Original Research Article

Early Breast Cancer Prediction Using Machine Learning Algorithm

Abstract

Breast cancer is one of the most prevalent and fatal form of cancer in India. In rural India it is the second most common cancer and first in urban. According to a report published by International Agency for Research on Cancer, there were more than 2.26 million new cases of breast cancer and almost 685,000 deaths from breast cancer worldwide. With youth being major section of population in India the number of women getting diagnosed with breast cancer is only going to increase and it will reach to daunting proportions. This is because of lack of awareness and delay in diagnosis. It is not possible to prevent breast cancer but we can increase the chances of survival by early detection and getting treatment at right time. This paper study usesK-NearestNeighbour (K-NN), Random Forest, Decision Trees (CART), Support VectorMachine (SVM), and NaïveBayes for numerical dataset which helps oncologist in identifying anddiagnosing the breast cancer and then helps in decision making in treatment method for the same purpose. In this paper we present a predictive model for early detection of breast cancer. Finally,we compare the results of used models for the detection of breast cancer.

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

Introduction

Breast cancer is the most prevalent form of cancer. This disease has become a major problem across the world including India. Breast cancer is characterized by uncontrolled growth of cells, which results in the formation of lumps within the breast. It is one of the treatable forms of cancers. If not detected early, it can be a life -threatening disease as it can also spread to other parts of the body. This paper study focuses on early detection of breast cancer using machine learning techniques. Strategies for early diagnosis focus on providing timely access to cancer treatment.The goal is to increase the proportions of breast cancer identified cases at an early stage, allowing for more effective treatment to be used as early as possible and this will reduce the risk of death from this disease. While breast cancer rates are higher among

women in more developed regions, rates are increasing in nearly every region globally. To improve this condition, early detection of breast cancer is critical.

The main contribution of this paper is to review the role of Machine Learning (ML) techniques in early detection of the breast cancer. Machine Learning (ML) can relate to an area of research based on the use of Artificial Intelligence (AI). It is devoted to algorithms that use automated enhancement by acquiring knowledge, i.e., through leaning. These complicated algorithms create a model by using samples of data. They aim to predict real situations or make decisions without being earlier programmed to do so. The most important benefit of machine learning in medical diagnosis is increased efficiency. While the AI becomes smarter through machine learning, medical staff has more time to focus on their work of providing correct treatment at right time. In field of medical diagnosis, machine learning helps not only in finding solutions but also in organizing data. Well structured pieces of information can an advantage for any doctor.

Since early detection of cancer plays crucial role in treatment of breast cancer, in this paper study we use various machine learning algorithms to predict the type of breast cancer as benign or malignant, based on feature provided by the data set.

Benign cells are neither cancerous and nor capable of spreading but malignant cells are cancerous and can spread to other parts of body. The main problem in this disease is that there is no proper machine which can diagnose it in early stage reducing the chances of death. Early diagnosis of any disease is often curable with a touch of human effort. A stitch in time saves nine. The lack of prognosis models results in difficulty for doctors to prepare a treatment plan that may prolong patient survival time.

The early methods of diagnosis which are available are ultrasound, mammogram, biopsy, and all these are very time consuming. There was a need for a fully computerized and automatic diagnosis technique, and we have used machine learning technique for this, in this paper. This methodology includes algorithms that detect it more accurately and take less time. We have used various machine learning algorithms namely Classification and Regression Tress (CART) which are used for classification and prediction, Support Vector Machine (SVM) which analyses data for classification and regression, Naïve Bayes, K-nearest Neighbour (K-NN) algorithms.

Data and Methodology

WDBC (Wisconsin Diagnostic Breast Cancer) repository provided the data set. Features are computed from a digitized images of a Fine Needle Aspirate (FNA) of breast mass. They describe the characteristics of cell nuclei present in the breast mass.

The dataset consists of 569 patients, 212 have an outcome of Malignancy and 357 are Benign. We have used dataset from UCI Machine Learning repository (WDBC). WDBC consists of 32 attributes (ID, diagnosis, 30 real valued input features).

Some attribute information (WDBC):

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)

Data pre-processing

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real world data is often incomplete, inconsistent, lacking in certain behaviours or trends, and likely to contain errors.

