Hyperparameter Tuning Based Machine Learning classifier for Breast Cancer Prediction

Md. Mijanur Rahman[1], Asikur Rahman Raju[2], Sumiea Akter Pinky[3], and Swarnali Akter[4]

[1]Assistant Professor, Dept. of Computer Science and Engineering, Southeast University. Dhaka, Bangladesh
[2, 3,4] Student, Dept. of Computer Science and Engineering Southeast University. Dhaka, Bangladesh

Abstract
Currently, the second most devastating form of cancer in people, particularly in women, is Breast Cancer (BC). In the healthcare industry, Machine Learning (ML) is commonly employed in fatal disease prediction. Due to breast cancer's favorable prognosis at an early stage, a model is created to utilize the Dataset on Wisconsin Diagnostic Breast Cancer (WDBC). Conversely, this model's overarching axiom is to compare the effectiveness of five well-known ML classifiers, including Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), and Naive Bayes (NB) with the conventional method. To counterbalance the effect with conventional methods, the overarching tactic we utilized was hyperparameter tuning utilizing the grid search method, which improved accuracy, secondary precision, third recall, and finally the F1 score. In this study hyperparameter tuning model, the rate of accuracy increased from 94.15% to 98.83% whereas the accuracy of the conventional method increased from 93.56% to 97.08%. According to this investigation, KNN outperformed all other classifiers in terms of accuracy, achieving a score of 98.83%. In conclusion, our study shows that KNN works well with the hyper-tuning method. These analyses show that this study prediction approach is useful in prognosticating women with breast cancer with a viable performance and more accurate findings when compared to the conventional approach.

Keywords:
Machine Learning, Breast Cancer Prediction, Grid Search, Hyperparameter Tuning.

1. Introduction

Cancer is one of the foremost mundane cognitive disorders that kill individuals. Breast cancer is the second-most prevalent malignancy globally, especially among women. Nearly 22.5 new instances of breast cancer per 100,000 females were reported in Bangladesh [1]. When compared to other types of cancer, Bangladeshi women have the greatest occurrence rate between the ages of 15 and 44 (19.3 per 100,000). According to WHO data published in 2020, Bangladesh's death rate has reached 6,808 or 0.95%. If breast cancer is discovered early, it can be treated easily and with fewer risks, which lowers the mortality rate by 25%.

To determine a patient's cancer status and whether they have it or not, the majority of clinicians perform a biopsy. Having benign cancer suggests the patient is safe because it is less harmful than malignant cancer. Benign cancer can be treated, in contrast to malignant cancer which is irreversible and spreads to other body parts [2]. For this cancer, indeed, neither a definitive cure nor even perfect outpatient care has been inferred. All doctors can currently only do this by saving the lives of those who are afflicted by this illness and giving them a second shot at life by stripping the ailing body part. Early detection and diagnosis are thus more important in lowering the mortality rate from breast cancer. After finding a breast tumor, the most arduous task is determining if the tumor is benign or malignant. Modern day breast cancer early detection uses a diversity of ML methods. ML techniques allow us to swiftly extract information from massive amounts of data, which then are used to predict outcomes. Therefore, ML classification is helpful in many sectors for early prediction and diagnosis. Many strategies are utilized to predict BC, however utilizing ML techniques, the prediction rate is soaring day by day. Data collection, selecting the optimal model, training the model, and testing are the four basic phases in ML for classification. According to a literature assessment of approaches employed by numerous researchers [2, 4-10, 15-17, 20] to predict breast cancer using the WDBC dataset, they all demonstrated how to evaluate the performance of a model via accuracy rate, precision, recall, and F1 score. However, more attention must be paid to this area if the accuracy rate is to be boosted, since this illness is extremely detrimental to every patient and is becoming more and more prevalent. Therefore, if the accuracy rate were raised to a level closer to 99%, it would aid healthcare professionals in predicting breast cancer early on before it becomes fatal.

