INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

# Breast Cancer using Machine Learning Techniques -A Survey

L.Srinivasa Rao [1], D Purnima [2], R Veeranja Neyulu [3]
[1] [2]Associate professor, [2] Assistant Professor, [3] Assistant professor
Department of Computer Science Engineering
Priyadarshini Institute of Technology & Science, Tenali,Guntur

Abstract - the critical challenge of predicting breast cancer, a leading cause of mortality among women. The paper explores the application of various machine learning (ML) algorithms, including Naïve Bayes, Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Decision Tree, and ensemble techniques like Random Forest, Adaboost, and XGBoost, for breast cancer prediction. These ML methods are essential for doctors and pathologists, providing automated tools to differentiate between malignant and benign tumors. The study evaluates these algorithms using different performance measures and finds that both Decision Tree and XGBoost classifiers achieve the highest accuracy at 97%. Additionally, XGBoost demonstrates the highest Area Under the Curve (AUC) value, indicating its excellent performance in breast cancer prediction with an AUC of 0.999. This research underscores the effectiveness of ML techniques in aiding medical professionals in breast cancer diagnosis and decision-making.

## 1. Introduction

GLOBOCAN 2020 report shows that how cancer is more dangerous as it recorded total 19.3 cases in that 10 million deaths in 2020 [1,2].Female breast cancer has recorded with 2.3 million new cases [1] and also indicates the mortality rate also increased. Breast cancer created huge impact on the life of women. Mortality rate can be decreased by increasing awareness, early prediction and diagnosis [3]. Today medical field generates large amount data of different diseases and that data helps to perform analysis and further predictions. Technology helps lot to doctors [4] pathologist for performing an accurate prediction that helps to avoid

ISSN: 2057-5688

further medical expenditure and to get proper treatment and in some cases early detection can save a person's life [5]. Different techniques have been used to classify breast cancer. Classification techniques play an important role in breast cancer prediction and diagnosis. Many research have demonstrated the importance of breast cancer prediction with different techniques and challenges [6,7], and they used a variety of data mining classification approaches to analyze the Wisconsin Diagnostic Breast Cancer (WDBC)[8,9] and Wisconsin Breast Cancer(WBC) [10]dataset and it found substantial result. Many data mining techniques [11] are used for classification such as KNN, NB, DT, SVM, LR, ensemble techniques such as RF, Adaboost, XGBoost. Our aim is to predict breast cancer by applying different classification algorithm and compare their performance and find out best algorithm. This paper organizes in following manner: Section 2 describes similar kind of work performed before on breast cancer datasets, section 3 describe the architectural overview of system with dataset, and classification techniques details, section 4 describes results and discussion of ML techniques with different performance measures.

## 2. Related Work

The provided text outlines the ongoing work in breast cancer prediction utilizing various machine learning techniques. Several studies have explored different algorithms such as KNN, NB, DT, SVM, LR, Random Forest, XGBoost, EXSA, and ensemble models for breast cancer diagnosis and risk prediction. Researchers have achieved notable accuracies, with methods like ANN, KNN, NB, SVM, and RF reaching high percentages in accuracy, often above 95%. Feature selection, ensemble models, and techniques like genetic programming have also been employed to enhance the performance of machine learning classifiers in breast cancer prediction. Moreover, the text highlights the importance of data preprocessing techniques such as scaling to standardize features and ensure that all characteristics are on the same magnitude level. Additionally, the text mentions the use of both linear and nonlinear approaches for data reduction

ISSN: 2057-5688

based on the dataset's feature correlations. This summary indicates the diverse array of machine learning techniques applied in breast cancer prediction, showcasing a range of sophisticated methods to improve accuracy and advance the field of medical diagnosis. Obaid et al. presented that SVM with quadratic kernel function achieved highest accuracy [9].Both linear and nonlinear approaches to data reduction are viable options, and which one is used will depend on the characteristics of the correlations that exist between the features of the dataset [18].The vast majority of ML algorithms, on the other hand, will calculate the Euclidean distance between any two data points they are given. It is necessary to bring all of the characteristics down to the same magnitude level. Scaling is a method that can be used to accomplish this goal [19].Ara et al. performed feature selection and apply ML on breast cancer data set for classification into benign and malignant and by using SVM and RF they got 96.5% accuracy [20].Jabbar et al. classifying the breast cancer data, an Ensemble model has been constructed in this work by utilising Bayesian networks and Radial Basis Function producing accurate classifications optimal accuracy of 97.42 % by combining these two different types of classifiers.Mahesh et al. used different machine learning techniques and also blended ensemble learning for breast cancer prediction and achieved 98.14% accuracy.

