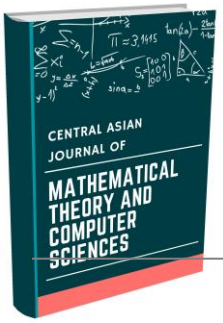

CENTRAL ASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES

https://cajmtcs.centralasianstudies.org

Volume: 04 Issue: 1 | Jan 2023 ISSN: 2660-5309

PREDICTION AND DIAGNOSIS OF BREAST CANCER USING MACHINE LEARNING AND ENSEMBLE CLASSIFIERS

Muhammad Waqas Arshad

Department of Computer Science and Engineering, University of Bologna, Italy

Muhammad.waqas.arshad.1@gmail.com

Abstract

There are considerably more breast cancer fatalities each year. The most common kind of cancer and the main cause of death in women worldwide is this one. A healthy life depends on every development in the prognosis and diagnosis of cancer sickness. The standard of treatment and patient survival rate must be updated, thus an accurate cancer prognosis is crucial. Research has demonstrated that machine learning approaches are effective for the early detection and prediction of breast cancer and have grown in popularity. Random Forest, Logistic Regression, Xtreme Gradient, and AdaBoost Classifier are trained on the Breast Cancer Wisconsin Diagnostic dataset, and their efficacy is assessed and compared in this study using ensemble classifier and machine learning. The major objective of this study is to identify the most effective ensemble and machine learning classifiers for breast cancer detection and diagnosis in terms of Accuracy and AUC Score.

ARTICLEINFO

Article history:
Received 6 Nov 2022
Revised form 5 Dec 2022
Accepted 17 Jan 2023

Keywords: *Prediction, Breast cancer, Machine Learning, Ensemble Classifier.*

---

1. INTRODUCTION

Breast cancer is defined as any kind of malignant tumor that arises in the breast as shows in Figure 1. It affects around 10% of all women at some point in their lives, making it the most common kind of cancer in women. Breast cancer is the second biggest cause of mortality among women, behind lung cancer (after lung cancer). After lung cancer, breast cancer is the second leading cause of death for women (after lung cancer). In the US, invasive breast cancer is anticipated to afflict 246,660 women in 2016, and the illness is anticipated to claim the lives of 40,450 women. The only approaches to stop this tumor from spreading are early discovery and prompt treatment. From 6.2 million cases in 2000 to 10 million cases in 2020, cancer mortality has similarly grown [1]. One in every six deaths is brought on by cancer. This demonstrates how crucial it is to provide funding for both the cancer battle and cancer prevention. Information and communication technology (ICT) must be employed effectively in medical practice in order to modernize the healthcare system, and particularly cancer therapy. Actually, the amount of data and the amount of value that can be derived from it have changed as a result of big data. Big data has greatly changed corporate

intelligence by analyzing a massive amount of unstructured, varied, non-standard, as well as healthcare data (BI). Since it not only predicts but also assists in decision-making, it is usually seen as a breakthrough in continuing innovation with the objective of enhancing patient care quality and reducing healthcare costs [2].

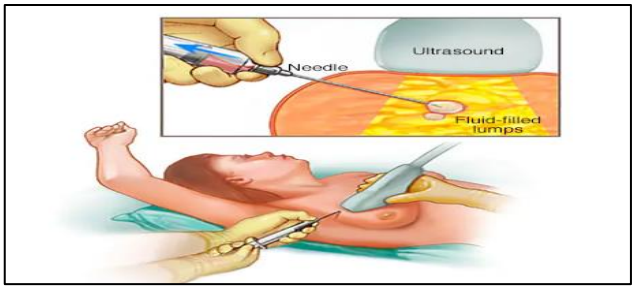

Figure 1. It shows the diagnosis process of the breast cancer [3].

