A Machine Learning Approach to Predict Breast Cancer Using Boosting Classifiers

Md. Mijanur Rahman
Department of Computer Science and Engineering
Dhaka 1213, Bangladesh
mijanur.rahman@seu.edu.bd

Zannatul Ferdousi
Department of Computer Science and Engineering
Dhaka 1213, Bangladesh
2019000000089@seu.edu.bd

Puja Saha
Department of Computer Science and Engineering
Dhaka 1213, Bangladesh
2019000000083@seu.edu.bd

Renesha Amin Mayuri
Department of Computer Science and Engineering
Dhaka 1213, Bangladesh
2019000000063@seu.edu.bd

Abstract

Breast cancer is a prevalent disease, with the second highest incidence rate among all types of cancer. The risk of death from breast cancer is increasing due to rapid population growth, and a dependable and quick diagnostic system can assist medical professionals in disease diagnosis and lower the mortality rate. In this study, various machine-learning algorithms are examined for predicting the stages of breast cancer, and most especially in the medical field, where those methods are widely used in diagnosis and analysis for decision-making. We focused on boosting classification models and evaluated the performance of XGBoost, AdaBoost, and Gradient Boosting. Our goal is to achieve higher accuracy by using boosting classifiers with hyperparameter tuning for the prediction of breast cancer stages, precisely the distinction between "Benign" and "Malignant" types of breast cancer. The Wisconsin breast cancer dataset is employed from the UCI machine learning database. The performance of our model was evaluated using metrics such as accuracy, sensitivity, precision, specificity, AUC, and ROC curves for various strategies. After implementing the model, this study achieved the best model accuracy, and 98.60% was achieved on AdaBoost.

***Keywords*: Machine Learning, Breast Cancer, Classification Algorithms, Xgboost, Adaboost, Gradient Boosting.**

1. Introduction

Cancer is considered the worst of all ailments. It is a group of illnesses that allow for erratic growth that could spread or invade certain bodily parts. Invasive breast cancer in women is anticipated to cause 245,299 new cases of diagnosis and 40,450 new mortality cases in the U.S cancer Statistics (CDC Report) in 2016. One type of cancer that begins in the breast is breast cancer. Benign and malignant are the two types of classes for cancer detection. A malignant tumor develops fast and damages its tissues by invading them [1]. Breast cancer cells typically develop a tumor frequently detectable on an x-ray or as a lump. Breast cancer can spread when cancerous cells enter the blood or lymphatic system and get carried to other parts of the body. Patients must endure breast cancer surgery, chemotherapy, radiotherapy, and endocrine treatments to stop cancer from spreading.

Worldwide, breast cancer is the second leading cause of death in women after heart disease. And it affects more than 8% of women at some point in their lives [2]. 97% of women can survive for more than 5 years with early detection despite the high survival rate. Statistics show that this condition has been responsible for a sharp rise in the number of fatalities in recent years. The major obstacle in treating it is early discovery. Therefore, some data science solutions must be included, in addition to medical remedies to address the death-causing problem. The objective of this research is to discover general trends that may aid in the selection of the best machine learning algorithm and hyperparameters, as well as to determine which traits are most effective in predicting whether a tumor is ‘benign’ or ‘malignant’.Turkki et al. [3] stated that prognostic assessment using machine learning techniques is possible without prior knowledge of breast cancer pathology. To do this, machine learning

classification algorithms are used to fit a function that can predict the discrete class of new data.ML is based on four steps: Collecting data, picking the model, training the model, and testing the model [4]. The research aims to distinguish between malignant and benign patients, categorize them, and plan how to parametrize our classification methods to use boosting classification algorithms with high accuracy. The Wisconsin Breast Cancer dataset is used for our data selection. In describing breast cancer, we investigated a variety of datasets and other machine learning techniques' potential. We aimed to maximize accuracy while lowering error rates. It came across multiple boosting classification techniques in machine learning, where different algorithms had been employed to predict the stages of breast cancer. We examined the supervised classification machine learning models and offered a useful technique for feature selection that lowers the features. The proposed model will be capable of classifying breast cancer, whether it is "Benign" or "Malignant".