## 3. Methodology

overall design of proposed methodology applied for detection of breast cancer. Different classification algorithms applied on breast cancer data but different classifier shows different performance on same data therefore we used an ensemble technique that uses bagging and boosting which combines results from different classifier also learns from previous classifiers. To perform this ,first step of this is data acquisition. The data then pre-processed for selection of attributes, after that data divided: 80% for training and 20 % for testing. Dataset is labelled dataset having labels malignant and benign and therefore supervised different

ISSN: 2057-5688

classification techniques applied on training data for building a model. Test data evaluated by using different classifier and finally compare the performance of different classifiers.

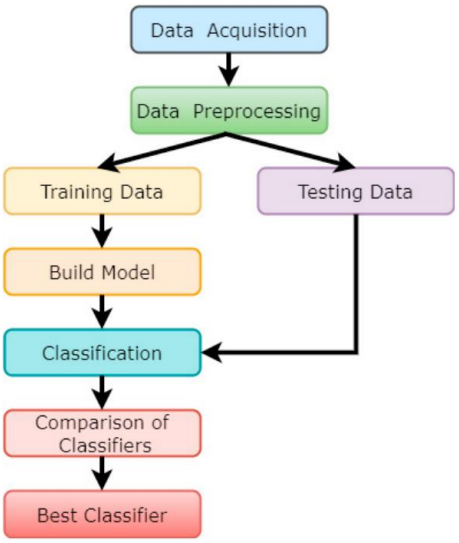


Fig.1. Proposed Flow Diagram

3.1. Dataset

For our experiments, we used the WDBC dataset. It is available on UCI repository [23].This dataset is used to determine whether a cancer is benign or malignant. A digitized image of an affine needle aspirate (FNA) of a mass in breast is used to extract features. It describes the properties of the nuclei in the image. This dataset contains 569 total records, 212 of which are malignant and 357 of which are benign. It has characteristics such as an ID number, a diagnosis, a radius, texture, a perimeter, an area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. There are a total of 13 features left. The mean, standard error, and worst or largest of these features are calculated for every image.

3.2. Model Building

The efficacy of medical treatment and the accuracy of the diagnosis are two factors that have a significant impact on a patient's chances of surviving cancer and avoiding a second bout with the disease. Data is randomly selected into training and testing and split in the ratio of 80:20 respectively. The effectiveness of the model was evaluated with the help of test data after it was trained with the help of training sets. The various features values will determine whether the person will be affected or not. The first thing that needs to be done is to collect the necessary data for pre-processing in order to enhance the data's quality. This can be done by using pre-processing Data is preprocessing performed by selecting only required attributes. Separate Diagnosis

ISSN: 2057-5688

column from remaining attributes, and label encoding is applied to categorical labels to convert them into numerical form. Standard scalar is applied that scales each feature to convert unit variance. After data preparation different machine models are used to train data. Once model trained its performance is evaluated on test data. In this different ML algorithms are used along with the ensemble classifier. The goal of ensemble learning is to produce a single superior predictive model by combining the results of multiple learning algorithms. It does this by combining several different kinds of supervised learners in order to boost the predictive ability of the model. Here we use Random forest bagging technique and Adaboost and XGBoost boosting techniques.

Following different Classification Algorithms and three ensemble learning techniques are applied on pre-processed breast cancer dataset to decide whether it is malignant or benign.

- Logistic Regression: It is an extension of linear regression model and used for classification problem. It provides the probabilities for two possible outcomes. In health care field this method is useful for prediction of likelihood of disease or illness.

- KNN: It uses all training data to classify new data point based on the similarity. To assign label to new data point it calculate its distance from different label point and finally it find nearest neighbour.

- SVM: This classification technique not required any prior distribution knowledge for classification. It uses hyper plane to classify data points

- Naïve Bayes: This techniques based on Bayes theorem. It performs prediction based on probability.