This study's axiom is to applies five ML classifiers to the WDBC dataset for the prognosis of breast cancer. These classifiers include logistic regression, decision trees, random forest, K-nearest neighbors, and Naive Bayes. In order to enhance performance and choose adequate

Manuscript received February 5, 2024
Manuscript revised February 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.2.24

classifier parameters, here we apply key tactic hyperparameters that have been fine-tuned using a grid search methodology. Every dataset does not perform well with the default settings of classifier algorithms, hence hyperparameter tuning is chosen. In order to obtain a more accurate result, the best parameters for the dataset were selected in this technique.

The following sections are included in the work: After introduction a related work is shown. Thirdly the research methodology, including data collection, data pre-processing, the algorithms utilized and their general introduction is described. Fourthly the experimental findings are displayed, and the overall conclusion reached together with suggestions for future research are presented, the acknowledgment and references are displayed in the rest of the paper.

2. Related works

The world's most hazardous and predominant illness that primarily distresses women is cancer. There are extensive forms of cancer, including breast, lung, ovarian, and brain diseases. Out of all these malignancies, breast cancer is the most damning form of the disease globally [3]. This section mostly provides a thematic summary of the contributions and attributes of the current breast cancer prediction techniques that have been made. Researchers have devised innumerable machine-learning classification strategies to predict breast cancer.

On the WBC dataset for the identification and diagnosis of breast cancer, Bazazeh et al. [4] analyze machine learning classifiers (SVM, RF, NB) and compare these classifiers with important characteristics similar to accuracy, precision, recall, and the ROC curve. The finding reveals that RF has the highest accuracy out of all of them when comparing the accuracy according to the classifiers SVM (96.6%), RF (99.9%), and NB (99.1%).

Chaurasiya et al. [5] scrutinize the accuracy values of four well-known ML classification models (LR, KNN, random forest tree (RDT, and SVM) while taking into account how well, each model performed on the WBCD dataset and among all the classifiers in this system, Random Forest Tree (RDT) achieved the greatest accuracy of 95%. Assegie [6] asserts a model for detecting breast cancer utilizing an improved KNN. To increase the model's accuracy in detecting breast cancer, conduct hyper-parameter tuning using a grid search to identify the best value of K, this method's accuracy was 94.35%, while the KNN default hyper-parameter value is 90.10%

Nurul et al. [7] examined the efficacy of several ML techniques to predict breast cancer survival. Furthermore, cross-validation of ten, five, three, and two-times procedures were used to attain the highest predictive performance on ML approaches, such as KNN, RF, SVM, and ensemble methods on WBCD datasets. AdaBoost ensemble approaches provided accuracy rates and cross-validation of 98.77% with 10 times, 98.41% with 2 times, and 98.24% with 3 times. SVM has the lowest error rate and the greatest accuracy rate at 98.60%, which is based on the results of 5-fold cross-validation.

Gupta et al. [8] advocate the application of deep learning (Adam Gradient Descent) and machine learning (DT, KNN, RF, LR, SVM) on malignant and benign cells on WBC datasets. Since deep learning combines the advantages of AdaGrad and RMSProp, which produces the most accurate results with the least amount of loss (98.24%). RMSProp performs well with nonstationary signals, while AdaGrad is ideally suited to computer vision issues.

The objectives of Ara et al. [9] is to analyse the WBC dataset, assess several classifiers for ml, and the effectiveness of breast cancer prediction using DT, SVM, K-NN, LR, RF, and NB. The finding shows an accuracy of 96.5%, RF and SVM perform better than other classifiers.

Amrane et al. [10] provide two distinct ML classifiers, which are Naive Bayes (NB) and k-nearest neighbor (KNN) on WBC and are two classifications that equate methods for breast cancer. Cross-validation is then used to assess the two significant and immediate outcomes and assess their correctness. In contrast to the NB classifier (96.19%), the findings show that KNN offers greater accuracy (97.51%) and a lower error rate.