Our target is to achieve higher accuracy by using boosting classifiers with hyperparameter tuning (Grid Search CV) for the prediction of breast cancer stages, specifically the distinction between "Benign" and "Malignant" types of breast cancer that will help the physician to identify the type of breast cancer correctly.

This paper is organized as follows. In section 2, Related Works is described. In section 3, the proposed methodology is presented, and the background study is described. In Section 4, the results of the study are discussed. Section 5 showed the discussion, and the conclusions are presented in Section 6.

2. Materials and methods

Related Works

Numerous innovative technologies for diagnosing breast cancer have been developed with the advancement of medical research. The following is a summary of the research in this field.

M. Raihan presented [5], Using two well-known ensemble machine learning methods, the breast cancer dataset was examined to provide predictions about breast cancer. Breast cancer was predicted using XGBoost and Random Forest. For this research, a total of 275 examples with 12 features were used. In this analysis, accuracy with the Random Forest method was 74.73%, and accuracy with XGBoost was 73.63%. V. L. Jyothi et al. [6] explained 4 machine learning classifiers Random Forest, Decision Tree, AdaBoost, and Gradient Boosting (GB) were compared to classify benign and malignant tumors. The performance of all four classifiers is assessed, and the results are utilized to compare them. The objective of the effort is to identify the best Machine Learning (ML) model for diagnosing breast cancer. The classification accuracy of GB, which outperformed all other models, was 95.82%.

Authors H. Gupta et al. [7] discussed and emphasized the importance of early detection of fatal diseases like breast cancer. In this regard, compared to other classifiers taken into consideration in this work, the CatBoost-based ML algorithm appears to be a better classifier for breast cancer prediction. It provides an accuracy of 97.8%, which is higher than the other ML techniques. Amrane et al. [8] compared, to accurately diagnose breast cancer patients, the two machine learning classifiers, Naive Bayes (NB) and K-Nearest Neighbors (K-NN). K-NN performed well in the comparison, scoring 97.51% accuracy, whereas NB scored 96.19% accuracy.

Sakri et al. [9] concentrated on improving the accuracy value and combining the ML techniques K-NNs, NB, and reduced error pruning (REP) tree with the feature selection algorithm particle swarm optimization (PSO). Their area of expertise includes the issue of breast cancer in Saudi Arabian women, which is one of the country's key issues. The Naive Bayes Classifier, RepTree Classifier, and K-NN classifier have achieved an accuracy of 70%, 76.3%, and 66.3%, respectively. They utilized the Weka tool to analyze their data. They have discovered four features that work best for this classification assignment when PSO is used. Yolanda et al. [10] Compared, with an accuracy of 74.14% when the GB machine learning algorithm was used and it is the most accurate classifier in predicting breast cancer using the CBCD.

In this study, M. Amine Naji et al [11] applied 5 machine-learning algorithms to the Breast Cancer Wisconsin Diagnostic dataset: Support Vector Machine(SVM), Random Forest, Logistic Regression, DT, and KNN. SVM performed better than all other classifiers and had the best accuracy of 97.2%. Subramanian PT et al [12] have highlighted Naive Bayes strategies in their study comparing Tree Augmented Naive Bayes (TAN), and Bayes Belief Network (BBN). Their research indicates that 91.7%, and 94.1% accuracy have been attained for BBN, and TAN, respectively, with the use of Gradient boosting.

On the WBCD dataset, [13] applied the Naive Bayes Classifier, Radial Basis Function Network, and Decision Tree Classifier algorithms. The accuracy for Naive Bayes was 97.36%, which is higher than the accuracy for the Radial Basis Function network and the Decision Tree Classifier, which were 96.77% and 93.41%, respectively. Using decision tree variants, Azar et al. [14] presented a technique for the prediction of breast cancer. The single decision tree (SDT), and decision tree forest (DTF) are the modalities employed in this method. The results

showed that, in the training phase, the accuracy acquired by single decision trees was 97.07%. In the testing phase, DTF had an accuracy of 97.51%, whereas SDT had an accuracy of 95.75%.

