Journal of Engineering Research and Reports

13(2): 48-54, 2020; Article no.JERR.58329
ISSN: 2582-2926

Early Breast Cancer Prediction using Artificial Intelligence Methods

Shawni Dutta[1] and Samir Kumar Bandyopadhyay[2*]

[1]Department of Computer Science, The Bhawanipur Education Society College, Kolkata, India.
[2]The Bhawanipur Education Society College, Kolkata, India.

Authors' contributions

This work was carried out in collaboration between both authors. Author SD designed the proposed method, coding and statistical work. Author SKB initiates the work and check the manuscript written by author SD. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JERR/2020/v13i217105
Editor(s):
(1) Dr. Tian- Quan Yun, South China University of Technology, China.
Reviewers:
(1) P. P. G. Dinesh Asanka, University of Kelaniya, Sri Lanka.
(2) Ayodele Abiola Periola, Bells University of Technology, Nigeria.
Complete Peer review History: http://www.sdiarticle4.com/review-history/58329

Original Research Article

Received 02 April 2020
Accepted 08 June 2020
Published 09 June 2020

ABSTRACT

In India, the death toll due to breast cancer is increasing at a rapid pace. Only early detection and diagnosis is the way of control but it is a major challenge in India due to lack of awareness and lethargy of Indian womentowards health care and regular check-up. But the major obstacle in India is expensive health care system and unavailability of proper infrastructure, especially in breast cancer treatment. This paper aims in obtaining an automated tool that will exploit patient's health records and predict the tendency of being affected in breast cancer. Gradient Boost classifier is used as an automated tool that predicts the chance of being affected in breast cancer disease. Early detection of this disease will assist health care systems to provide counter measures in order to save patients' life. The proposed model is evaluated against other peer classifiers such as Support Vector Machine (SVM) and K-Nearest Neighbour (K-NN), Naïve bayes classifier, Adaboost classifier, Decision Tree (DT) classifier, and Random Forest (RF) Classifier. The proposed method achieves encouraging result with an accuracy of 97.34%, F1-Score of 0.97 Cohen-Kappa Score of 0.94 and MSE of 0.0266. The Gradient Boost algorithm attains the lowest error rate along with highest efficiency which might be the best choice of algorithm for this problem and prediction of disease.

**Corresponding author: Email: 1954samir@gmail.com;*

Keywords: Breast cancer disease; gradient boosting; automated tool.

1. INTRODUCTION

Breast cancer develops from cells lining the milk ducts and slowly grows into a lump or a tumor. Breast cancer may be invasive or non-invasive. Invasive cancer spreads from the milk duct or lobule to other tissues in the breast, whereas, non-invasive ones lack the ability to invade other breast tissues. Non-invasive breast cancer is called "*in situ*" and may remain inactive for entire lifetime [1].

Data mining and knowledge discovery approaches are applied in enormous amount of data that automatically finds out patterns and relationship among of data. While diagnosing breast cancer diseases, Data mining and knowledge discovery approaches are explored [1]. The purpose of using these approaches is to generate new information based currently existing records. In order to identify hidden relationship among interfering factors that causes breast cancer, data mining and knowledge discovery approaches are utilized. The proposed system automatically captures previous health records of patient and detects whether the patient can be affected by breast cancer disease or not. Early prediction of this disease is required since cancer is often known as silent killer that develops without any symptoms. This paper applies Machine learning (ML) which is a supervised learning algorithm to identify patients with cancer disease severity. Given a set of messages, ML methods are capable of obtaining information and later use the acquired information to classify unknown new messages. Early cancer disease may be predicted by utilizing supervised machine learning approaches those takes patient's record as input. The predictive models can act as a tool to analyze the information of patients about their past health history records and predict their chances of having in breast cancer. This prediction will in turn help the doctors to take informed decisions and prescribe medicines and surgeries accordingly.