Many machine learning approaches may be used to detect and diagnose breast cancer. Other machine learning approaches include Random Forest, Logistic Regression, Xtreme Gradient, and AdaBoost Classifier. Academics have used a variety of datasets in their research on breast cancer, including the SEER dataset, mammography images preserved in databases, the Wisconsin Dataset, and datasets from other institutions. Authors may complete their study by extracting and selecting unique data from these databases. These are fascinating research. The author uses 3D images to demonstrate the categorization of breast cancer using several supervised machine learning algorithms, concluding that SVM is the best choice overall [4]. Research on a comparison study of Relevance vector machine, on the other hand, has revealed that RVM is better to other machine learning algorithms for identifying breast cancer even when the variables are limited and reaches 97 percent accuracy [5]. In compared to other machine learning approaches utilized for breast cancer diagnosis, RVM has a low processing cost. The Support Vector Machine (SVM) predicts and detects breast cancer with the highest accuracy and lowest error rate [6]. Our research focuses on evaluating machine learning approaches and algorithms to determine the best strategy for breast cancer detection and prediction.

The body of this research paper is structured as follows. Section 2 describes the techniques and findings of prior research on breast cancer diagnosis. The proposed methodology and recommended technique for research is described in Section 3. Section 4 provides and elaborates on the outcomes of the experiments. Section 5 brings the paper at end with conclusion.

2. LITERATURE REVIEW

For the purpose of identifying and forecasting breast cancer, there are several machine learning techniques accessible. A few machine learning techniques include the Random Forest, Logistic Regression, Xtreme Gradient, and AdaBoost Classifier. A number of datasets, including the SEER dataset, mammography picture databases, the Wisconsin Dataset, and datasets from other institutions, have been used by several researchers to study breast cancer. To finish their inquiry, authors extract and choose distinguishing traits from numerous databases. These are significant studies. The author employs 3D images to demonstrate the use of several supervised machine learning algorithms in the categorization of breast cancer, and he concludes that SVM is the best choice based on his overall performance [4]. Research comparing Relevance vector machine to other machine learning approaches for breast cancer diagnosis, on the other hand, finds that it has a low processing cost. This helps to explain why RVM outperforms other machine learning algorithms for identifying breast cancer, even with less variables, and reaches 97 percent accuracy [5]. Support vector machine (SVM) demonstrates its value in breast cancer diagnosis and prediction with the highest accuracy and lowest error rate [6]. Our study examines multiple machine learning methodologies and algorithms in order to determine the best choice for early breast cancer detection and diagnosis.

A potent data analytics strategy that may be effective with datasets associated to breast cancer. The technique takes into account both cancer survivability and patient survival. For identification, the Surveillance, Epidemiology, and End Results (SEER) tool was used, and for classification, the Self-

Organizing Map (SOM) and the (DBSCAN) tools were used. Table 1 lists the objectives and methods of current, relevant studies.

Recent Related Studies.

Table. 1: The relative study of the Past research work.

Objective	Reference	Methods and Approaches Adopted
Clarification of data science and applications	[07]	Related Work
	[08]	Naïve Bayes, decision tree, support vector machine
	[09]	neural networks, SVM, decision tree, Naïve Bayes
	[10]	k-nearest neighbors , SVM, naïve Bayes, decision tree (c4.5),
	[11]	Neural network, c4.5 decision tree, Naïve Bayes
Prediction of breast cancer	[12]	Neural networks, Logistic regression, nearest neighbors ,decision tree,
	[13]	Naïve Bayes classifier, Support vector machine (SVM), adaboost tree, artificial neural network (ANN),
	[14]	The J48 decision trees and Naïve Bayes.
	[15]	Multilayer-perceptron , Naïve Bayes and support vector machine-sequential minimal optimization,

3. METHODOLOGY

Since our main objective in doing this research was to find the most effective and accurate approach for detecting breast cancer, we employed machine learning and ensemble classifiers. Random Forest, Logistic Regression, Xtreme Gradient, and AdaBoost Classifier are utilized to assess the results and establish which model is more accurate using the Breast Cancer Wisconsin Diagnostic dataset.