The results of the extensive literature investigations are shown in Table 1 The reference numbers are displayed in column 1. The year appears in column 2. The datasets are given in column 3, the research algorithms employed are displayed in column 4, and finally, column 5 illustrates the efficiency of the algorithms used.

Table 1: Comparison of publicly available prediction models.

Ref. No.	***Period***	***Datasets***	***Algorithm***	***Accurateness (%)***
[21]	2022	WDBC and BCCD	SVM, LR, KNN and EC	99.3%, 98.06%, 97.35%, and 97.61% (WDBC)
[5]	2022	WDBC	KNN, SVM,	91.25%, 92.5%,

			LR and Random Forest Tree (RFT)	93.75% and 95%
[11]	2022	Regional OncologyCenter in Meknes, Morocco.	SVM, KNN, LR and NB	90.6%, 86.1%, 80.6% and 51.7%
[2]	2021	WDBC	LR, SVM, KNN, DT Classifier, RF Classifier and NB Classifier.	98.2%, 98.2%, 96.8% ,91.4%, 97.4% and 97.1%
[13]	2021	UCSB and BreakHis	c and ANN	89.1%, 85.2%, 82.4% and 86.27%
[14]	2020	WDBC	LR and DT	94.4% and 95.1%
[15]	2020	(WBC) and (WDBC)	NB, SVM, KNN and LR,	92% ,96%, 97% and 99% (WBC) and 96%, 94%, 96% and 98% (WDBC)
[16]	2020	WBC	NB, LR, and Neural Networks (NN)	95% training and 93% testing and 98% training and 97% testing
[28]	2019	WDBC	DT and KNN	92% and 95.95%
[17]	2019	WBCD	MLP, KNN, CART, Gaussian Naive Bayes (NB) and SVM	99.12%, 95.61%, 93.85% 94.73% and 98.24%
[18]	2019	WDBC	LR, NB and RF	95.61%, 96.49% and 97.36%
[10]	2018	WBC	NB and KNN	96.19% and 97.51%
[20]	2018	BCCD and WBCD	DT, SVM, RF, LR, NN and DT, SVM, RF, LR, NN	68.3%, 76.3%, 78.5%, 73.7%, 74.8% (BCCD) 96.3%, 97.7%, 98.9%, 98.1%, 98.5% (WBCD)
[19]	2017	BCD	NB and KNN	96.19% and 97.51%
[4]	2016	WBC	SVM, Bayesian Networks (BN), and RF	96.6%, 99.2%, and 99.9%

3. Methodology

To ascertain if the tumor is either cancerous (malignant) or harmless (benign), we have set up a series of methods to get the most trustworthy results and information for decision-making. The subsections can be used to present our general methodology: Dataset Description, Data Collection, Data Pre-processing, and Feature Selection.

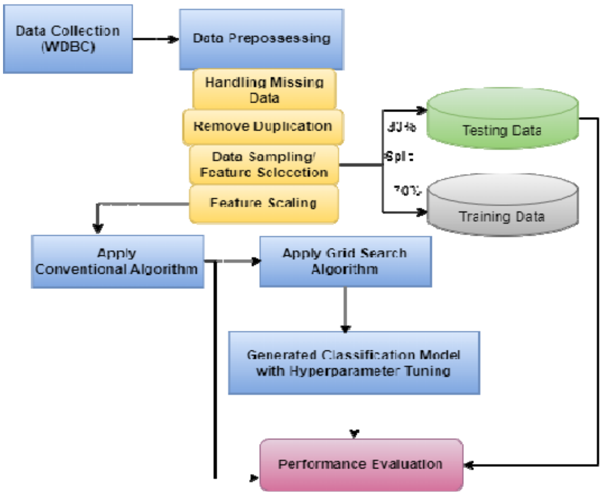


Fig. 1 Model for researched system.

