P-ISSN: 2715-2448 | E-ISSN: 2715-7199
Vol.5 No.2 June 2024
Buana Information Technology and Computer Sciences (BIT and CS)

Early Breast Cancer Detection in Coimbra Dataset Using Supervised Machine Learning (XGBoost)

Ahmed Sami Jaddoa
Business Informatics College, University of Information Technology and Communications, Iraq
Email: ahmed.sami@uoitc.edu.iq

Abstract

Worldwide, breast cancer (BC) represents one of the serious health concerns for adult females. The early detection and accurate prediction of risks are vital for the provision of optimum care and enhancement of patient outcomes. In the past few years, promising large data merging and ensemble learning algorithms appeared for the purpose of classification and prediction of BC risk. In the area of medical applications, methods of machine learning (ML) are crucial. Early diagnosis is necessary for a more efficient carcinoma treatment. This study's aim is to classify the carcinoma with the use of the 10 predictors that are found in Breast Cancer Coimbra dataset (BCCD). Presently, early diagnoses are necessary. The rates of cancer survival could be raised in the case where it is discovered early. Methods of machine learning offer effective way for data classifying and making early disease diagnoses. This study utilizes BCCD for the classification of BC cases utilizing XGBoost algorithm. Based on performance criteria, early detection of BC is the primary goal. The XGBoost classifier in this research achieved 98% precision, 98.32% accuracy, 99% f1-score, and 97% recall.

Keyword: Machine Learning, XGBoost, Z-score, BCCD.

1. Introduction

Humans are more vulnerable than ever to many forms of cancer in the last few years. An estimated one in six deaths globally result due to cancer, which makes it one of the top death causes globally. BC is the most prevalent type of cancer in terms of newly diagnosed cases. About 40,920 women died from BC alone in the year 2018. The World Health Organization (WHO) estimated that 2.90 million women receive a BC diagnosis annually. No less than 100 diseases that affect various bodily parts are referred to be cancer [1]. The most prevalent cancer all over the world is BC. In the year 2020, there will likely be no less than 2.26 million new cases of BC, according to WHO research on the disease's current and prospective impact [2]. The retrieval and maintenance of patients' electronic medical records and related devices are examples of healthcare technology.

It has never been easy to diagnose and treat hematological diseases when cancer is present. Nowadays, a staggering portion of the populace suffers from one or more diseases. Medical research has made enormous strides in the last few years. Even with such advancements, the general population still knows incredibly little about disease and health. It's possible that a sizable section of the populace has health problems, some of which could be lethal [3]. Through creating a prediction model, it is possible to diagnose diseases early on and provide patients with more effective care. In earlier research, ML-based models were employed to identify BC, and they show noteworthy efficacy [4]. One aspect of AI that enables the system to obtain information without explicit expertise is ML. Supervised algorithms are employed in many classification applications because they leverage human outputs and inputs to improve prediction accuracy and streamline the training process. As a result, the use of ML in healthcare has expanded [5][6]. ML is

emerging as a major diagnostic tool for patients in the medical field. In the case when a task is large and difficult to program, ML is used as an analytical technique. Examples of such tasks include anticipating pandemics, evaluating genomic data, and turning medical records into knowledge [7].

2. Literature Review

For classifying BC patients, many different kinds of studies were done recently.

Research by Sakri et al. [8] looked into a number of data mining (DM) methods to predict the recurrence of BC. They employed particle swarm optimization (PSO) as feature selection technique for K-Nearest Neighbor (K-NN), naïve Bayes (NB), and rapid decision tree (DT) learners in order to increase the accuracy of the prediction model. The results of this study have shown that the prediction model of the recurrence of BC through using fast decision tree learner (REPTree) as classifier had higher accuracy rate, which has been equal to 76.3i% in comparison to K-NN and NB classifiers.

Keles *etal*. [9] used non-invasive and painless approaches for BC prediction and diagnosing with the use of DM algorithms. They have studied several algorithms for BC classification utilizing ten-fold cross-validation method for the assessment of each one's predictive power for the relative results, using readings from antenna as a dataset. The optimal algorithms were IBK, Bagging, Random Committee, RF, and Simple Classification and Regression Tree (SimpleCART), which had an over 90% detection accuracy.

Through using patient data that is related to BC, Ferroni *etal*. [10] have made an identification of ML-based decision support system (DSS) significance combined with the random optimization (RO). The DSS model was built with the use of multiple kernel learning (MKL), which has first been created for assessing risks of cancer-related thrombosis. The MKL was expanded for the prediction of disease progression risks for patients with Breast Cancer in oncology setting. As it has been foreseen, their suggested model had produced a 10.90 hazard ratio (HR) and a C-index for PFS (i.e., progression-free survival) of 0.84, with a 86% rate accuracy.