This paper attempts to utilize Gradient Boosting algorithm [2] to be applied on Breast Cancer Wisconsin (Diagnostic) Data Set and obtain prediction results. However, other classification algorithms such as K-Nearest Neighbour (K-NN) [3], Support Vector Machine (SVM) [4], Naïve Bayes Classifer (NB) [5], Decision Tree (DT) [6] classifier and also with ensemble classifiers such as Adaboost [7], Random Forest classifiers (RF) [8] are also implemented in this paper those are used as baseline for comparing with Gradient Boosting algorithm.

2. RELATED WORK

Three algorithms like Decision Tree (C4.5), Artifical Neural Networks (ANN) , and Support Vector Machine (SVM) are implemented in [9] in order to find classification accuracy in breast cancer dataset. Comparative study analysis show that SVM produces higher accuracy in classification.An extensive study was carried out in [10] by varying the values of k for k- Nearest Neighbor classification technique in order to enhance classification accuracy. Experiments were implemented on breast cancer dataset for early disease detection[10].

Delen et al.investigatedthe use ofartificial neural networks, decision trees and logistic regression to develop prediction models for breast cancer survival [11]. 10-fold cross-validation methods are explained to measure the unbiased estimate of the three prediction models for performance comparison purposes. The results indicated that the decision tree (C5) turns out to be the best predictor with 93.6% accuracy [11]. An investigation is made in using Naïve Bayes algorithmto test the classification accuracy of breast cancer dataset with respect to specificity, sensitivity and mean accuracy [12]. The results indicate that the Naive Bayes classifier provides equivalent performance as compared to other machine learning algorithms with low computational effort and high speed.

Two models namely Logistic Regression and ANN was implemented [13]. They were used to compare prediction accuracy results for detecting breast cancer via mammography. Comparative study concludes that logistic regression provides superior results in terms of prediction [13]. A diagnosticsystem is proposed for detecting breast cancer by implementing RepTree, RBF Network and Simple Logistic. In test stage, 10-fold cross validation method was applied for evaluating the proposed system performances. The correct classification rate of proposed system is attains 74.5% of efficiency [1].

From the existing carried out works specified in this breast cancer prediction domain, it is necessary to consider efficiency of these predictive models. In this paper, the main objective is to detect breast cancer patients with sufficiently enhanced efficiency. Considering related factors that may cause breast cancer with increased efficiency as well as lower prediction error rates is the main focused area of this research.

3. METHODOLOGY

The proposed methodology aims todetect patients with probable breast cancer tendency problem. The exact process of prediction using Gradient Boosting ensemble methodalong with other classifiers are illustrated through a series of steps as follows-

3.1 Data Collection and Preprocessing

In this framework, Breast Cancer Wisconsin (Diagnostic) Data Set from UC Irvine (UCI) machine learning repository [14] is utilized for predicting cardiac trouble tendency of a patient. The dataset can be formulated as collection of attributes that include several criteria for detecting heart disease tendency such as radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values),perimeter, area, smoothness (local variation in radius lengths),compactness (perimeter^2 / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, fractal dimension ("coastline approximation" - 1), diagnosis. However, the attribute ‘diagnosis’ is utilized as output class of the prediction which contains the class either benign or malignant. Fig.1 shows overall understanding of the dataset. Distribution of all attributes present in the dataset is illustrated in Fig. 1. For obtaining a balanced dataset, preprocessing techniques such as missing value handling, scaling some attributes are performed. Existing ‘nan’ values are also handled for this data. The attribute ‘id’ and ‘Unnamed :32’ will not contributing for prediction purpose which instantiate for elimination. An encoding process is applied on this pre-processed data to transform non-numeric data into numeric data. Performing these techniques will yield a transformed dataset that can be fitted to classifier. Next, feature scaling of relevant attributes are performed which enhances the efficiency while fitting to a classifier. The transformed dataset is partitioned into training set and testing dataset which is obtained by partitioning the transformed dataset with the ratio of 7:3.