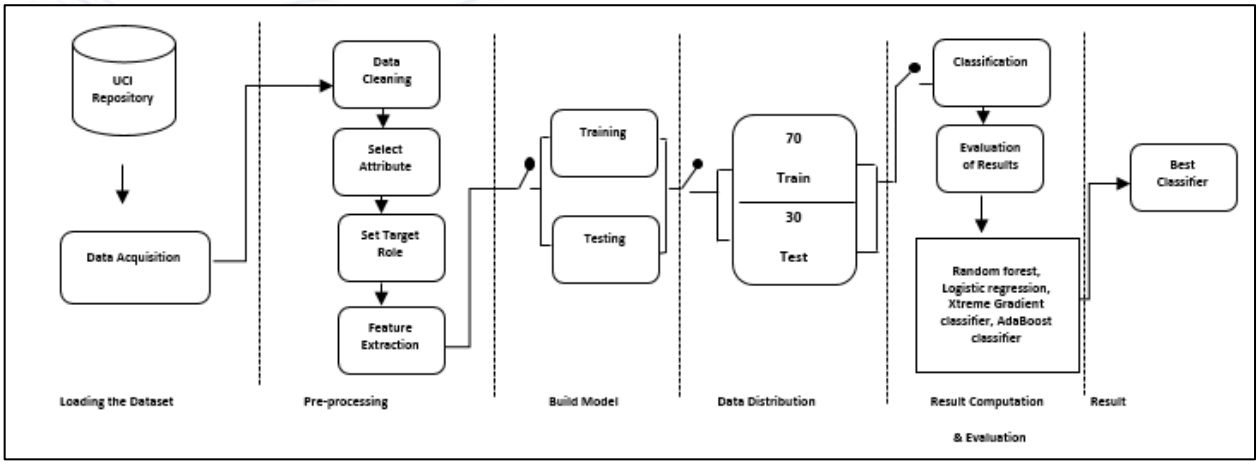

Figure 2. The proposed methodology of our research work describe the overall process from dataset start to its final result.

Pre-processing, which is broken down into four stages and includes data cleaning, attribute selection, target Role creation, and feature extraction, is the second step in our strategy, as shown in Figure 2. Data collection is the first step in our strategy. The machine learning model based on the processed data was successful in identifying breast cancer using a fresh set of measurements. In order to assess the performance of the algorithms, we continuously feed the model new data along with labels. This is often accomplished by

utilizing the Train Test Split approach to split the labelled data that we have gathered into two halves, with the ratios 70:30 being used for training and testing, respectively.

3.1. Algorithms

In our study, we realize the ensemble classifier and machine learning predictive analysis. The algorithms employed in our project are as follows:

- The class that would correspond to the mean of each prediction is predicted by random forests regression after training a large number of decision trees for classifying or regression. Random choice forests are a solution to the issue of overfitting a training set for decision trees.
- Logistic regression, a form of linear regression, is a particularly successful modelling approach [16]. Logistic regression is used to anticipate the chance of an illness and other health problems based on a risk factor (and variables). Using both basic and multivariate logistic regression, the relationship between an independent variable (s) (Xi), also known as the exposure and predictor variables, and a binary dependent variable (Y), also known as the outcome or a response variable, is investigated. It is often used to forecast binary or multiclass dependent variables.
- Extreme Gradient Classifier: Gradient boosting is an ensemble machine learning technique that may be used to problems with classification and regression predictive modelling. A quick open-source variation of the gradient boosting technique is called "extreme gradient boosting," or "XGBoost." Because of this, XGBoost is an open-source project, a Python library, and an algorithm. It is meant to have very high computational efficiency, maybe exceeding the most current open-source versions. The technological goal of going above the computational resource limit for boosted tree algorithms is referred to as "Xgboost."
- One of the meta-estimators, the AdaBoost Classifier, fits multiple instances of a classifier on the same dataset, modifies the weights of instances that were incorrectly classified, and then applies the classifier to the provided data [17]. This allows subsequent classifiers to focus on challenging cases.