Akben [11] had presented a Decision Tree model for the diagnosis of BC by utilizing Coimbra data-set. This DT had employed Gini index in their research for ascertaining attribute importance degree. Compared with existing models, which include K-NN, ANN, SVM, NB, adaptive boosting (AdaBoost), and so on, the proposed diagnostic approach had a 90.52% accuracy rate, according to results.

An ensemble model for the prediction of breast cancer has been applied by Nanglia et al. [12] to the Coimbra data-set. In this study, stacking has been utilized for the construction of ensemble model that consists of the three ML algorithms SVMs, DT, and KNN. Comparing their suggested ensemble model to other classifiers they employed in the research; it achieved the highest accuracy score of up to 78%. Moreover, the chi-square approach was used for the determination of the top five features, which included glucose, insulin, BMI, resistin, and homeostasis model assessment (HOMA) values.

3. Materials And Methods

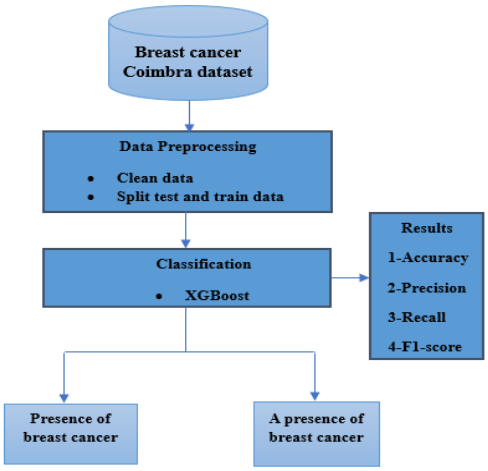

Fig.1. Process of BC Detection

A. Dataset and Attributes

In the case when related works have been analyzed, various approaches were utilized to diagnose BC. Multiple datasets are accessible for the identification of BC. UCI ML Repository provided the BCCD, which consists of 4000 observations and 10 attributes, one of which is a class variable (1 = Healthy, 2 = Patient). The dataset's attributes are listed in Table 1.

Table1. Data-set Attributes

No	Attribute name	Type
1	Age	Numeric
2	BMI	Numeric
3	Glucose	Numeric
4	Insulin	Numeric
5	HOMA	Numeric
6	Leptin	Numeric
7	Adiponectin	Numeric
8	Resistin	Numeric
9	MCP.1	Numeric
10	Classification	Numeric

B. Dataset Preprocessing

One of the most important phases of ML classification is data preprocessing, since cleaner data typically yields higher classification results. The model's pre-processing methods are outlined as follows: reduce noisy data: In ML, attribute noise as well as class noise are the two types of data noise. Nonetheless, attribute noise is decreased in the suggested model to improve accuracy. The data-set is split up into sections for process analysis training and testing. Thirty percent is utilized for testing and seventy percent is used for training.

C. Classification

Through classifying the records into predefined classes, classification is a model utilized for predicting the future behavior regarding the data. With the use of testing and training data sets, accurate disease detection is possible in classification. To construct the prediction, it suggests using two ML models. Test data is used after training data across ML models, i.e., XGBoost predicts using each learned model individually. Among the criteria are recall, f1-score, accuracy, and precision.

a. Extreme gradient boosting (XGBoost) Classifier

XGBoost can be defined as a gradient-boosting method that makes use of DTs and ensemble ML methods. This ML method can yield very important results because of its scalability and high processing speed. XGBoost is applied to both classification and regression tasks. The concept of this method is to find weak classifiers iteratively in order to get a correct classification. It creates customized DTs through the gradient descent technique through first establishing a range of threshold values, which are after that updated repeatedly through reducing residuals throughout tree construction. Regression trees can be considered as the weak learners when gradient boosts are applied, and each one maps an input dataset that shows one of its leaves contains a continuous mark. Regularized function (L1 & L2) that is minimized is a convex loss function that is based upon the difference between the goal output and predictions. In order to predict the mistakes or residues of earlier trees, the training iteratively adds new trees that are subsequently integrated with the earlier trees to provide the final prediction [13].

b. Performance metric evaluation method

Model performance is assessed using the Accuracy, F1-score, Recall, and Precision metrics. Eqs. (1) and (2) characterize the accuracy value regarding the classification model, respectively; comparably, Eqs. (3) and (4) denote the mathematical formula of Recall and the F1-score, respectively [14].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1-score} = \frac{2*(Precision*Recall)}{(Precision+Recall)} \quad (4)$$