- Decision Tree: In this technique instances are classified using features value. For splitting it uses Gini Index, Information Gain. Leaf node indicates the label.

- Random Forest: It is ensemble technique. It creates multiple DT on different sample data generates majority votes and perform classification

- AdaBoost: Adaptive boosting is simple technique that uses decision tree. Multiple models are created; errors found in previous

ISSN: 2057-5688

models are corrected in next model. It assigns weights to incorrectly classified instances and subsequent model predicts that values correctly

- XGBoost: It stands for eXtreme Gradient Boosting and based on DT algorithm and reduce overfitting.

3.3 Implementation

All machine learning algorithm used in this paper are implemented on Google Colab that provide Jupyter Notebook environment using Scikit learn library available in python. Numpy, Pandas, Matplotlib libraries are also used.

4. Results and Discussions

On the Wisconsin Breast Cancer dataset, various models are used, and their performance is measured using Accuracy and AUC, precision, and recall. We compared the performance of various models by using AUC. We divided our results into two sections: standard ML algorithms and ensemble techniques. Confusion matrix includes actual and predicted labels as well as True Negative (TN), False Negative (FN) True Positive (TP) and False Positive (FP).Precision is defined as the number of positive class predictions that are actually positive class predictions, as shown in equation 1. The recall is the number of correct positive class predictions made out of all correct positive examples in the dataset and it is calculated as shown in equation 2. The F1 Score is derived from the weighted average of Precision and Recall; it is calculated as shown in equation 3.Accuracy is calculated by using confusion matrix as shown in equation 4; it tells how the tuple in training and testing data are correctly classified.\

Table 1. Classifier Performance

| Classification Technique | Parameter | Confusion matrix | Class | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|---|---|---|
| K-Nearest Neighbor | n_neighbors=5 | [67 0]<br>[ 5 42] | benign<br>malignant | 0.93<br>1.00 | 1.00<br>0.89 | 0.96<br>0.94 | 0.96 |
| Support Vector Machine | C=5.0<br>Kernel=linear | [63 4]<br>[ 2 45] | benign<br>malignant | 0.97<br>0.92 | 0.94<br>0.96 | 0.95<br>0.94 | 0.95 |
| Decision Tree | criterion = "gini" | [66 1]<br>[ 2 45] | benign<br>malignant | 0.97<br>0.98 | 0.99<br>0.96 | 0.98<br>0.97 | 0.97 |
| NaiveBayes | default<br>GaussianNB() | [61 6]<br>[ 5 42] | benign<br>malignant | 0.92<br>0.88 | 0.91<br>0.89 | 0.92<br>0.88 | 0.90 |
| Logistic Regression | solver='liblinear' | [65 2]<br>[ 2 45] | benign<br>malignant | 0.97<br>0.96 | 0.97<br>0.96 | 0.97<br>0.96 | 0.96 |

ISSN: 2057-5688

Table 2: Ensemble Techniques Results

| Classification Technique | Parameter | Confusion matrix | Class | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|---|---|---|
| Random Forest | n_estimators=30 | [65 2]<br>[ 2 45] | benign<br>malignant | 0.97<br>0.96 | 0.97<br>0.96 | 0.97<br>0.96 | 0.96 |
| Adaboost | n_estimators=40 | [64 3]<br>[ 1 46] | benign<br>malignant | 0.98<br>0.94 | 0.96<br>0.98 | 0.97<br>0.96 | 0.96 |
| XGBoost | objective="binary:logistic | [66 1]<br>[ 2 45] | benign<br>malignant | 0.97<br>0.98 | 0.99<br>0.96 | 0.98<br>0.97 | 0.97 |

Table 1 shows the performance of different classifier on test dataset. It shows the different classifier with their respective hyper parameters. It includes precision, recall, F1 score for each class and accuracy of the model. It indicates that decision tree classifier with highest accuracy 97% and NB with lowest accuracy 90%. Fig.2. shows the ROC curve for different ML classification techniques applied on breast cancer dataset. Logistic regression classifier has highest AUC score 0.993 and Decision tree has lowest AUC score 0.938.Similarly Table 2 shows performance of ensemble techniques: RF, Adaboost, XGBoost on same dataset and Fig.3. shows the ROC curve for the same. XGBoost has achieved maximum accuracy 97% and highest AUC 0.999.Table 3 and Fig .4. shows comparison with previous approaches used for breast cancer detection in terms of accuracy .Our ensemble model accuracy is slightly less than mentioned in [11] but Fig.5. shows the comparison of ROC curves between [11] and ours. Our XGBoost classifier achieved 97% accuracy and AUC-ROC 0.999.