3.2. Dataset acquisition

We used the Breast Cancer Wisconsin Diagnostic data from the University of Wisconsin Hospitals Madison Breast Cancer Network to carry out our study [14]. The characteristics of the dataset were extracted from a digital picture of a breast cancer sample obtained by fine-needle aspiration (FNA). These features specify the characteristics of the cell nuclei in the picture. Wisconsin has recorded 569 cases of breast cancer as shows in figure 3, with two categories (62.74 benign and 37.26 malignant) and 11 integer-valued characteristics (-Id -Texture -Area -Perimeter -Diagnosis -Radius -Smoothness -Compactness -Concavity -Perimeter -Concave points). Fractal dimension and symmetry).

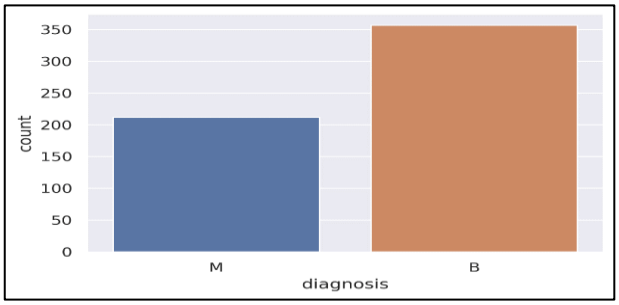

Figure 3. Data from Wisconsin Breast Cancer Diagnostic reveals that 357 benign(B) cases and 212 malignant (M) cases have been diagnosed.

3.3. Experiment Environment

The Scikitlearn package and the Python programming language were used for all of the testing on the machine learning algorithms discussed in this research. Scikit-learn, sometimes referred to as sklearn, is a Python-based open source machine learning toolkit [17]. It is designed to operate with Python's NumPy and

SciPy scientific and numerical libraries, and contains the classifiers Random Forest, Logistic Regression, Xtreme Gradient, and AdaBoost Classifier.

4. RESULTS AND DISCUSSION

The use of machine learning methods on the Wisconsin Diagnostic breast cancer dataset. We analyzed and assessed the models using the performance criteria Accuracy and AUC Score in order to choose the best algorithm for breast cancer prediction. A method for assessing classification task success when the result may be two or more unique classes is the confusion matrix as shows in figure 4 . An example of a confusion matrix is a table having the columns "Predicted," "Actual," "False Negatives (FN), "False Positives (FP)," and "True Positives (TP)" and "True Negatives (TN)". Accuracy is the most typical performance metric for classification algorithms. The ratio of occurrences that were accurately anticipated to all other expected events is how it is defined. The amount of precise documents that our ML model successfully discovered when doing document retrieval is known as precision. The sensitivity of a machine learning model refers to how many successful outcomes it produces. The F1 score offers a harmonic mean of accuracy and sensitivity. The weighted average of accuracy and sensitivity is used to get the F1 score. The accuracy percentages for the Wincson Breast Cancer Diagnostic datasets are shown in Tables 2. The results of the training and testing sets affect the accuracy of each classifier, although the logistic regression has a greater accuracy of 96.491 and AUC Score is 0.994 than other classifiers as shows in figure 5.