4. Results

Following the algorithms described in the preceding section, the following outcomes were attained:

Table 2. Performance measure of **XGBoost** classifier

Classifier	Accuracy	Precision	Recall	F1-score
XGBoost	98.32%	98%	97%	99%

5. Conclusion

A predictive model for the likelihood of BC development was reported in this paper. The suggested model's study and evaluation demonstrate how effective and simple it is to apply. For patients with classified BC, the XGBoost algorithm is used on the BCCD. With the use of Z-score outlier detection method, one can identify outliers in a clean dataset during data preparation and apply Robust Scaling to address the outliers. Afterward, XGBoost algorithm was used. Based on the examination of test outcomes, the XGBoost algorithm has attained the maximum accuracy of 98.32%.

References

[1] S. Aamir *et al.*, "Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques," *Comput. Math. Methods Med.*, vol. 2022, 2022, doi: 10.1155/2022/5869529.

[2] R. Gonzales Martinez and D. M. van Dongen, "Deep learning algorithms for the early detection of breast cancer: A comparative study with traditional machine learning," *Informatics Med. Unlocked*, vol. 41, no. April, p. 101317, 2023, doi: 10.1016/j.imu.2023.101317.

[3] Y. Amethiya, P. Pipariya, S. Patel, and M. Shah, "Comparative analysis of breast cancer

detection using machine learning and biosensors," *Intell. Med.*, vol. 2, no. 2, pp. 69–81, 2022, doi: 10.1016/j.imed.2021.08.004.
[4] G. Alfian *et al.*, "Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method," *Computers*, vol. 11, no. 9, 2022, doi: 10.3390/computers11090136.
[5] A. sami Jaddoa, S. J. Saba, and E. A.Abd Al-Kareem, "Liver Disease Prediction Model Based on Oversampling Dataset with RFE Feature Selection using ANN and AdaBoost algorithms," *Buana Information Technology and Computer Sciences (BIT and CS)*, vol. 4, no. 2. pp. 85–93, 2023, doi: 10.36805/bit-cs.v4i2.5565.
[6] C. Li, Y. Weng, Y. Zhang, and B. Wang, "A Systematic Review of Application Progress on Machine Learning-Based Natural Language Processing in Breast Cancer over the Past 5 Years," *Diagnostics*, vol. 13, no. 3, 2023, doi: 10.3390/diagnostics13030537.
[7] Ahmed Sami Jaddoa, "Heart Disease Prediction System Using (SMOTE Technique)," vol. 050006, 2023.
[8] S. B. Sakri, N. B. Abdul Rashid, and Z. Muhammad Zain, "Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction," *IEEE Access*, vol. 6, pp. 29637–29647, 2018, doi: 10.1109/ACCESS.2018.2843443.
[9] M. Kaya Keleş, "Breast cancer prediction and detection using data mining classification algorithms: A comparative study," *Teh. Vjesn.*, vol. 26, no. 1, pp. 149–155, 2019, doi: 10.17559/TV-20180417102943.
[10] P. Ferroni, F. M. Zanzotto, S. Riondino, N. Scarpato, F. Guadagni, and M. Roselli, "Breast cancer prognosis using a machine learning approach," *Cancers (Basel).*, vol. 11, no. 3, pp. 1–9, 2019, doi: 10.3390/cancers11030328.
[11] S. B. Akben, "Determination of the Blood, Hormone and Obesity Value Ranges that Indicate the Breast Cancer, Using Data Mining Based Expert System," *Irbm*, vol. 40, no. 6, pp. 355–360, 2019, doi: 10.1016/j.irbm.2019.05.007.
[12] S. Nanglia, M. Ahmad, F. Ali Khan, and N. Z. Jhanjhi, "An enhanced Predictive heterogeneous ensemble model for breast cancer prediction," *Biomed. Signal Process. Control*, vol. 72, no. July 2021, p. 103279, 2022, doi: 10.1016/j.bspc.2021.103279.
[13] M. M. Hassan *et al.*, "A comparative assessment of machine learning algorithms with the Least Absolute Shrinkage and Selection Operator for breast cancer detection and prediction," *Decis. Anal. J.*, vol. 7, no. May, p. 100245, 2023, doi: 10.1016/j.dajour.2023.100245.
[14] Z. Salod and Y. Singh, "Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol," *J. Public health Res.*, vol. 8, no. 3, pp. 112–118, 2019, doi: 10.4081/jphr.2019.1677.