In Fig. 1 The WDBC dataset was initially compiled. The data was then examined to determine if there were any duplicate or missing data. The data was separated into training and testing after being checked. The feature scaling was performed using standard scaling. Then, in order to assess and contrast their performances, we constructed both the traditional method and the hyper tuned parameter algorithm.

3.1 Dataset Description

The WDBC dataset has been generated by Dr. William H. Wolberg of the University of Wisconsin Hospital in Madison, Wisconsin, in the United States. It contains 32 columns, “ID” is the first and the second is the “diagnosis outcome” (0-benign and 1-malignant). The rest of the columns (3–32) contain 3 measurements (Mean, SD, and Worst-Case Mean) for each of the remaining 10 attributes. They exhibit more variability in the qualities of the size and form of the intended cancer cell's nucleus. In a biopsy test, a breast sample of cells is taken using the Fine Needle Aspiration (FNA) technique. In a pathology lab, each cell's nucleus is examined under a microscope to detect these traits. All feature values are maintained with a maximum of 4 meaningful digits. No null value was observed within the sample. The ten genuine qualities are given in Table 2.

Table 2: Description of WDBC dataset.

Feature Name	***Feature Description***
Radius	The average distance between the spots at the circumference's center and edges.
Texture	Grayscale value's SD. Perimeter Gross separation exists between the snake's points.
Perimeter	Gross separation exists at the snake's tip and between.
Area	Total amount of pixels inside the snake, plus one-half of each pixel outside its body.
Smoothness	Measured locally by computing the length difference, the variation in radius length.

3.2 Data Collection

The WDBC dataset was aggregated from Kaggle and is used to predict breast cancer; it has 569 instances with a total of 32 features.

3.3 Data Pre-processing

The WDBC dataset is checked before working with this data at first, and then the unnecessary features such as the id and unnamed column are extracted. Since variables like ID and nameless objects are redundant for predicting breast cancer, they have been removed from the dataset to improve the exploit and increase veracity.

3.4 Feature selection

Benign vs Malignant cells: There are 569 records in the dataset, 357 (62.7%) of which are Benign, and 212 (37.3%) are Malignant. The comparison of benign and malignant cells in this study data is shown in Fig. 2.

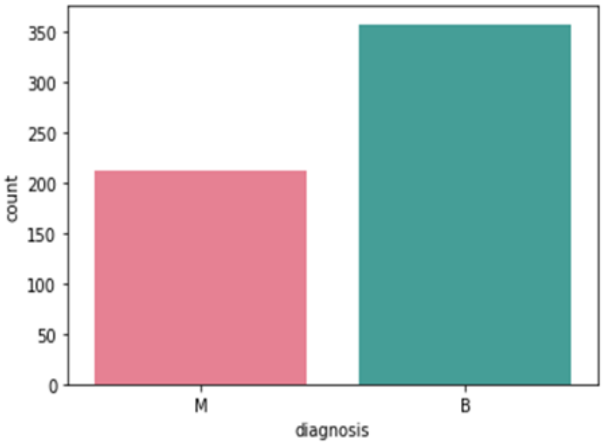


Fig. 2 Benign vs Malignant cell

3.5 Algorithm Used

In this section, we explored the WDBC dataset to determine which algorithm performs best with this small dataset. In this study, five of the most well-liked ML algorithms are used, but KNN and DT performed well on small datasets while RF, NB, and LR performed well on large datasets. The paramount goal is to benchmark each approach against one another and determine the most efficient and robust technique for the WDBC dataset.

K-Nearest Neighbor (KNN): The simplest technique used for classification is K-Nearest Neighbor. As this algorithm does not learn anything from its dataset and attributes [15]. During the training phase, this algorithm stores new data sets and classifies them into a well-suited category that is most similar to the available category [22]. KNN can be a suitable option for smaller datasets but may not be applicable for larger ones.

