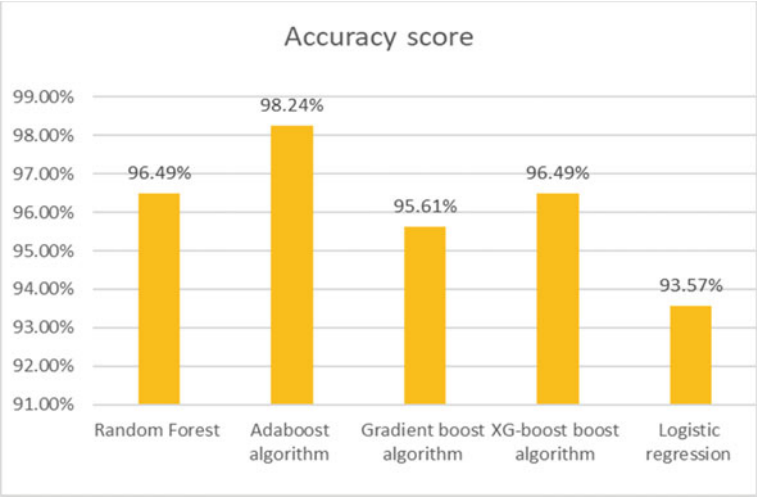

Fig. 4 The plot of the accuracy score of all the algorithms used before tuning

Table 1 The accuracy score of all the algorithms used before tuning

Algorithms	Accuracy scores (%)
Random Forest	96.49
Adaboost algorithm	98.24
Gradient boost algorithm	95.61, 96.49 (with exponential function)
XG-boost boost algorithm	96.49
Logistic regression	93.57

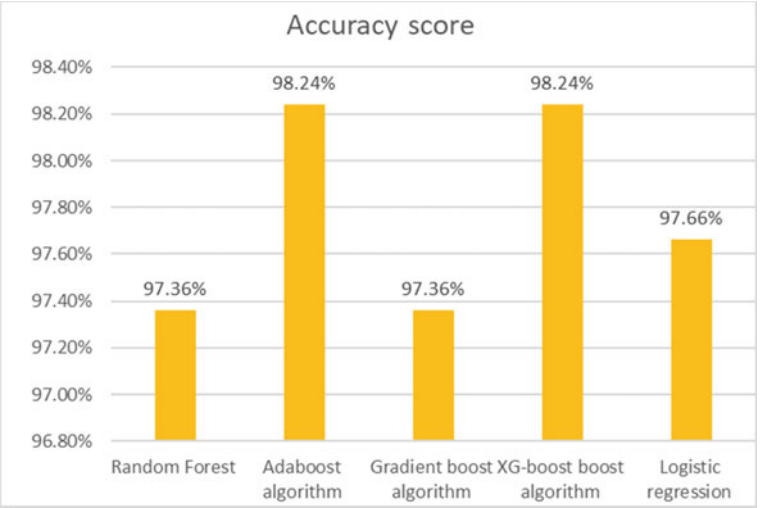

Fig. 5 The plot of accuracy score of all the algorithms used after tuning

Table 2 Accuracy score of all the algorithms used after tuning

Algorithms	Accuracy scores (%)
Random Forest	97.36
Adaboost algorithm	98.24
Gradient boost algorithm	97.36
XG-boost boost algorithm	98.24
Logistic regression	97.66

4 Conclusion

We hereby conclude that the prediction of breast cancer and classification of cancer as either benign or malignant was successful and the highest accuracy was obtained by Adaboost. Other algorithms namely Random forest, Gradient boost, XG-boost, and Logistic Regression improved their accuracy after tuning and gave us comparatively greater results, where Random Forest gave 97.36% which was 0.87% better than the earlier similarly Gradient boost gave 97.36% which was 1.75% better, XG-boost gave 98.24% which had improved by 1.75%, followed by Logistic Regression that gave 97.66% that improved 4.09%. The future work of this paper shall be to build a mobile application and make the correct and efficient medication available to every woman when the required data matches the symptoms of breast cancer. Along with this, the application should have a legend should have a list of hospitals specialized in the field of treating breast cancer.

References

Bhagwani H, Agarwal S, Kodipalli A, Martis RJ (2021, December) Targeting class imbalance problem using GAN. In: 2021 5th international conference on electrical, electronics, communication, computer technologies and optimization techniques (ICEECCOT). IEEE, pp 318–322

Bhoomika R, Shreya Shahane, Siri T C, Trupthi Rao, Dr. Ashwini K, Pradeep Kumar Chodon, "Ensemble Learning Approaches for Detecting Parkinson's Disease", 2022.