3.2 Classification Method and Implementation

A classifier model maps input variable to target classes after learning from training data. The objective of using classifier is to predict whether a patient has malignant breast cancer tendency or not. A brief description of all classifiers used in this paper is provided as follows-

1. **Support Vector Machines -** *Support Vector machine* (SVM) [4] belongs to the category of linear as well as non-linear classifiers. It identifies different classes by separating samples with the help of decision boundary known as hyperplane. Both linear as well as non-linear data can be classified with the help of SVMs. It is also known as Maximum margin classifier since it can minimize the empirical classification error and maximize the geometric margin simultaneously.SVM is often advantageous in handling classification tasks with execellent generalization performance.The exhibited generalization ability by SVM is controlled by two different factors, that is the training error and the capacity of the learning machine measured. By changing the features in the classifiers, the training error rate can be controlled.
2. **Naïve Bayes Classifier-** The Naive Bayes classifier [5] is a supervised classification tool that exemplifies the concept of Bayes Theorem [15] of Conditional Probability. The decision made by this classifier is quite effective in practice even if its probability estimates are inaccurate. When features are independent or features are completely functionally independent are the two scenarios where this classifier provides very promising result. The accuracy of this classifier is not related to feature dependencies rather than it is the amount of information loss of the class due to the independence assumption is needed to predict the accuracy [5].
3. **Decision Tree Classifier-** A Decision Tree (DT) [6] is a classifier that exemplifies the use of tree-like structure. It gains knowledge on classification. The decision node or non-leaf node indicates certain test. The outcomes of these tests are signified either of the branches of that decision node. Each target class is denoted as a leaf node of DT. Classification result is obtained from this classifier by starting from the beginning of the corresponding nodes of the tree is

traversed through the tree until a leaf node is reached.

4. **K-nearest neighbor classifier-** K-Nearest Neighbour Classifiers (K-NN) [3] are often known as lazy learners. The classifier proceeds by identifying objects based on closest proximity of training examples in the feature space. This classifier considers k number of objects as the nearest object while determining the class. The main challenge of this classification technique relies on picking the appropriate value of k.
5. **Ensemble based classifier-** Ensemble approach facilitates several machine learning algorithms to work in harmony to attain higher accuracy of the entire system.

a. **Random Forest Classifier-** Random forest [8] exploits the concept of ensemble learning approach and applies regression technique for classification based problems. This classifier is a combination several tree-like classifiers which are applied on various sub-samples of the dataset and each tree cast its vote to the most appropriate class for the input.

b. **Adaboost Classifier-** Boosting [7] is an efficient technique that is applied on combination of several unstable learners in order to improve accuracy of classification. Boosting technique applies classification algorithm to the reweighted versions of the training data and chooses the weighted majority vote of the sequence of classifiers. AdaBoost [7] is a good example of boosting technique that produces improved output even when the performance of the weak learners is inadequate.

c. **Gradient Boost Classifier-** Gradient boosting algorithm [2] is another boosting technique based classifier that exemplifies the use of decision tree. It also minimizes the prediction loss. It checks models which decreases the loss function obtained from trained samples. From these calculations the errors are measured and analysed for optimal prediction of results. Loss function calculates the range of detected rate which compares with desired target. Onward stepwise process is most popular method for updating different with various attributes. The accuracy is optimized by reducing loss function and adding base learners at all stages.

Success in Machine learning methods may not always give accurate results. It depends on the dataset used for implementing the methods. All machine learning methods are not applicable to all situations. There are some limitations. It depends on the kinds of problems to recognize specific applications. The problem is advisable to solve different machine learning methods for any given set of data. It also requires comparing results using different Machine Learning methods to obtain as far as accurate prediction. The key point of the proposed method is compare results with the existing methods for obtaining high accuracy rate.

Implementation- The above specified classifiers are implemented by considering and adjusting appropriate hyper-parameters for obtaining the maximized performance. The SVM classifier utilizes ‘rbf’ kernel and regularization parameter C=1. The K-NN classifier gives a promising result for the value k=5 considering all the evaluating metric. For naïve bayes classifier, multinomial naïve bayes classifier is employed. The decision tree classifier implemented in this paper uses Gini index while choosing objects from dataset. The nodes of the decision tree are expanded until all leaves are pure or until all leaves contain less than minimum number of samples. In this case, minimum number of samples is assigned a value as 2. On the other hand, ensemble classifiers, such as, AdaBoost, Random Forest and Gradient Boost classifiers are built based on 500 numbers of estimators on which the boosting is terminated.