Decision Tree (DT): A supervised ML approach known as a decision tree is utilized for both classification and regression [23]. It looks like a tree structure according to its name for classifying different classes. This tree has three entities. One is decision nodes, which is used to make any decision by applying features of the dataset. The second one is brunches, which are used for any kind of decision rule. And the last one is the leaf node; it represents the output [2]. The output is taken by a yes/no question and answer. DT works well for the classification which has fewer class labels.

Random Forest (RF): Building numerous DTs on different subsets of the supplied dataset and taking the average to increase the prediction accuracy of the dataset at training time constitutes the Random Forest ensemble approach, [22] which is used for classification, regression, and other applications. [24] Random Forest is good for large datasets.

Naive Bayas (NB): This is one of the most well-known and straightforward classification algorithms for predictive modeling. It is also known as a probabilistic classifier that is used for quick prediction where one needs to make a prediction based on the probability of a particular task [22]. As this is a powerful algorithm, it works well on large datasets.

Logistic regression (LR): This is a machine learning method from the statistics world used for solving classification problems [3]. It mostly applies to binary classification problems and forecasts a binary dependent variable using a logistic function. This algorithm works well on very large datasets.

4. Experimental Result

In this section, we examined the effectiveness of the dataset after constructing the ML algorithms. This is accomplished by running the algorithms on the test dataset that was previously established. The test dataset contained 30% of the total dataset. To determine the accuracy, precision, recall, and F1 score for each method utilized, a confusion matrix made up of TP, FP, TN, and FN is constructed for the actual and predicted results. The interpretation of the terms is listed below.

- TP: True Positive (Correctly Identified)
- FP: False Positive (Correctly Rejected)
- TN: True Negative (Incorrectly Identified)
- FN: False Negative (Incorrectly Rejected)

4.1 Accuracy

Accuracy tells you how many times the ML model was correct overall. It is determined as the sum of all the data set's occurrences divided by the number of precise forecasts. It is important to note that the accuracy varies for various testing sets depending on the classifier's threshold selection. For calculating accuracy, use the formula (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \qquad (1)$$

4.2 Precision

Precision is how good the model is at predicting a specific category. Utilizing the proportion of all expected positives to actual positives, the mathematical formula is shown in equation (2).

$$Precision = \frac{TP}{TP + FP} \times 100\% \qquad (2)$$

4.3 Recall

Recall refers to the number of correctly predicted data that were recognized (found), i.e., the number of perfect finds that were also identified. The mathematical formula is shown in equation (3).

$$Recall = \frac{TP}{TP + FN} \times 100\% \qquad (3)$$

4.4 F1 Score

This refers to the merging variables that would normally be in opposition, recall, and precision. This simply summarizes the prediction capability of a model. The mathematical formula is shown in equation (4).

$$F1\ score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \qquad (4)$$

Performance	***Hyperparameter Tuning***	***DT***	***RF***	***KNN***	***NB***	***LR***
Accuracy	With	94.74%	97.08%	98.83%	95.91%	97.08%
	Without	94.15	97.0	96.4	95.9	96.49

		%	8%	9%	1%	%
Precision	With	95%	97%	99%	96%	97%
	Without	94%	97%	96%	96%	97%
Recall	With	94%	97%	99%	96%	96%
	Without	94%	97%	96%	96%	97%
F1 Score	With	94%	97%	99%	96%	97%
	Without	94%	97%	96%	96%	97%

