Prediction and Diagnosis of Breast Cancer using Machine Learning Algorithms

Abstract

Breast cancer is one of the most prevalent and fatal forms of cancer in India. It ranks the second most common cancer in rural areas and the most common in urban areas. According to a report by the International Agency for Research on Cancer, there were over 2.26 million new breast cancer cases and nearly 685,000 deaths from breast cancer globally. With a significant portion of India's population being young, the number of women diagnosed with breast cancer is expected to increase, reaching alarming levels due to a lack of awareness and delays in diagnosis. While breast cancer cannot be prevented, early detection and timely treatment can significantly improve survival rates. This study uses K-Nearest Neighbour (K-NN), Random Forest, Decision Trees (CART), Support Vector Machine (SVM), and Naïve Bayes to aid oncologists in identifying and diagnosing breast cancer, thereby assisting in treatment decision-making. We present a predictive model for the early detection of breast cancer and compare the results of the employed models for effective detection.

Keywords: Machine Learning, Breast Cancer, Classification, Prediction.

Introduction

"Breast cancer is the most prevalent form of cancer, posing a significant global health challenge, including in India. Characterized by the uncontrolled growth of cells leading to the formation of lumps in the breast, it is one of the treatable forms of cancer. However, if not detected early, it can become life-threatening as it may spread to other parts of the body. This study focuses on the early detection of breast cancer using machine learning techniques. Early diagnosis strategies aim to provide timely access to treatment, increasing the proportion of cases identified at an early stage, which allows for more effective treatment and reduces the risk of death. Although breast cancer rates are higher among women in developed regions, the incidence is rising globally"[20,21,22]. Thus, early detection is critical for improving outcomes.

The main contribution of this paper is to review the role of Machine Learning (ML) techniques in the early detection of breast cancer. Machine Learning, a branch of Artificial

Intelligence (AI), focuses on developing algorithms that improve through learning from data. These sophisticated algorithms create models using data samples to predict real-world situations or make decisions without explicit programming. The primary benefit of ML in medical diagnosis is increased efficiency, allowing medical staff more time to provide accurate treatment.

In medical diagnosis, ML not only aids in finding solutions but also in organizing data, which can be advantageous for doctors. Given the critical importance of early cancer detection for effective breast cancer treatment, this study employs various ML algorithms to predict whether breast cancer is benign or malignant based on dataset features. Benign cells are non-cancerous and do not spread, whereas malignant cells are cancerous and can metastasize.

One major issue is the lack of reliable early-stage diagnostic machines, which reduces the chances of survival. Early diagnosis is often curable with prompt intervention—"a stitch in time saves nine." The absence of prognostic models complicates doctors' efforts to create effective treatment plans that could extend patient survival.

Traditional diagnostic methods like ultrasound, mammograms, and biopsies are time-consuming, highlighting the need for a fully automated diagnostic technique. This paper addresses this need by applying ML techniques to improve accuracy and reduce diagnostic time. The study utilizes several ML algorithms, including Classification and Regression Trees (CART) for classification and prediction, Support Vector Machine (SVM) for data analysis in classification and regression, Naïve Bayes, and K-Nearest Neighbour (K-NN) algorithms.

Numerous machine learning algorithms are available for the prediction and diagnosis of breast cancer, including Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree, and K-Nearest Neighbors (KNN) etc. Researchers have conducted extensive studies on breast cancer using various datasets such as the SEER dataset, mammogram images, the Wisconsin Dataset, and datasets from different hospitals. By leveraging these datasets, researchers extract and select various features to complete their studies, leading to significant research contributions.

Nayak and Gope (2017) demonstrated “the use of various supervised machine learning algorithms for classifying breast cancer using 3D images, finding that SVM performed best overall”. Gayathri and Sumathi (2016) conducted “a comparative study on the Relevance Vector Machine (RVM), highlighting its low computational cost and 97% accuracy, making

it superior to other ML algorithms for diagnosing breast cancer with reduced variables". Asri et al. (2016) showed that "SVM excels in breast cancer prediction and diagnosis, achieving 97.13% accuracy with high precision and low error rate.Khoudfi and Bahaj(2018) compared ML algorithms and found SVM to be the best classifier with 97.9% accuracy, compared to K-NN, RF, and NB, using a multilayer perceptron with five layers and ten-fold cross-validation". Latchoumietand Parthiban (2017) achieved "a 98.4% classification value by optimizing the weighting of particle swarm optimization (WPSO) based on SSVM". Osman (2017)proposed"a solution for diagnosing Wisconsin breast cancer (WBCD) with 99.10% prediction accuracy using SVM combined with a clustering algorithm and an efficient probabilistic vector support machine".