Das S, Chatterjee S, Sarkar D, Dutta S (2022) Comparison-based analysis and prediction for earlier detection of breast cancer using different supervised ML approach

Dhanush N, Prajapati PR, Revanth M, Ramesh R, Kodipalli A, Martis RJ (2021, September) Prediction of gold price using deep learning. In: 2021 IEEE 9th region 10 humanitarian technology conference (R10-HTC). IEEE, pp 1–5

Feng X, Song L, Wang S (2019) Accurate prediction of neoadjuvant chemotherapy pathological complete remission (PCR) for the four subtypes of breast cancer

Guha S, Kodipalli A, Rao T (2022) Computational deep learning models for detection of COVID-19 using chest X-ray images

Guha S, Kodipalli A, Rao T (2022) Computational deep learning models for detection of COVID-19 using chest X-ray images. In: Emerging research in computing, information, communication and applications: proceedings of ERCICA 2022. Springer Nature Singapore, Singapore, pp 291–306

Gururaj V, Shriya VR, Ashwini K (2019) Stock market prediction using linear regression and support vector machines. Int J Appl Eng Res 14(8):1931–1934

Gururaj V, Ramesh SV, Satheesh S, Kodipalli A, Thimmaraju K (2022) Analysis of deep learning frameworks for object detection in motion. Int J Knowl-Based Intell Eng Syst 26(1):7–16

https://analyticsindiamag.com/primer-ensemble-learning-bagging-boosting/

https://www.geeksforgeeks.org/python-seaborn-pairplot-method/

https://www.kaggle.com/code/jagannathrk/predicting-breast-cancer-logistic-regression/data

https://www.webmd.com/cancer/features/top-cancer-killers

Kodipalli A, Devi S (2021) Prediction of PCOS and mental health using fuzzy inference and SVM. Front Public Health 1804

Kodipalli A, Guha S, Dasar S, Ismail T (2022) An inception-ResNet deep learning approach to classify tumours in the ovary as benign and malignant. Exp Syst e13215

Kodipalli A, Devi S, Dasar S, Ismail T (2022) Segmentation and classification of ovarian cancer based on conditional adversarial image to image translation approach. Exp Syst e13193

Rachana PJ, Kodipalli A, Rao T (2022) Comparison between ResNet 16 and Inception V4 network for Covid-19 prediction

Rachana PJ, Kodipalli A, Rao T (2022) Comparison between ResNet 16 and Inception V4 network for COVID-19 prediction. In: Emerging research in computing, information, communication and applications: proceedings of ERCICA 2022. Springer Nature Singapore, Singapore, pp 283–290

Raj A, Umrani NR, Shilpashree GR, Audichya S, Kodipalli A, Martis RJ (2021, July) Forecast of covid-19 using deep learning. In: 2021 IEEE international conference on electronics, computing and communication technologies (CONECCT). IEEE, pp 1–5

Rao T, Devamane S, Moumen A (2022) Machine learning approaches for stratification of Parkinson's disease

Ruchitha PJ, Richitha YS, Kodipalli A, Martis RJ (2021, December) Segmentation of ovarian cancer using active contour and random walker algorithm. In: 2021 5th international conference on electrical, electronics, communication, computer technologies and optimization techniques (ICEECCOT). IEEE, pp 238–241

Ruchitha PJ, Sai RY, Kodipalli A, Martis RJ, Dasar S, Ismail T (2022, October) Comparative analysis of active contour random walker and watershed algorithms in segmentation of ovarian cancer. In: 2022 international conference on distributed computing, VLSI, electrical circuits and robotics (DISCOVER). IEEE, pp 234–238

Sagarnal C, Devamane SB, Hosamani R, Rao T (2021) Deep learning approaches for COVID-19 diagnosis

Sakib S, Yasmin N, Tanzeem AK, Shorna F, Hasib KMd, Alam SB (2022) Breast cancer detection and classification: a comparative analysis using machine learning algorithms

Sanjana S, Sanjana S, Shriya VR, Vaishnavi G, Ashwini K (2021) A review on various methodologies used for vehicle classification, helmet detection and number plate recognition. Evol Intel 14(2):979–987

Shaikh K, Krishnan S, Thanki R (2021) Artificial intelligence in breast cancer early detection and diagnosis. Exclusive license to Springer Nature Switzerland AG

Vandana L, Radhika K (2021) Detailed review on breast cancer diagnosis using different ML algorithms

Vergis S, Bezas K, Exarchos TP (2021) Decision support system for breast cancer detection using biomarker indicators

Zacharia S, Kodipalli A (2022) Covid vaccine adverse side-effects prediction with sequence-to-sequence model. In: Emerging research in computing, information, communication and applications: proceedings of ERCICA 2022. Springer Nature Singapore, Singapore, pp 275–281