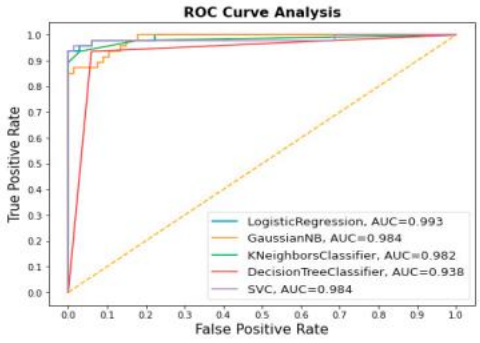


Fig.2. AUC-ROC Curve for ML Algorithm

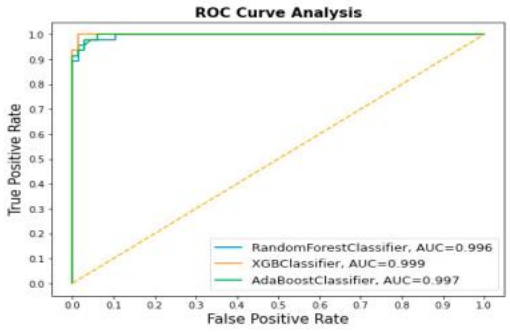


Fig.3. AUC-ROC Curve For Ensmeble Techniques

ISSN: 2057-5688

Table 3: Comparison for Various Approaches on breast

| Author | Technique | Accuracy |
|---|---|---|
| Naji et al.[11] | SVM | 97.2 |
| Amrane et al.[16] | MLP | 95.4 |
| Ara et al.[20] | RF | 96.5 |
| Chaurasia et al[.24] | J48 | 93.41 |
| Ours | XGBoost | 97 |

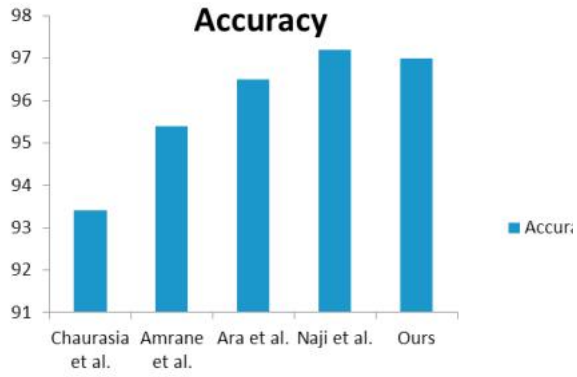


Fig.4. Accuracy comparison

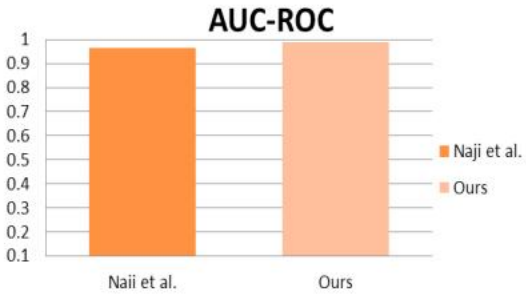


Fig.5. AUC-ROC Curve compariosn with previous approach

## 5. Conclusion

The passage highlights the significance of breast cancer research and the crucial role of technology, particularly machine learning (ML) algorithms, in reducing the mortality rate associated with breast cancer. Despite numerous classification methods developed for breast cancer data analysis, challenges remain, particularly in terms of accuracy. To address this, the study proposes a classification model using six ML techniques: Decision Tree (DT), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), and Logistic Regression (LR), as well as ensemble techniques. The study focuses on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset and evaluates these techniques. Among the standard ML algorithms, the Decision Tree classifier using the Gini index criterion achieved the highest accuracy at 97%, and the LR classifier attained the highest Area Under the Curve (AUC) value of 0.996. For ensemble techniques, XGBoost achieved the highest accuracy of 97% with an AUC of 0.99. The proposed system holds promise for aiding cancer specialists in cancer recognition. Additionally, the study identifies a future

ISSN: 2057-5688

research direction: performing hyperparameter tuning to further enhance the model's performance. This work emphasizes the ongoing efforts to harness machine learning for accurate and effective breast cancer classification, reflecting the continuous advancements in this critical area of research.