In this section, we present a comparative study of several machine learning techniques for predicting breast cancer, such as SVM, KNN, Random Forest, SDT, CatBoost, and logistic-regression, Naïve Bayes, etc. The technique for learning the basic features and working principles of each machine is different. The highest accuracy achieved by Catboost was 97.8%.

3. Methodology

We started by collecting the data we intended to use for pre-processing and applying the classification methods. A data mining approach called "data pre-processing" entails putting raw data into a comprehensible format. Real-world data is frequently insufficient, inconsistent, and uncertain, containing several inaccuracies. Pre-processing data is a tried-and-true way to fix these problems. Raw data is prepared for subsequent processing by data pre-processing. We pre-processed the UCI dataset by using the standardization approach. This stage is crucial because the quality and volume of data you collect will directly affect how effective your prediction model will be.
The collection consists of 32 attributes and 569 examples. The two primary cancer designations for it are benign (B) and malignant (M). These are all the classes. A total of 357 benign instances, or 62.7% of all cases, outnumber 212 malignant cases, which account for 37.3% of all cases. The summary information for the classes is displayed below.

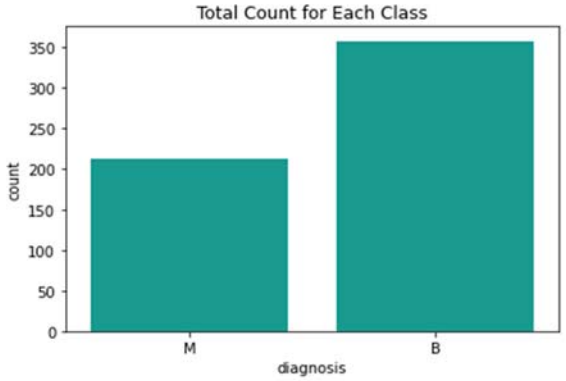

Figure 01: Class Distribution

The main workflow of this research is shown in Figure 02. We discussed the origin of the dataset, its features, and the surrounding background. This section concludes with a brief discussion of specific boosting classifier models and assessment techniques. Various pre-processing techniques and optimization techniques have boosted the performance of this work.

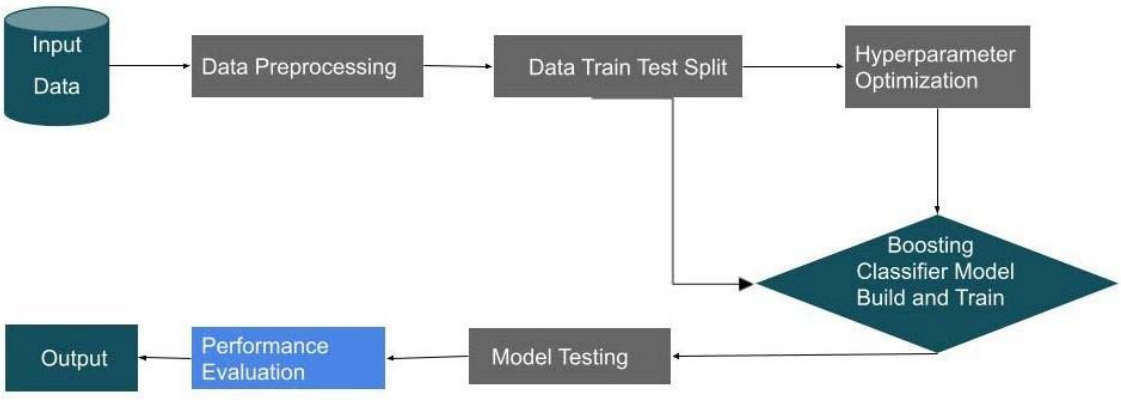

Figure 02: Research Methodology

Each feature's data type has been checked to ensure that there are no inconsistencies between data type and value. But as our dataset has no missing values and the class label is not imbalanced, we didn't have to take any action. We performed the MinMaxScaler technique. It basically helps to bring the dataset into a particular range of 0 to 1. Then, divide the dataset into "train" (75% of it) and "test" (25% of it), and use the training dataset to build three boosting machine learning models. The performance of three different models was evaluated on the testing dataset. Then, we performed the GridSearchCV hyperparameter technique to select the best set of parameters for our machine-learning models. Finally, we evaluated the performance of all algorithms on the testing dataset.