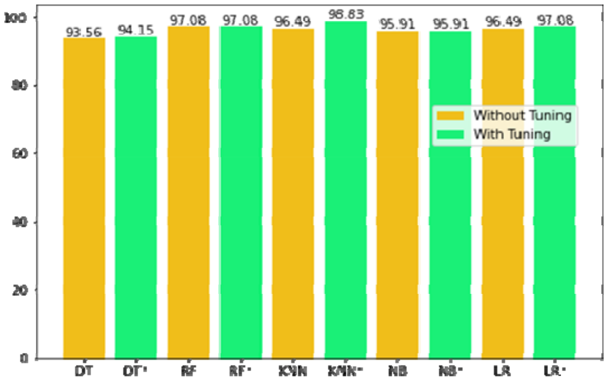


Fig. 3 Result Analysis

The results shown in Table 3 demonstrate that the KNN classifier performs well on this study (hyper tuning) according to accuracy, precision, recall and F1 score. Based on the findings, the KNN model is the most accurate classifier among the five suggested classifiers for predicting breast cancer. According to this Fig. 3 shows a graphical representation for better understanding.

Table 4: Result comparison with existing work.

Study using WDBC	*Accuracy of KNN Classifier*
Our paper	98.83%
Rasool et al. [21], 2022	97.35%
Roy et al. [2], 2021	96.8%
Roy et al. [15], 2020	96%
Rajaguru et al. [28], 2019	95.95%

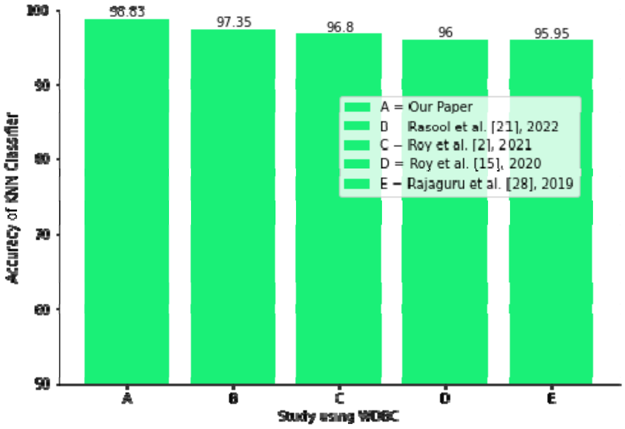


Fig. 4 Result comparison with existing work.

Table 4 compares the effects of the study model, hyperparameter tuning BC prediction using the WDBC only with the accuracy of KNN. Finally, we draw the conclusion that the suggested method surpasses all other approaches mentioned in the literature by comparing the results of KNN with other state-of-the-art studies in Table 4. According to this Fig. 4 shows a graphical representation for better understanding.

5. Conclusion and Future Work

The leading cause of mortality in women is breast cancer. This study integrated a postulated method for forecasting breast cancer. There are five different ML classifiers using WDBC dataset with LR, DT, RF, KNN, and NB to produce the breast cancer prognostic model. When it comes to tuning hyperparameters using grid search, the study is isolated from the conventional system. While the accuracy rates of the DT, RF, KNN, NB, and LR classifiers without hyperparameter adjustment are 93.56%, 97.08%, 96.49%, 95.91%, and 96.49%, respectively. However, the DT, RF, KNN, NB and LR classifiers in the improved set take the accuracy rate of 94.15%, 97.08%, 98.83%, 95.91% and 97.08% using the hyperparameters tuning approach. We compared the classifiers and discovered that KNN provides the highest accuracy (98.83%) and works well with the study approach.

By expanding the data size in the future, this accuracy can be robustically enhanced and also more work can be carried out not only in cancer prediction but also in detecting the stage of a cancer patient.

Acknowledgments

We would like to use this chance to express our gratitude to the Center for Artificial Intelligence and Robotics

(CAIR).

References

[1] Knowledge, Attitude and Practice of Bangladeshi Women towards Breast Cancer: A Cross Sectional Study. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30755557/

[2] S. Roy, R. Gawande, A. Nawghare, and S. Mistry, “Comparative Study of Machine Learning Algorithms for Detecting Breast Cancer”, International Journal of Computer Science Trends and Technology (IJCST), Vol. 9, No. 3, pp. 103-111, 2021.

[3] K. Hashi, E. and M.S. Zaman, “Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction”, Journal of Applied Science & Process Engineering, Vol. 7, No. 2, pp. 631-647, 2020.