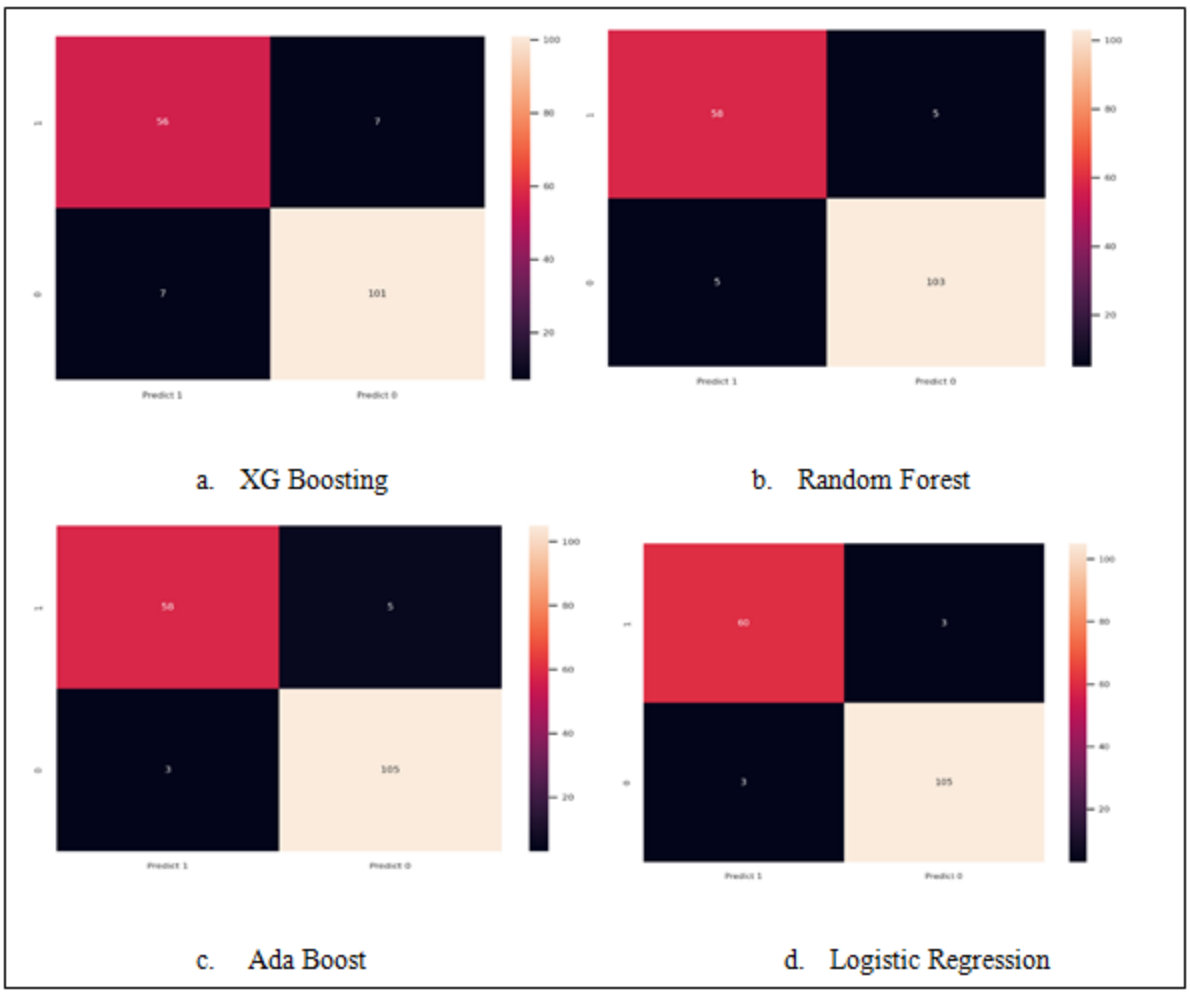

Figure 4. Illustrate the result of the models includes ‘a’ XG Boosting, ‘b’ Random Forest Classifier, ‘c’ Ada Boost and ‘d’ Logistic Regression. On the basis of confusion matrix.