## References

[1] S Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 71(3), 209-249.

[2] Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., & Bray, F. (2021). Cancer statistics for the year 2020: An overview. International Journal of Cancer, 149(4), 778-789.

[3] Patil S, Kirange D, Nemade V. Predictive modelling of brain tumor detection using deep learning. Journal of Critical Reviews. 2020;7.

[4] Nemade, V., Pathak, S., Dubey, A. K., & Barhate, D. (2022).A Review and Computational Analysis of Breast Cancer Using Different Machine Learning Techniques,12(3)111-118.

[5] Chaturvedi P, Jhamb A, Vanani M, Nemade V. Prediction and classification of lung cancer using machine learning techniques. InIOP Conference Series: Materials Science and Engineering 2021 Mar 1 (Vol. 1099, No. 1, p. 012059). IOP Publishing.

[6] Abdelhafiz, D., Yang, C., Ammar, R., & Nabavi, S. (2019). Deep convolutional neural networks for mammography: advances, challenges and applications. BMC bioinformatics, 20(11), 1-20.

[7] Nemade, V., Pathak, S., & Dubey, A. K. (2022). A Systematic Literature Review of Breast Cancer Diagnosis Using Machine Intelligence Techniques. Archives of Computational Methods in Engineering, 1-30.

[8] Dora L, Agrawal S, Panda R, Abraham A.(2017) Optimal breast cancer classification using Gauss–Newton representation based algorithm. Expert Systems with Applications,85:134-45.

[9] Obaid OI, Mohammed MA, Ghani MK, Mostafa A, Taha F. (2018) Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer. International Journal of Engineering & Technology,7(4.36):160-6.

[10] Yellamma, P., Chowdary, C. S., Karunakar, G., Rao, B. S., & Ganesan, V. (2020). Breast Cancer Diagnosis Using MLP Back Propagation. International Journal, 8(9).

[11] Naji, M. A., El Filali, S., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021). Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. Procedia Computer Science, 191, 487-492.

ISSN: 2057-5688

[12] Kabiraj, S., Raihan, M., Alvi, N., Afrin, M., Akter, L., Sohagi, S. A., & Podder, E. (2020, July). Breast cancer risk prediction using XGBoost and random forest algorithm. In 2020 11th international conference on computing, communication and networking technologies (ICCCNT) (pp. 1-4). IEEE.

[13] Liu, P., Fu, B., Yang, S. X., Deng, L., Zhong, X., & Zheng, H. (2020). Optimizing survival analysis of XGBoost for ties to predict disease progression of breast cancer. IEEE Transactions on Biomedical Engineering, 68(1), 148-160.

[14] Nanglia, S., Ahmad, M., Khan, F. A., & Jhanjhi, N. Z. (2022). An enhanced Predictive heterogeneous ensemble model for breast cancer prediction. Biomedical Signal Processing and Control, 72, 103279.

[15] Islam, M., Haque, M., Iqbal, H., Hasan, M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: a comparative study using machine learning techniques. SN Computer Science, 1(5), 1-14 [16] Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018, April). Breast cancer classification using machine learning. In 2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT) (pp. 1-4). IEEE.

[17] Dhahri, H., Al Maghayreh, E., Mahmood, A., Elkilani, W., & Faisal Nagi, M. (2019). Automated breast cancer diagnosis based on machine learning algorithms. Journal of healthcare engineering.

[18] Kharya S, Soni S. Weighted naive bayes classifier: a predictive model for breast cancer detection. International Journal of Computer Applications. 2016 Jan;133(9):32-7

[19] Huang Q, Chen Y, Liu L, Tao D, Li X. On combining biclustering mining and AdaBoost for breast tumor classification. IEEE Transactions on Knowledge and Data Engineering. 2019 Jan 14;32(4):728-38. [20] Ara S, Das A, Dey A. Malignant and benign breast cancer classification using machine learning algorithms. In2021 International Conference on Artificial Intelligence (ICAI) 2021 Apr 5 (pp. 97-101). IEEE.