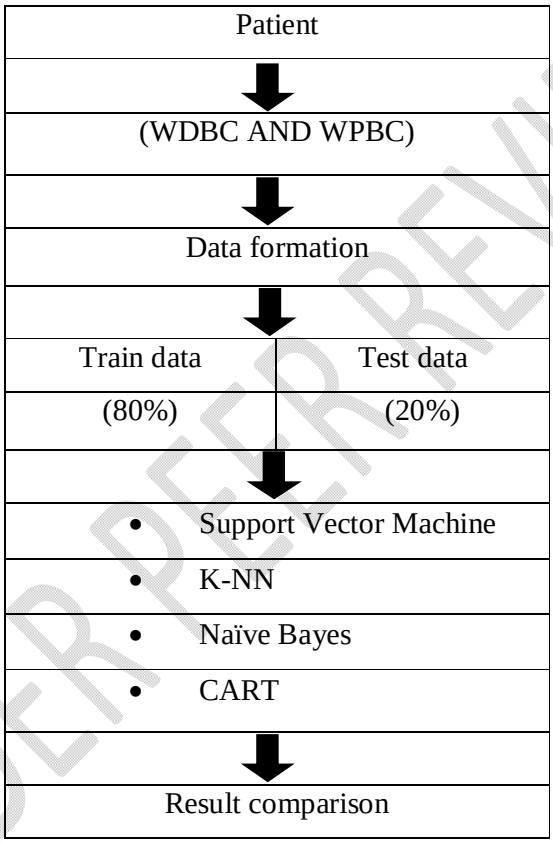
Training and testing phase

Training set is the actual data set from which a model trains i.e., the model sees and learns from this data to predict the outcome to make the right decisions. Most of the training data is pre-processed and organized to provide proper performance of the model.

The testing data set is independent of the training set but has a somewhat similar type of probability distribution of classes and is used as a bench mark to evaluate the model, it is used only after the training of the model is complete. Testing set is usually a properly organized dataset having all kinds of data for scenarios that the model would probably be facing when used in a real world.

We have used 80% of our data for training and 20 % for testing.

Table 1. Testing data set



Support Vector Machine (SVM)

SVM is one of the most popular supervised machine learning algorithms, which is used for classification and regression. It is very strong and sophisticated ML algorithm especially when it comes to predictive analysis. In this paper study, we have implemented SVM using two kernels: *linear* and *gaussian*. When we have to classify separable data set, we prefer linear kernel and for non-linear classification of dataset we prefer for kernel selection such as Gaussian polynomial.The focus of SVM is in determining the *hyperplane* such that it divides the region into two 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)

K-NN algorithm uses 'feature similarity' to predict the values of new datapoints which further means that the new data will be assigned a value based on how clearly it matches the points in the training set. K-NN algorithm is a non-parametric method is used for classification and regression. Both for classification and regression it can be useful to weight the contributions of the neighbours, so that the nearer neighbours contribute more to the average than the more distant ones. For example, a common weighting scheme consist in giving each neighbour a weight of 1/d, where d is the considered Euclidean and Manhattan distance measures to assign weight and determine the neighbours.

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 an assumption of independence between predictors which is known a Byes' theorem. In simple terms, a Naïve Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Naïve Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. Using Bayes' Theorem, conditional probability can be decomposed as:

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

Where, P(c|x) is posterior probability according to the predictor (x) for the class (c). P(c) is the prior probability of class, P(x) is prior probability of the predictor, and P(x|c) is the probability of the predictor for the particular class (c).

Hence, we implemented the classifier by considering *the normal distribution* and *kernel classifier*.

Classification and Regression Tree (CART)

The CART algorithm is a type of classification algorithm that is required to build a decision tree on basis of Gini's impurity index. When the target variable is continuous, the classification tree is used to find the "class" into which the target variable is most likely to fell, whereas Regression trees are used to forecast the value of a continuous variable.

In decision tree, the nodes are split into sub nodes based on a threshold value of an attribute. The CART algorithm does that by searching for the best homogeneity for the sub nodes, with the help of the Gini index criterion.

The root node is taken as the training set and is split into two by considering the best attribute and threshold value. Further, the subsets we also split using the same logic. This continuous till the last pure sub-set is found in the tree or the maximum number of leaves possible in that growing tree.

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 2: 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.