3.1 Gradient Boosting

A Gradient Boosting Decision Tree, is a particular case of boosting algorithms where errors are minimized by a gradient descent algorithm and produce a model in the form of weak prediction models decision trees. The major difference between boosting and gradient boosting is how both algorithms update models (weak learners) from wrong predictions. Each iteration adjusts the weight of each poor learner based on how well the learning went. The performance of the ensemble learner on the training datasets and validation datasets will gradually improve. Gradient boosting creates additive regression models by iteratively fitting current "pseudo"-residuals via least squares to a basic parameterized function. Most typical issues like regression, classification, and ranking can be handled with various loss functions.

3.2 XGBoost

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed decision gradient Boosting tree (GBDT) machine learning library. It provides a parallel tree boosting and is the leading machine-learning library for regression, classification, and ranking problems. It's vital to understand XGBoost first to grasp the machine learning concepts and algorithms that XGBoost builds upon supervised machine learning, decision trees, ensemble learning, and gradient Boosting.

3.3 AdaBoost

One of the first boosting models created was adaptive boosting (AdaBoost) [21]. Every time the boosting process is repeated, it adjusts and seeks to be self-correct. AdaBoost initially assigns each dataset the same weight. After each decision tree, it automatically modifies the data points' weights. This is done by providing misclassified cases to be updated with increased weights after an iteration. It keeps doing this until the residual error, or the gap between actual and anticipated values is below a certain level that is deemed acceptable.

3.4 Hyper Parameter Tuning

Finding the optimal combination of hyperparameters to enhance the model's performance is known as hyperparameter tuning (or hyperparameter optimization). It operates by conducting numerous trials within a single training procedure. Some advanced processes are required to perform it- Randomized Search and Grid Search CV, etc.

4. Results and discussion

4.1 Results

It is important to thoroughly evaluate the performance of our models to understand their strengths and weaknesses. One common way to do this is through the use of classification reports, which provide a summary of the model's performance in terms of various evaluation metrics. Some common evaluation metrics that may be included in a classification report include accuracy, precision, specificity, and recall. Another evaluation metric that may be included in a classification report is the AUC of the ROC curve. Overall, a thorough analysis of the results in terms of these evaluation metrics and other factors can provide valuable insights into the performance of the classification model and inform improvements in future iterations.

Table 01 Classification reports without hyperparameter tuning

Algorithm Names	Accuracy	Precisions	Specificity	Recall	AUC
AdaBoost	95.10%	0.988	0.981	0.932	0.99
XGBoost	96.50%	0.966	0.944	0.977	0.99
Gradient Boosting	95.80%	0.966	0.944	0.966	0.98

Table 02: Classification reports with hyperparameter tuning

Algorithm Names	Accuracy	Precisions	Specificity	Recall	AUC
AdaBoost	98.60%	0.988	0.981	0.988	1.00
XGBoost	97.20%	0.967	0.944	0.988	0.99
Gradient Boosting	95.80%	0.977	0.962	0.955	0.98

Using hyperparameter tuning, Table: 02, the performance of all three algorithms improved. The greatest improvement was seen in AdaBoost, which increased from 95.10% accuracy without tuning to 98.60% accuracy with tuning. XGBoost also improved, increasing from 96.50% accuracy before tuning to 97.20% after tuning. Gradient Boosting, on the other hand, had no change in accuracy after tuning, remaining at 95.80%.

Overall, it appears that hyperparameter tuning successfully improved the accuracy of both AdaBoost and XGBoost, with AdaBoost seeing the greatest improvement. It is possible that further improvements could be achieved through additional hyperparameter tuning or by trying different algorithms.
A confusion matrix makes it simple to summarize the performance of a classification method. Calculating a confusion matrix will help us better understand.