Table 2. Comparison of Models on the basis of Accuracy and AUC Score

Sr No	Model Name	Accuracy (%)	AUC Score
1	Logistic Regression	96.491228	0.994709
2	AdaBoost	95.321637	0.992798
3	Random Forest Classifier	94.152047	0.985670
4	XGBoost	91.812865	0.982216

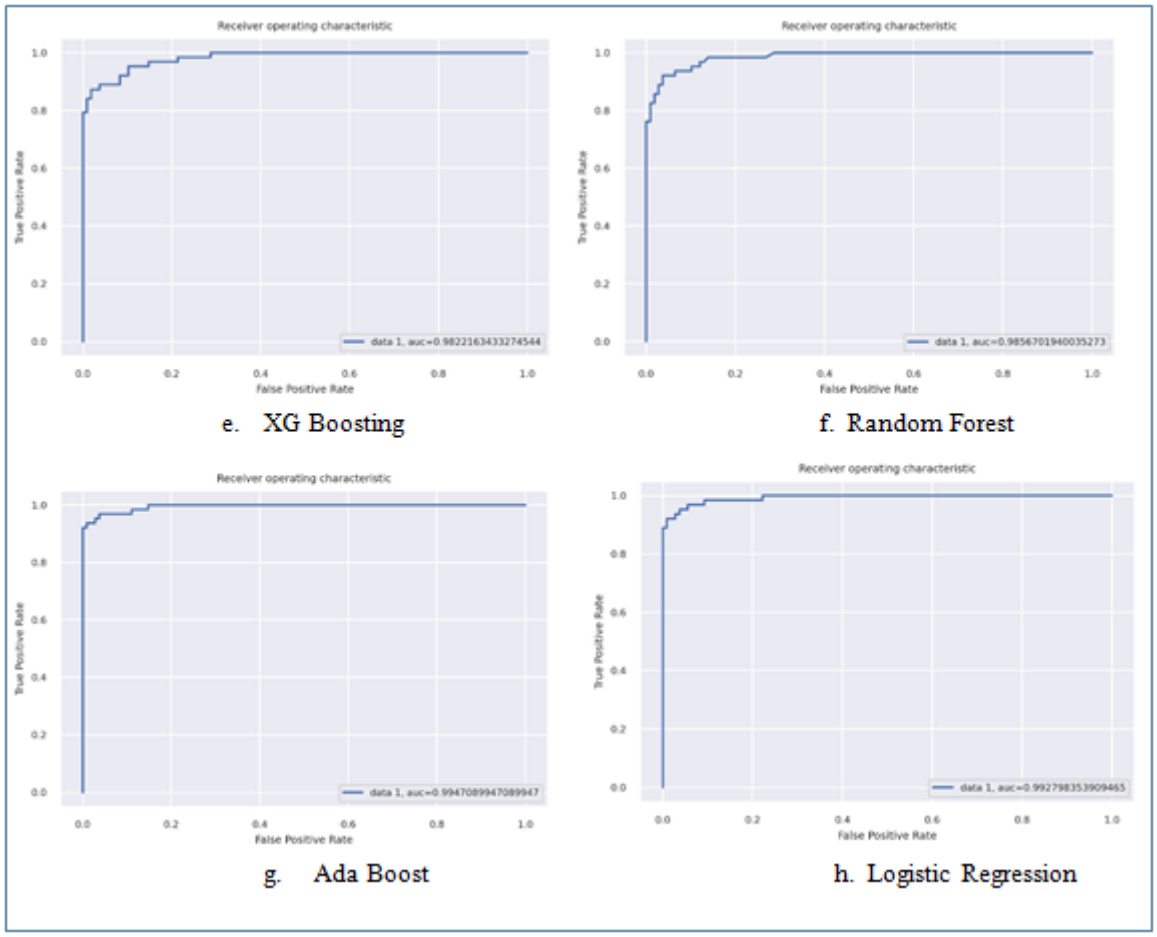

Figure 5. It shows the Area under curve for all model by representing ‘e’ by XG Boosting, ’f’ Random Forest , ‘g’ Ada Boost and ‘h’ for Logistic Regression. AUC score range in value from 0 to 1, value close to 1 means better prediction and score close to 0 shows wrong prediction [18].