Our research focuses on evaluating these machine learning algorithms and approaches to determine the most effective methodology for breast cancer prediction and diagnosis.

Data and Methodology

The study used secondary data set downloaded from UCI Machine Learning Repository. Features are derived from digitized images of Fine Needle Aspirates (FNA) of breast masses, characterizing the cell nuclei within the breast tissue. The dataset includes 569 patients, with 212 diagnosed with malignant tumours and 357 with benign tumours.It consists of 32 attributes (ID, diagnosis, 30 real valued input features).Some attribute information are:

1. ID number
2. Diagnosis (M= malignant, B= benign)

   (3-32) ten real-valued features are computed for each cell nucleus:

   i. Radius (mean of distances from centre to points on perimeters)
   ii. Perimeter
   iii. Area
   iv. Texture (standard deviation of grey-scale values)
   v. Compactness (perimeter^2/area-1.0)
   vi. Smoothness (local variation in radius lengths)
   vii. Concave points (number of concave portions of contour)
   viii. Concavity (severity pf concave portions of the contour)
   ix. Symmetry
   x. Fractal dimension ("coastline approximation"-1)

Our methodology begins with data acquisition followed by pre-processing, which involves four steps: data cleaning, attribute selection, target role setting, and feature extraction. The prepared data is then used to build machine learning algorithms to predict breast cancer for new measurements. To evaluate the performance of these algorithms, we test them on the data with known labels. This is typically achieved by splitting the labelled data into two parts using the train and test split method: 80% of the data is used to build the machine learning model (training set), and 20% is used to evaluate its performance (test set). After testing the models, we compare the results to select the algorithm with the highest accuracy and identify the most predictive algorithm for breast cancer detection.

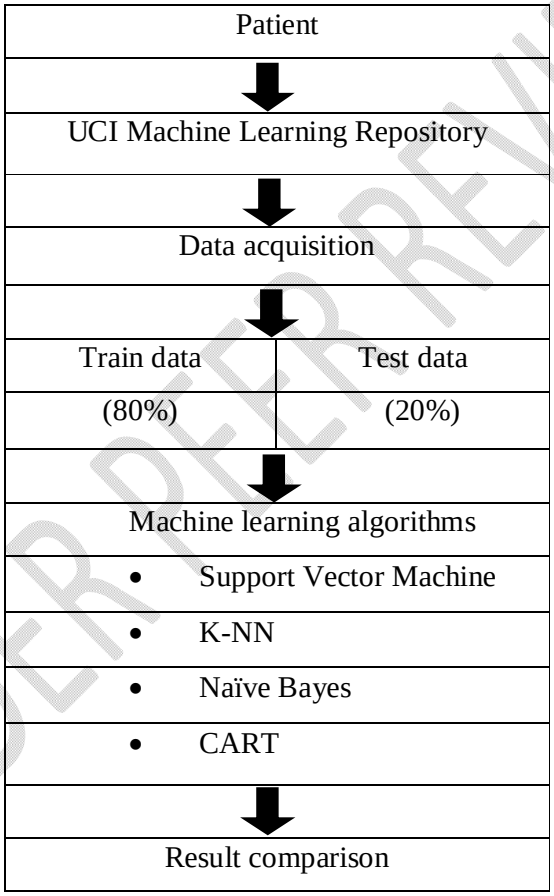

Chart 1:Process flowchart

Support Vector Machine (SVM)

SVM is one of the most popular supervised machine learning algorithms, used for both classification and regression. It is particularly powerful and sophisticated for predictive analysis. In this study, we have implemented SVM using two kernels: linear and Gaussian. For classifying linearly separable datasets, we prefer the linear kernel, while for non-linear

datasets, we use the Gaussian polynomial kernel. The primary goal of SVM is to determine a hyperplane that effectively divides the data into two distinct classes.

$$hyperplane\ eq(y) = w*X' + b$$

The above equation is dependent on weight vector (w), bias element (b) and support vector (X). Vectors that lie closest to the hyperplane are support vectors. Support vectors are responsible for determining the *hyperplane*.