[4] D. Bazazeh, and R. Shubair, “Comparative study of machine learning algorithms for breast cancer detection and diagnosis”, 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), pp. 1–4, 2016.

[5] S. Chaurasiya, and R. Rajak, “Comparative Analysis of Machine Learning Algorithms in Breast Cancer Classification”, pp. 1–12, 2022.

[6] T.A. Assegie, “An optimized K-Nearest Neighbor based breast cancer detection” Journal of Robotics and Control (JRC), Vol. 2, No. 3, pp. 115-118, 2021.

[7] N. Mashudi, S. Rossli, N.Ahmad, and N.M. Noor, “Comparison on Some Machine Learning Techniques in Breast Cancer Classification” 2020 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES), pp. 499-504, 2020.

[8] P. Guptaa, and S. Garga, “Breast Cancer Prediction using varying Parameters of Machine Learning Models” Third International Conference on Computing and Network Communications (CoCoNet’19), Vol. 171, pp. 593-601, 2020.

[9] S. Ara, A. Das, and A. Dey, “Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms”, 2021 International Conference on Artificial Intelligence (ICAI), pp. 97-101, 2021.

[10] M. Amrane, S. Oukid, I. Gagaoua, and T. ENSAR, “Breast Cancer Classification Using Machine Learning”, 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), pp. 1-4, 2018.

[11] E. Merouane, A. Said, and E.F. Nour-eddine, “Prediction of Metastatic Relapse in Breast Cancer using Machine Learning Classifiers”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 13, No. 2, pp. 176-181, 2022.

[12] M. Amrane, S. Oukid, I. Gagaoua, and T. ENSAR, “Breast Cancer Classification Using Machine Learning”, 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), pp. 1-4, 2018.

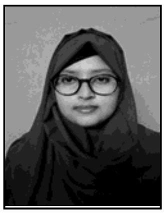
[13] M.A. Aswathy, and M. Jagannath, “An SVM approach towards breast cancer classification from H&E-stained histopathology images based on integrated features”, International Federation for Medical and Biological Engineering, pp. 1-11, 2021.

[14] P. Sengar, M. Gaikwad, and A. Nagdive, “Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction” 2020 Third International Conference on Smart Systems and Inventive Technology (ICCSIT), pp. 796-801, 2020.

[15] B.R. Roy, M.; Pal, S. Das, and A. Huq, “Comparative Study of Machine Learning Approaches on Diagnosing Breast Cancer for Two Different Dataset”, 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), pp. 29-34, 2020.

Md. Mijanur Rahman, Assistant Professor, Computer Science, and Engineering and Director, of Central for Artificial Intelligence and Robotics at Southeast University, and his Research area is Deep Learning, Machine Learning, and Blockchain.

Ashikur Rahman Raju was born on May 21, 1999 and raised in Dinajpur, Bangladesh. He is currently enrolled in Southeast University in Dhaka, Bangladesh, where he is pursuing a B.S.C. in computer science and engineering (CSE). This is his final semester before graduating (2023). Machine learning, deep learning, Data Science, and its use in biomedical and health informatics are some of his areas of study interest.



Sumiea Akter was born and raised in Dhaka, Bangladesh on 24 August, 1999. Currently she is enrolled in Southeast University in Dhaka, Bangladesh, where she is pursuing a B.S.C. in computer science and engineering (CSE). This is her final semester before graduating (2023). Machine learning, deep learning, and its use in biomedical and health informatics are some of his areas of study interest.

Swarnali Akter born and brought up in Dhaka, Bangladesh on 4 April, 1998. Currently she is enrolled in Southeast University in Dhaka, Bangladesh, where she is pursuing a B.S.C. in computer science and engineering (CSE). This is her final semester before graduating (2023). Machine Learning, Deep Learning, IOT and its use in biomedical and health informatics are some of his areas of study interest.