5. CONCLUSION

In order to calculate, analyses, and evaluate the various outcomes obtained based on accuracy, and AUC score, we employed four main algorithms to the Wisconsin Breast Cancer Diagnostic dataset (WBCD): Random Forest, Logistic Regression, Xtreme Gradient, and AdaBoost Classifier. Finding the most reliable, accurate, and high-accuracy algorithm was the goal. With the help of the scikit-learn package and the Anaconda environment, all algorithms were created in Python. A detailed study of our models shows that, with an 70:30 train-to-test split, logistic regression outperforms all other techniques with a greater efficiency of accuracy 96.49%, and AUC score of 0.9947. Additionally, logistic regression has shown its effectiveness in the recognition and forecasting of breast cancer, obtaining the best accuracy. It is important to keep in mind that all of the results are solely linked to the WBCD database in order to validate the findings collected from the database. Therefore, it is crucial to think about how similar techniques and strategies may be used

to other databases in future projects. Our next attempts will also make use of fresh variables and machine learning techniques on sizable data sets.

Conflict of Interest

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speaker's bureaus; membership, employment, consultancies, stock ownership, or other equity interest' and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Funding

"The authors declare that no funds, grants, or other support were received during the preparation of this manuscript."

REFERENCES

1. 'WHO | Breast cancer', WHO. http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ (accessed Feb. 18, 2020).
2. Datafloq - Top 10 Data Mining Algorithms, Demystified. https://datafloq.com/read/top-10-data-mining-algorithmsdemystified/1144. Accessed December 29, 2015.
3. "Breast Cancer - Diagnosis And Treatment - Mayo Clinic". Mayoclinic.Org, 2022, https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475. Accessed 23 July 2022.
4. S. Nayak and D. Gope, "Comparison of supervised learning algorithms for RF-based breast cancer detection," 2017 Computing and Electromagnetics International Workshop (CEM), Barcelona, 2017, pp.
5. B.M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1-5.
6. H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, 'Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis', Procedia Computer Science, vol. 83, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224
7. Dhar V. Data science and prediction. *Commun. ACM.* 2013;56:64–73. doi: 10.1145/2500499.
8. Aruna S., Rajagopalan S., Nandakishore L. Knowledge based analysis of various statistical tools in detecting breast cancer. *Comput. Sci. Inf. Technol.* 2011;2:37–45.
9. Chaurasia V., Pal S. Data mining techniques: To predict and resolve breast cancer survivability. *Int. J. Comput. Sci. Mob. Comput.* 2014;3:10–22.
10. Asri H., Mousannif H., Al Moatassime H., Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Comput. Sci.* 2016;83:1064–1069. doi: 10.1016/j.procs.2016.04.224.
11. Delen D., Walker G., Kadam A. Predicting breast cancer survivability: A comparison of three data mining methods. *Artif. Intell. Med.* 2005;34:113–127. doi: 10.1016/j.artmed.2004.07.002.
12. Bernal J.L., Cummins S., Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: A tutorial. *Int. J. Epidemiol.* 2017;46:348–355.
13. Wang H., Yoon W.S. Breast cancer prediction using data mining method; Proceedings of the 2015 Industrial and Systems Engineering Research Conference; Nashville, TN, USA. 30 May–2 June 2015.

14. Williams T.G.S., Cubiella J., Griffin S.J. Risk prediction models for colorectal cancer in people with symptoms: A systematic review. *BMC Gastroenterol.* 2016;16:63. doi: 10.1186/s12876-016-0475-7.

15. Nithya R., Santhi B. Classification of normal and abnormal patterns in digital mammograms for diagnosis of breast cancer. *Int. J. Comput. Appl.* 2011;28:0975–8887. doi: 10.5120/3391-4707.

16. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York, NY: Springer-Verlag;2001.

17. Chengsheng, Tu & Huacheng, Liu & Bing, Xu. (2017). AdaBoost typical Algorithm and its application research. MATEC Web of Conferences. 139. 00222. 10.1051/matecconf/201713900222.

18. "Understanding AUC - ROC Curve". Medium, 2021, https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5. Accessed 23 July 2022.