K-Nearest Neighbour (K-NN)

The K-NN algorithm uses 'feature similarity' to predict the values of new data points, meaning new data is assigned a value based on how closely it matches points in the training set. K-NN is a non-parametric method used for both classification and regression. In both cases, it can be beneficial to weight the contributions of the neighbors so that closer neighbors contribute more to the prediction than more distant ones. For instance, a common weighting scheme involves assigning each neighbor a weight of 1/d, where d is the distance. Euclidean and Manhattan distance measures are typically used to assign weights and determine the neighbors. If we consider two points x_i and $y_i$in n dimensional space then:

$$\text{Euclidean}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

$$\text{Manhattan}(x,y) = \sum_{i=1}^{n} | x_i - y_i |$$

Naïve Bayes

Naïve Bayes is a classification technique based on Bayes' theorem, assuming independence between predictors. Simply put, a Naïve Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. This technique constructs classifiers that assign class labels to problem instances, represented as vectors of feature values, with class labels drawn from a finite set.Using Bayes' Theorem, the conditional probability can be expressed as:

$$P(c|x) = P(x|c)P(c)/P(x)$$

where (P(c|x)) is the posterior probability of class (c) given predictor (x), (P(c)) is the prior probability of class (c), (P(x)) is the prior probability of predictor (x), and (P(x|c)) is the likelihood of predictor (x) given class (c).We implemented the classifier using both the normal distribution and the kernel classifier.

Classification and Regression Tree (CART)

The CART algorithm is a classification algorithm that constructs decision trees based on Gini's impurity index. When the target variable is continuous, a classification tree is employed to determine the "class" most likely to encompass the target variable, while regression trees are utilized to predict the value of a continuous variable.

In a decision tree, nodes are divided into sub-nodes based on the threshold value of an attribute. The CART algorithm accomplishes this by seeking the best homogeneity for the sub-nodes using the Gini index criterion. Initially, the root node comprises the training set and is divided into two based on the optimal attribute and threshold value. This process iteratively continues, with subsets being further divided using the same logic, until either the last pure subset is reached in the tree or the maximum number of leaves allowed in the growing tree is achieved.

Results and Discussion

Based on above explanation of the proposed methodology, analysis of the numerical dataset by using four different machine learning algorithms results in namely SVM, K-NN, Decision Tree (CART), and Naïve Bayes yielded following results.

Model	accuracy value
CART	0.91
SVM	0.97
Naïve Bayes	0.92
K-NN	0.97

Table 1: Predicted accuracy values of estimated models (train data)

The table presents the accuracy values of different models when trained on a dataset. Accuracy is a measure of how often the model's predictions are correct. In this context, a higher accuracy value indicates that the model's predictions closely match the actual outcomes.

CART (Classification and Regression Trees) has an accuracy value of 0.91, suggesting that it correctly predicts outcomes about 91% of the time on the training data.

SVM (Support Vector Machine) has an accuracy value of 0.97, indicating that it is highly accurate in its predictions, with a correctness rate of 97% on the training data.

Naïve Bayes has an accuracy value of 0.92, indicating that it accurately predicts outcomes about 92% of the time on the training data.

K-NN (K-Nearest Neighbors) has an accuracy value of 0.97, suggesting that it is highly accurate, with a correctness rate of 97% on the training data. (Table 2)

In summary, these accuracy values provide insight into the performance of each model when trained on the dataset, with SVM achieving the highest accuracy, followed closely by K-NN and Naïve Bayes, while CART exhibits slightly lower accuracy but still performs well.

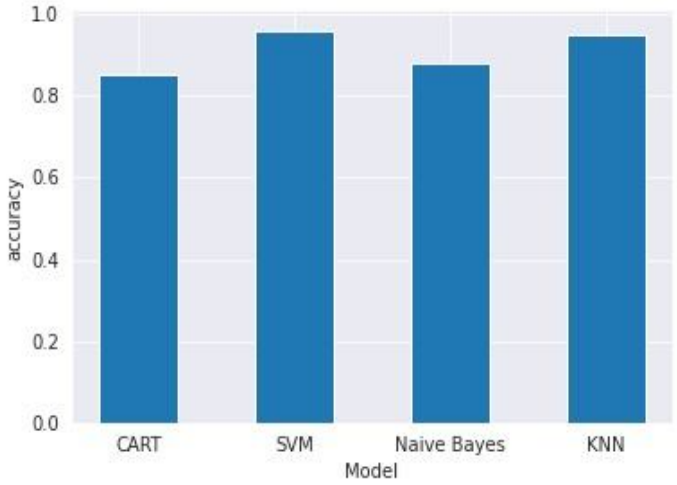

Figure 1: Accuracy values for trained data using different models

Model	accuracy value (test data)
CART	0.90
SVM	0.99
Naïve Bayes	0.93
K-NN	0.98

Table 2: Predicted accuracy values of estimated models (test data)

The table presents the accuracy values of different models when tested on a dataset. Accuracy is a measure of how often the model's predictions are correct. In this context, a higher accuracy value indicates that the model's predictions closely match the actual outcomes.