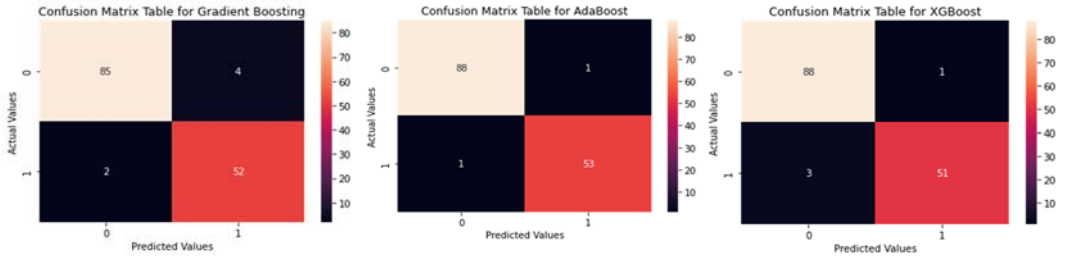

Figure 03: Confusion Matrix After Tuning

Area Under the Curve (AUC) calculates the two-dimensional area under the entire Receiver Operating Characteristic (ROC) curve ranging from (0,0) to (1,1).

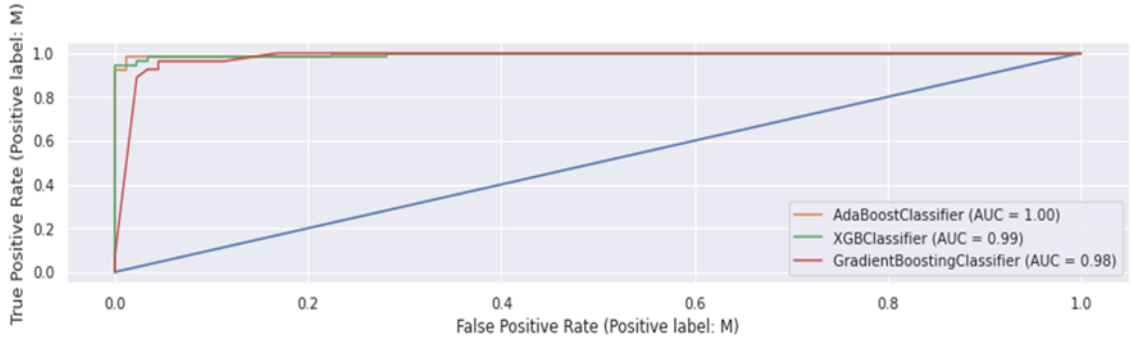

Fig 04: AUC & ROC Curve After Tuning

Our findings indicated that the suggested model AdaBoost Classifier had the highest AUC, which was 1.00. and XGBoost and Gradient Boosting had AUCs of 0.99 and 0.98.

4.2 Discussion

The study results indicate that AdaBoost and XGBoost performed better than Gradient Boosting in terms of accuracy. Afer hyperparameter tuning implementation, AdaBoost had the highest accuracy of 98.6%, followed by XGBoost with 97.2% and Gradient Boosting with 95.8%. In terms of other performance metrics such as precision, specificity, and recall, AdaBoost and XGBoost also performed better than Gradient Boosting. With respect to the confusion matrix, AdaBoost had the highest number of true positives and true negatives, indicating that it had the highest number of correct predictions. It also had the lowest number of false positives and false negatives, indicating that it had the lowest number of incorrect predictions. Overall, the results suggest that AdaBoost is the most effective machine-learning algorithm for this particular study. However, it is important to note that the performance of any machine learning algorithm depends on various factors, such as the quality of the data, the choice of features, and the hyperparameter tuning. Therefore, these results should be interpreted with caution, and further studies may be needed to confirm these findings.
In table 03, we made a comparison table for available prediction machine-learning models on breast cancer.