After implementing the above specified classifiers training dataset is fitted into the classifier and later prediction results are obtained. A general structure of classifier model is depicted in Fig. 2.

Later the results are evaluated with the actual observations with respect to predefined metrics which as discussed as following section.

3.3 Performance Measure Metrics

While evaluating performance of a model, performance measure metrics are used. Following are the metrics those are required to justify the performance of the given model.

Accuracy [16] is a metric that ascertains the ratio of true predictions over the total number of instances considered. However, evaluating a model in terms of accuracy may not be enough since it does not consider wrong predicted cases. For addressing the above mentioned problem, we yield two more metrics known as, Recall and Precision. *Precision* [16] identifies the ratio of correct positive results over the number of positive results predicted by the classifier. *Recall* [16] represents the number of correct positive results divided by the number of all relevant samples. *F1-Score* or *F-measure* [16] is a parameter that is calculated as the harmonic mean of precision and recall. *Cohen-Kappa Score* [17] is also taken into consideration as an evaluating metric in this paper. This metric is a statistical measure that finds out inter-rate agreement for qualitative items for classification problem.

Mean Squared Error (MSE) [16] is another evaluating metric which is used for measuring absolute differences between the prediction and actual observation of the test samples.

The Matthews correlation coefficient (MCC) [18] is another evaluating metric that is defined as a measure of the quality of binary (two-class) classifications. It considers true and false positives and negatives and is generally regarded as a balanced measure which is useful even if the classes are of very different sizes.

A model showing higher values of accuracy, MCC, F1-Score, Cohen-Kappa Score and lower MSE value indicate a better performing model.