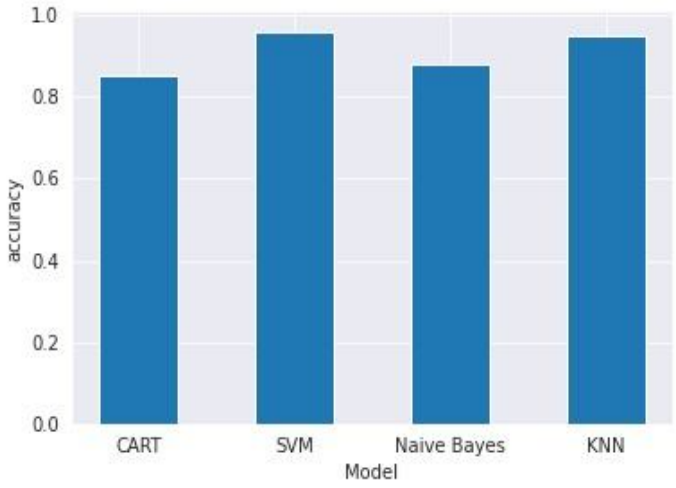


Figure 1: Accuracy values for trained data using different models

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

Table 3: 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.

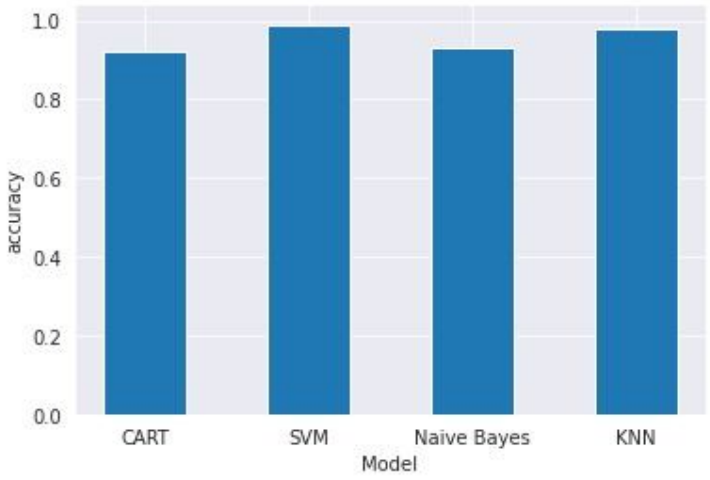


Figure 2: Accuracy values for test data using different models

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

Table 4: Confusion Matrix

The confusion matrix is commonly used to evaluate the performance of a classification model.

True Positive (TP): The model correctly predicted 74 instances as positive.

False Negative (FN): The model incorrectly predicted 1 instance as negative when it was actually positive.

False Positive (FP): The model incorrectly predicted 0 instances as positive when they were actually negative.

True Negative (TN): The model correctly predicted 39 instances as negative.

From this confusion matrix, we can calculate various performance metrics such as:

Accuracy= 0.986, or 98.6%,

Precision = 1.0, or 100%,

Recall (Sensitivity) = 0.987, or 98.7%,

Specificity = 1.0, or 100%.

Overall, the SVM model demonstrated high accuracy, precision, recall, and specificity, indicating strong performance in correctly identifying both positive and negative instances.

Conclusion

Our study indicates that machine learning methods trained via an end-to-end approach can achieve remarkably high levels of accuracy and are potentially adaptable to various mammography platforms. In this study, we explored different machine learning techniques for detecting breast cancer. We conducted an analysis comparing CART, SVM, Naïve Bayes, and KNN to understand their similarities and differences. It was found that SVM outperforms other used approaches in terms of accuracy, precision, and data utilization. Furthermore, our methodology can be applied to address additional challenges in medical imaging.

References

1. Allugunti VR. Breast cancer detection based on thermographic images using machine learning and deep learning algorithms. Int J Eng Comput Sci. 2022 Jan 1;4(1):49–56.
2. 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.
3. 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
4. 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.
5. Ghosh P, Azam S, Hasib KMd, 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/
6. 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/
7. 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
8. 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.
9. 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

10. 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.
11. 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.
12. Surendhar SP, Vasuki RJ. Breast cancers detection using deep learning algorithm. Materials Today: Proceedings. 2021.
13. 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.