Table: 03 Comparison of publicly available prediction model

Ref No	Period	Dataset Name	Algorithm Name	Accuracy (%)
[5]	2020	UCI BCD	RF, XGBoost	74.73%,73.63%
[7]	2021	WBCD	CatBoost, XGBoost, DT, KNN	97.80%,97.08%,95.60%,97%
[15]	2021	WBCD	DT, AdaBoost,	90.20%,96.50%
[16]	2022	WBCD	GB, KNN	95.34%,75.96%
[17]	2022	WBCD	XGB, AdaBoost	98.24%,94.73%
[18]	2020	WBCD	AdaBoost, GB	96.81%,97.34%
[20]	2020	WBCD	GB, XGBoost, AdaBoost	95.96%,97.19%,95.96%
Our paper	2022	WBCD	GB XGBoost AdaBoost	95.80% 97.20% 98.60%

In this research, the performance of several boosting classifiers was evaluated on the WBCD dataset. The classifiers used were Decision Tree, AdaBoost, Gradient Boosting, XGBoost, CatBoost, and KNN. The accuracy of each classifier was measured and recorded. The results show that XGBoost and AdaBoost had the highest accuracy rates, with 98.60% and 97.20% respectively. GB also performed well, with a 95.80% accuracy rate. Compared with previous research results on the WBCD dataset, XGBoost and AdaBoost performed better than the other classifiers in this dataset. GB also performed similarly well compared with a 95.34% and 97.34% accuracy rate of other previous research. In conclusion, the results of this research suggest that AdaBoost is an effective classifier for the WBCD dataset to predict breast cancer.

5. Conclusions

The study aims to distinguish between malignant and benign patients, which can be highly helpful for patients and doctors in prescribing the right medications. The suggested approach can serve as an alternative to the current testing requirements and aid in the early diagnosis of breast cancer. Additionally, a side-by-side comparison is shown in this study, and the best classifier for the model that offers dependability is selected. With this research paper, we can see that among Gradient Boosting, XGBoost, and AdaBoost, the AdaBoost is the most accurate algorithm for the best accurate result for the detection of breast cancer type with an efficiency of 98.60%. In the medical industry, the diagnosing process is very time-consuming. The system proposed that a clinical assistant may be used to diagnose breast cancer using machine learning techniques.

In the future, to attain high accuracy, we plan to parametrize our categorization methods. In this study, our dataset is the shortest, but our research could have been done with a larger dataset, so we are investigating a variety of datasets and the potential applications of machine learning techniques to describe breast cancer further and want to maximize accuracy with a larger dataset while lowering error rates.

Conflict of interest

All authors declare no conflicts of interest in this paper.

6. References

[1] Key, T. J., Verkasalo, P. K., & Banks, E. Epidemiology of breast cancer. The lancet oncology, 2(3), 133-140, 2001.
[2] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, National Cancer Institute; 2012.
[3] Turkki, R.; Byckhov, D.; Lundin, M.; Isola, J.; Nordling, S.; Kovanen, P.E.; Verrill, C.; von Smitten, K.; et al. Breast cancer outcome prediction with tumor tissue images and machine learning. Breast Cancer Res. Treat., 177, 41–52, 2019. [CrossRef] [PubMed]
[4] S. Gokhale., "Ultrasound characterization of breast masses", The Indian journal of radiology & imaging, Vol. 19, pp. 242-249, 2009.
[5] Sajib Kabiraj,M. Raihan,Nasif Alvi,Marina Afrin,Laboni Akter,Shawmi Akhter Sohagi,Etu Podder. Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm. 10.1109/ICCCNT49239.2020.9225451, 2020.
[6] Kurian, Babymol & Jyothi, V.L. Comparative Analysis of Machine Learning Methods for Breast Cancer Classification in Genetic Sequences. Journal of Environmental and Public Health. 1-6. 10.1155/2022/7199290, 2022.
[7] H. Gupta, P. Kumar, S. Saurabh, K. Mishra, B. Appasani, A. Pati, C. Ravariu, A. Srinivasulu. Category Boosting Machine Learning Algorithm for Breast Cancer Prediction. Vol. 66, 3, pp. 201–206, Bucarest, 2021.
[8] M. Amrane, S. Oukid, I. Gagawa, and T. Ensari, “Breast cancer classification using machine learning,” in 2018 Electric Electronics, Computer Science, Biomedical Engineerings’ Meeting (EBBT), pp. 1–4,2018.
[9] Sakri SB, Rashid NBA, Zain ZM. Particle swarm optimization feature selection for breast cancer recurrence prediction. IEEE Access.6:29637–47 , 2018.