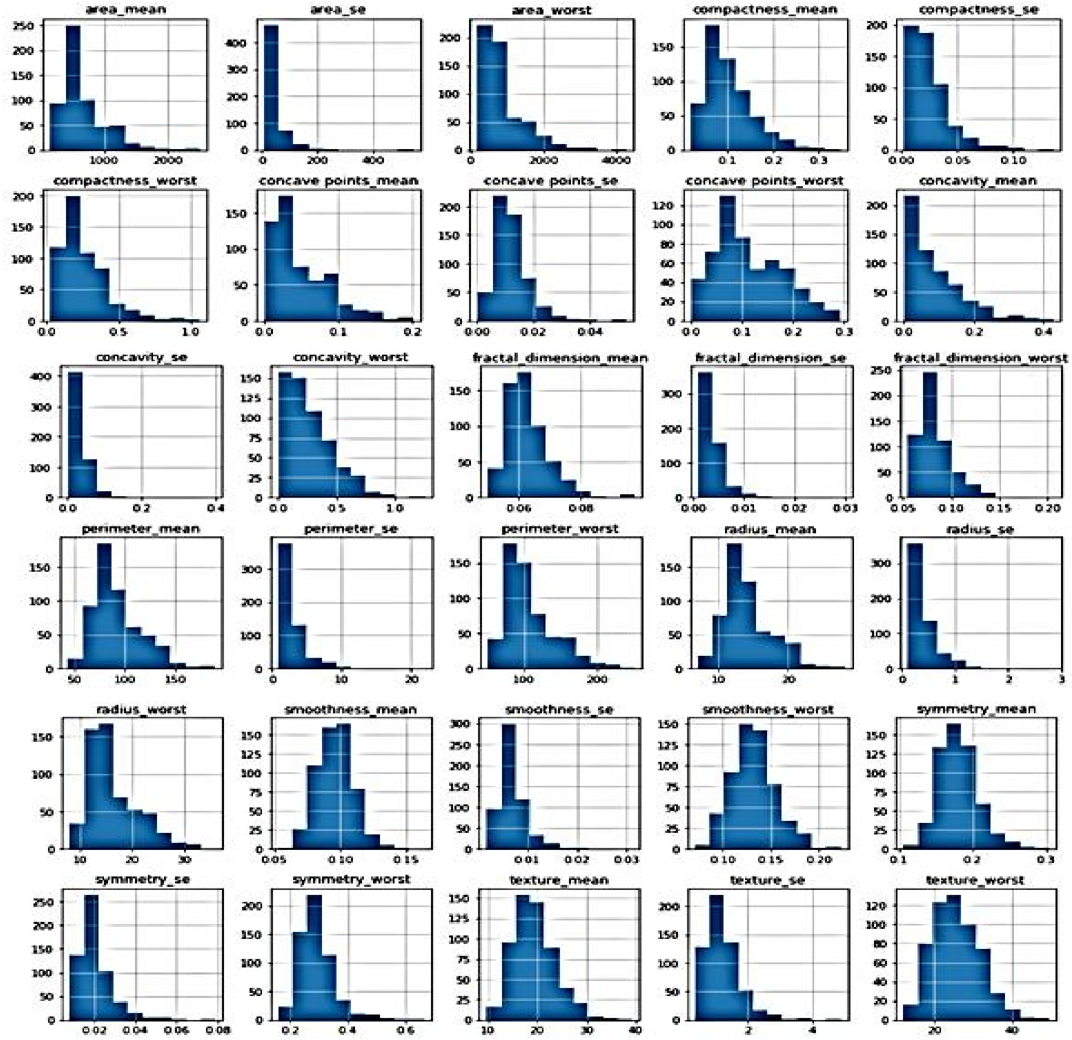

Fig. 1. Visual representation of collected dataset

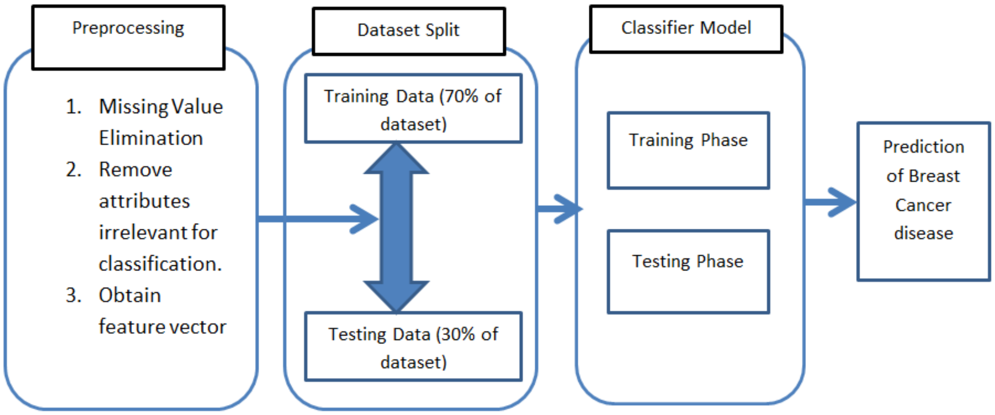
Preprocessing
1. Missing Value Elimination
2. Remove attributes irrelevant for classification.
3. Obtain feature vector
Dataset Split
Training Data (70% of dataset)
Testing Data (30% of dataset)
Classifier Model
Training Phase
Testing Phase
Prediction of Breast Cancer disease

Fig. 2. General structure of classifier model

Table 1. Comparative analysis of all specified classifier models

Performance measure metrics	**SVM**	**K-NN**	**Naïve bayes classifier**	**Decision tree classifier**	**AdaBoost**	**Random forest**	**Gradient boost**
Accuracy	96.81%	95.74%	92.55%	93.09%	96.81%	96.28%	97.34%
MCC	0.9264	0.90	0.825	0.840	0.925	0.913	0.938
F1-Score	0.97	0.96	0.93	0.93	0.97	0.96	0.97
Cohen-Kappa Score	0.92	0.9	0.82	0.84	0.93	0.91	0.94
MSE	0.0319	0.0426	0.0745	0.0691	0.0319	0.04	0.0266

4. RESULTS AND DISCUSSION

The Gradient Boosting method is implemented and evaluated in terms of the above mentioned metrics. This model is later compared with other benchmark classifiers known as other classifier models such as SVM, K-Nearest Neighbor (K-NN), NB Classifier, DT Classifier, Adaboost Classifier, RF classifier. The comparative study is shown in Table 1. From the comparative study it is clear that the proposed model indicates much promising result over other classifiers in terms of Accuracy, F1-Score, Cohen-kappa score and MSE.