CART has an accuracy value of 0.90, suggesting that it correctly predicts outcomes about 90% of the time on the test data.

SVM has the highest accuracy value of 0.99, indicating that it is highly accurate in its predictions, with a correctness rate of 99% on the test data.

Naïve Bayes has an accuracy value of 0.93, indicating that it accurately predicts outcomes about 93% of the time on the test data.

K-NN has an accuracy value of 0.98, suggesting that it is highly accurate, with a correctness rate of 98% on the test data. (Table 3)

In conclusion, SVM appears to be the best model overall based on its performance on the test data. However, the choice of the best model may also depend on other factors such as the nature of the problem and the cost of errors.

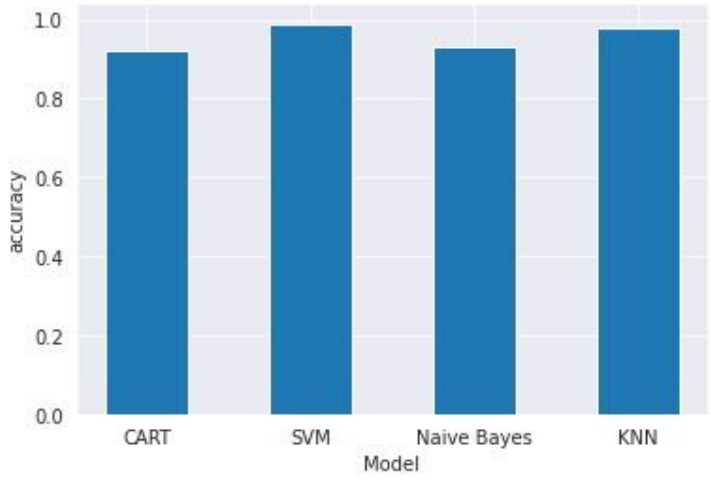

Figure 2: Accuracy values for test data using different models

		True Class	
		Positive	**Negative**
Predicted class	**Positive**	**74**	**1**
	Negative	**0**	**39**

Table 3: Confusion Matrix

In this confusion matrix, the rows represent the actual (true) classes of the data points, and the columns represent the classes predicted by the classifier. The diagonal elements (True Positive and True Negative) represent the number of data points that were correctly

classified. The off-diagonal elements (False Positive and False Negative) represent the number of data points that were misclassified. Further from table entries:

True Positive:74 data points that were correctly classified as positive.**False Negative:**01 data points that were actually positive but were predicted as negative by the classifier. These are also known as Type II errors.**False Positive:**39 data points that were predicted as positive by the classifier but were actually negative. These are also known as Type I errors.**True Negative:**0 data points that were correctly classified as negative.

Based on this confusion matrix, we can see that the classifier has a good performance overall, with a high number of True Positives (74) and True Negatives (0). However, there are also some misclassifications (1 False Negative and 39 False Positives).

Conclusion

Our study suggests that machine learning models trained through an end-to-end approach can achieve notably high levels of accuracy and possess the potential for adaptation across various mammography platforms. We investigated different machine learning techniques for breast cancer detection in this study. Through an analysis comparing CART, SVM, Naïve Bayes, and KNN, we aimed to discern their similarities and differences. Our findings revealed that SVM excels over the other methods employed in terms of accuracy, precision, and data utilization. Furthermore, our methodology can be applied to address additional challenges in medical imaging.

Disclaimer (Artificial intelligence)

Option 1:

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

Option 2:

Author(s) hereby declare that generative AI technologies such as Large Language Models, etc have been used during writing or editing of manuscripts. This explanation will include list the name, version, model, and source of the generative AI technology and as well as the all input prompts provided to a generative AI technology

Details of the AI usage are given below:

1.

2.

3.