[10] Austria, Yolanda & Goh, Marie & Jr, Lorenzo & Lalata, Jay-Ar & Goh, Joselito & Vicente, Heintjie.Comparison of Machine Learning Algorithms in Breast Cancer Prediction Using the Coimbra Dataset. International journal of simulation: systems, science & technology.10.5013/IJSSST.a.20.S2.23 , 2019 .
[11] M. Amine Naji, S. El Filali, K. Aarika, EL Habib, RAit Abdelouhahid, Debauche, Machine Learning Algorithms for Breast Cancer Prediction And Diagnosis 2021.07.062.
[12] Banu AB, Subramanian PT. Comparison of Bayes classifiers for breast cancer classification. Asian Pac J Cancer Prev (APJCP). 19(10):2917–20, 2018.
[13] Chaurasia V, Pal S, Tiwari B. Prediction of benign and malignant breast cancer using data mining techniques. J Algorithms Comput Technol. 12(2):119–26 ,2018.
[14] Azar AT, El-Metwally SM. Decision tree classifiers for automated medical diagnosis. Neural Comput Appl. 23(7–8):2387–403, 2012.
[15] Tsehay Admassu Assegie , R. Lakshmi Tulasi , N. Komal Kumar.Breast cancer prediction model with decision tree and adaptive boosting.Vol. 10, No. 1, pp. 184-19, March 2021.
[16] Md. Samiul Islam, Md. Ashikuzzaman,Joy Mojumdar..Breast Cancer Detection using Machine Learning Techniques.Vol 184– No.39, 2022.
[17] Manav Mangukiya, Anuj Vaghani, Meet Savani . Breast Cancer Detection with Machine Learning 2321-9653, 2022.
[18] Shawni Dutta and Samir Kumar Bandyopadhyay,.Early Breast Cancer Prediction using Artificial Intelligence Methods.JERR.58329,2020.
[19] Malik, Shubham & Harode, Rohan & Singh, Akash. XGBoost: A Deep Dive into Boosting (Introduction Documentation). 10.13140/RG.2.2.15243.64803 , 2020..
[20] Pulung Hendro Prastyo, GedeYudi Paramartha, Michael S. Moses Pakpahan, Igi Ardiyanto.Predicting Breast Cancer: A Comparative Analysis of Machine Learning Algorithms. Volume 3, April 2020.
[21] Tharwat, Alaa. AdaBoost classifier: an overview. 10.13140/RG.2.2.19929.01122 , 2018 .
[22] Piegorsch, Walter. Confusion Matrix. 10.1002/9781118445112.stat08244 , 2020

Authors Profile

Md. Mijanur Rahman is currently an assistant professor in the department of Computer Science and Engineering and Director, of the Central for Artificial Intelligence and Robotics at Southeast University. His research interests are deep learning, machine learning, medical image processing, data science and recommendation system, and blockchain. Rahman has published in conferences and journals with more than 20 pepper-reviewed journals.

Zannatul Ferdousi received his Bachelor's in Computer Science and Engineering at Southeast University in 2023. Her research interests are Artificial Intelligence, Deep Learning, Machine Learning, Computer Vision, and Big Data Analytics

Puja Saha received his Bachelor's in Computer Science and Engineering at Southeast University in 2023. Her research interests are Artificial Intelligence, Machine Learning, and Computer Vision.

Renesha Amin Mayuri received a Bachelor's Degree in Computer Science and Engineering at Southeast University in 2023. Her areas of interest are artificial intelligence, database systems, networking, and machine learning.