5. CONCLUSIONS

Breast cancer affecting the women is known to cause high mortality unless detected in time. Detection requires a simple procedure of Mammography followed by biopsy of the tumour or lesions present in the breast tissue.Early prediction of breast cancer is one of the most essential works in the follow-up process. The objective of this study is to detect the feasibility of utilising previous medical records and determine the probability of being affected by malignant breast cancer disease. Gradient Boosting Ensemble method is utilised for this purpose to obtain and detect patients with severity. Comparative Analysis shows that the proposed method achieves encouraging result with an accuracy of 97.34%, MCC of 0.938, F1-Score of 0.97 Cohen-Kappa Score of 0.94 and MSE of 0.0266. The discussion carried out through this paper addresses the problem of early breast cancer detection. The use of machine learning techniques is recognized to be suitable for detecting the onset of early breast cancer detection. The performance of different classifiers described using different indexes are also presented in the paper.In conclusion, the Gradient Boost algorithm attains the lowest error rate along with highest efficiency which might be the best choice of algorithm for this problem and prediction of disease.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Chaurasia V, Pal S. International Journal of Computer Science and Mobile

Computing Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability. Int J Comput Sci Mob Comput [Internet]. 2014;3(1):10–22. Available:www.ijcsmc.com

2. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013;7(DEC).
3. Cunningham P, Delany SJ. K -Nearest Neighbour Classifiers. Mult Classif Syst. 2007;1–17.
4. Osuna E, Platt J. Support vector machines; 1978.
5. Rish I. An empirical study of the naïve bayes classifier an empirical study of the naive bayes classifier. 2014;41–6.
6. Sharma H, Kumar S. A survey on decision tree algorithms of classification in data mining. Int J Sci Res. 2016;5(4):2094–7.
7. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: A statistical view of boosting. Ann Stat. 2000;28(2): 337–407.
8. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
9. LG A, AT E. Using Three machine learning techniques for predicting breast cancer recurrence. J Heal Med Informatics. 2013; 04(02).
10. Ahmed Medjahed S, Ait Saadi T, Benyettou A. Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules. Int J Comput Appl. 2013;62(1):1–5.
11. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: A comparison of three data mining methods. Artif Intell Med. 2005;34(2):113–27.
12. Dumitru D. Prediction of recurrent events in breast cancer using the Naive Bayesian classification. Ann Univ Craiova, Math Comp Sci Ser. 2009;36(2): 92–6.
13. Ayer T, Chhatwal J, Alagoz O, Kahn CE, Woods RW, Burnside ES. Informatics in radiology: Comparison of logistic regression and artificial neural network models in breast cancer risk estimation. Radiographics. 2010;30(1):13–22.
14. Michael Kahn, St. Louis. UCI Machine Learning Repository.
Available:http://archive.ics.uci.edu/ml
Irvine. CA. University of California, School of Information and Computer Science.
15. Walters DE. Bayes's theorem and the analysis of binomial random variables. Biometrical J. 1988;30(7):817–25.
16. MH MNS. A review on evaluation metrics for data classification evaluations. Int J Data Min Knowl Manag Process. 2015;5(2):01–11.
17. Vieira SM, Kaymak U, Sousa JMC. Cohen's kappa coefficient as a performance measure for feature selection. 2010 IEEE World Congr Comput Intell WCCI; 2010.
18. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: An overview. Bioinformatics. 2000;16(5): 412–24.

© 2020 Dutta and Samir; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/58329