References

1. Osman AH. An enhanced breast cancer diagnosis scheme based on two-step-SVM technique. Int. J. Adv. Comput. Sci. Appl. 2017 Apr 1;8(4):158-65.
2. Allugunti VR. Breast cancer detection based on thermographic images using machine learning and deep learning algorithms. Int J EngComput Sci. 2022 Jan 1;4(1):49–56.
3. Alzu'bi A, Najadat H, Doulat W, Al-Shari O, Zhou L. Predicting the recurrence of breast cancer using machine learning algorithms. Multimed Tools Appl. 2021 Apr;80(9):13787–800.
4. Arya KN, Pandian S, Joshi AK, Chaudhary N, Agarwal GG, Ahmed SS. Sensory deficits of the paretic and non- paretic upper limbs relate with the motor recovery of the poststroke subjects, Topics in Stroke Rehabilitation. 2023; DOI: 10.1080/10749357.2023.2253629
5. Gayathri BM, Sumathi CP. Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer. In2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) 2016 Dec 15 (pp. 1-5). IEEE.
6. Chauhan A, Kharpate H, Narekar Y, Gulhane S, Virulkar T, Hedau Y. Breast Cancer Detection and Prediction using Machine Learning. In2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) 2021 Sep 2 (pp. 1135-1143). IEEE.
7. Ghosh P, Azam S, HasibKMd, Karim A, Jonkman M, Anwar A. A Performance Based Study on Deep Learning Algorithms in the Effective Prediction of Breast

Cancer. In: 2021 International Joint Conference on Neural Networks (IJCNN) [Internet]. Shenzhen, China: IEEE; 2021. p. 1–8. Available from: https://ieeexplore.ieee.org/document/9534293/

8. Asri H, Mousannif H, Al Moatassime H, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science. 2016 Jan 1;83:1064-9.
9. Kajala A, Jain VK. Diagnosis of Breast Cancer using Machine Learning Algorithms-A Review. In: 2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3) [Internet]. Lakshmangarh, India: IEEE; 2020 [cited 2023 Aug 4]. p. 1–5. Available from: https://ieeexplore.ieee.org/document/9117320/
10. Latchoumi TP, Parthiban L. Abnormality detection using weighed particle swarm optimization and smooth support vector machine. Biomedical Research. 2017 Jan 1;28(11):4749-51.
11. Mandal JK, Bhattacharya D, editors. Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018 [Internet]. Singapore: Springer Singapore; 2020 [cited 2023 Aug 4]. (Advances in Intelligent Systems and Computing; vol. 937). Available from: http://link.springer.com/10.1007/978-981-13-7403-6
12. Naji MA, Filali SE, Aarika K, Benlahmar EH, Abdelouhahid RA, Debauche O. Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. Procedia Computer Science. 2021;191:487–92.
13. Rabiei R. Prediction of Breast Cancer using Machine Learning Approaches. J Biomed Phys Eng [Internet]. 2022 Jul 1;12(3). Available from: https://jbpe.sums.ac.ir/article_48331.html
14. Nayak S, Gope D. Comparison of supervised learning algorithms for RF-based breast cancer detection. In2017 Computing and Electromagnetics International Workshop (CEM) 2017 Jun 21 (pp. 13-14). IEEE.
15. Sadhukhan S, Upadhyay N, Chakraborty P. Breast cancer diagnosis using image processing and machine learning. InEmerging Technology in Modelling and Graphics 2020 (pp. 113-127). Springer, Singapore.
16. Sharma A, Mishra PK. Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis. International Journal of Information Technology. 2022 Jun;14(4):1949-60.

17. Surendhar SP, Vasuki RJ. Breast cancers detection using deep learning algorithm. Materials Today: Proceedings. 2021.
18. Tahmooresi M, Afshar A, Rad BB, Nowshath KB, Bamiah MA. Early detection of breast cancer using machine learning techniques. Journal of Telecommunication, Electronic and Computer Enginesssering (JTEC). 2018 Sep 26;10(3-2):21-7.
19. Khourdifi Y, Bahaj M. Applying best machine learning algorithms for breast cancer prediction and classification. In2018 International conference on electronics, control, optimization and computer science (ICECOCS) 2018 Dec 5 (pp. 1-5). IEEE.
20. Kumar, U., Singh, A., Chandra, K., Atreya, K., Singh, R., & Kumar, M. (2022). Molecular Classification of Breast Carcinoma Based on the Prognostic Marker: A Clinico-pathological Correlation. Journal of Advances in Medicine and Medical Research, 34(23), 237–247. https://doi.org/10.9734/jammr/2022/v34i234859
21. Lathishna, A. G. and Kamal, V. S. (2021) "Breast Cancer Screening Awareness, Practice and Knowledge among Women Attending Out Patient in Tertiary Care Centre", Journal of Pharmaceutical Research International, 33(54A), pp. 146–150. doi: 10.9734/jpri/2021/v33i54A33732
22. Islam MM, Haque MR, Iqbal H, Hasan MM, Hasan M, Kabir MN. Breast cancer prediction: a comparative study using machine learning techniques. SN Computer Science. 2020 Sep;1:1-4.