Journal of Computer and Communications, 2023, 11, 149-165
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Hyperparameter Tuning Based Machine Learning Classifier for Breast Cancer Prediction

Md. Mijanur Rahman, Asikur Rahman, Swarnali Akter, Sumiea Akter Pinky

Department of CSE, Southeast University, Dhaka, Bangladesh
Email: mijanur.rahman@seu.edu.bd, 2018000000041@seu.edu.bd, 2018000000046@seu.edu.bd, 2018000000045@seu.edu.bd

How to cite this paper: Rahman, M.M., Rahman, A., Akter, S. and Pinky, S.A. (2023) Hyperparameter Tuning Based Machine Learning Classifier for Breast Cancer Prediction. *Journal of Computer and Communications*, **11**, 149-165.
https://doi.org/10.4236/jcc.2023.114007

Received: March 2, 2023
Accepted: April 25, 2023
Published: April 28, 2023

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 favourable 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, F1 score and finally the AUC & ROC curve. In this study of 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 6808 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 if 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.

For the purpose of predicting breast cancer, Roy *et al.* employed the WDBC Dataset in ML (LR, K-NN, SVM, NB, DT, and RF). Support vector machines and logistic regression are the most efficient algorithms we've looked at so far. SVM and LR have been shown to be the most accurate algorithms, with LR and SVM both scoring 98.245% accuracy [2]. Indeed, this study has the potential to use a new methodology and dataset to increase their performance.

According to Chaurasiya *et al.* [3] analysis of the accuracy ratings on the WDBC dataset of four popular ML classification models (LR, KNN, random forest tree (RDT), and SVM), Random Forest Tree (RDT) has the highest accuracy of 95% out of all the classifiers. To make a more conclusive generalization and further lower the incidence of misclassification, this study's shortcomings are its lack of use of other classification algorithms on various and comparably extensive data sets.

In this investigation, Kim *et al.* [4] presented a simple to use machine learning prediction tool for pathological Complete Response (pCR) in breast cancer survivors medicated with Neoadjuvant Chemotherapy (NAC) and generated their training set by using Two-class Bayes point machine technique. They made use of information from clinical traits and gene xpression patterns. The accuracy was 0.875 and the AUC of the ROC curve was 0.909 in this gene-based prediction model. The AUC of the ROC curve and accuracy were both 0.800 in a different model absent gene data. The first drawback of this study is the small number of patients who were recruited for it. A second limitation is that only internal validation has been conducted.

According to a literature assessment of approaches employed by numerous researchers [2]-[14] 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 through a different method, data preprocessing and so on. Since this illness is extremely detrimental to every patient and is becoming more and more prevalent. Therefore, if the accuracy rate was raised to a higher level, it would aid healthcare professionals in predicting breast cancer early on before it becomes fatal.

This study's axiom is to apply 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 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 is presented, the acknowledgment and references are displayed in the rest of the paper.

2. Related Work

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 [15]. 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.* [5] 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.* [3] 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 WDBC 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 analyze 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 (%)
[16]	2022	WDBC and BCCD	SVM, LR, KNN and EC	99.3%, 98.06%, 97.35%, and 97.61%

Continued

[3]	2022	WDBC	KNN, SVM, LR and Random Forest Tree (RFT)	91.25%, 92.5%, 93.75% and 95%
[17]	2022	Regional Oncology Center 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%
[18]	2021	UCSB and BreakHis	c and ANN	89.1% and 86.27%
[19]	2020	WDBC	LR and DT	94.4% and 95.1%
[14]	2020	(WBC) and (WDBC)	NB, SVM, KNN and LR,	92%, 96%, 97% and 99% (WBC) and 96%, 94%, 96% and 98% (WDBC)
[12]	2020	WBC	NB, LR, and Neural Networks (NN)	95% training and 93% testing and 98% training and 97% testing
[20]	2019	WDBC	DT and KNN	92% and 95.95%
[13]	2019	WBCD	MLP, KNN, CART, Gaussian Naive Bayes (NB) and SVM	99.12%, 95.61%, 93.85% 94.73% and 98.24%
[21]	2019	WDBC	Kernel SVM, LR KNN, DT, NB and RF	98.24%, 96.49%,95.61%,88.59%,85.09% and 92.98%
[10]	2018	WBC	NB and KNN	96.19% and 97.51%
[14]	2018	BCCD and WBCD	DT, SVM, RF, LR, NN 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)
[22]	2017	BCD	NB and KNN	96.19% and 97.51%
[5]	2016	WBC	SVM, Bayesian Networks (BN), and RF	96.6%, 99.2%, and 99.9%
[23]	2013	WDBC	K-SVM (Hybrid), ACO-SVM, GA-SVM and PSO-SVM	97.38%, 95.96%, 97.19% and 97.37%

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.

In Figure 1, the WDBC dataset was initially compiled. The data was then examined to determine if there were any duplicates or missing data. Handling missing data was omitted since no missing data was discovered. The data was separated into training (70%) and testing (30%) after being checked. The feature scaling was performed using standard scaling. Then, in order to assess and contrast the performances, we constructed both the traditional method and the hyper-tuned parameter algorithm.

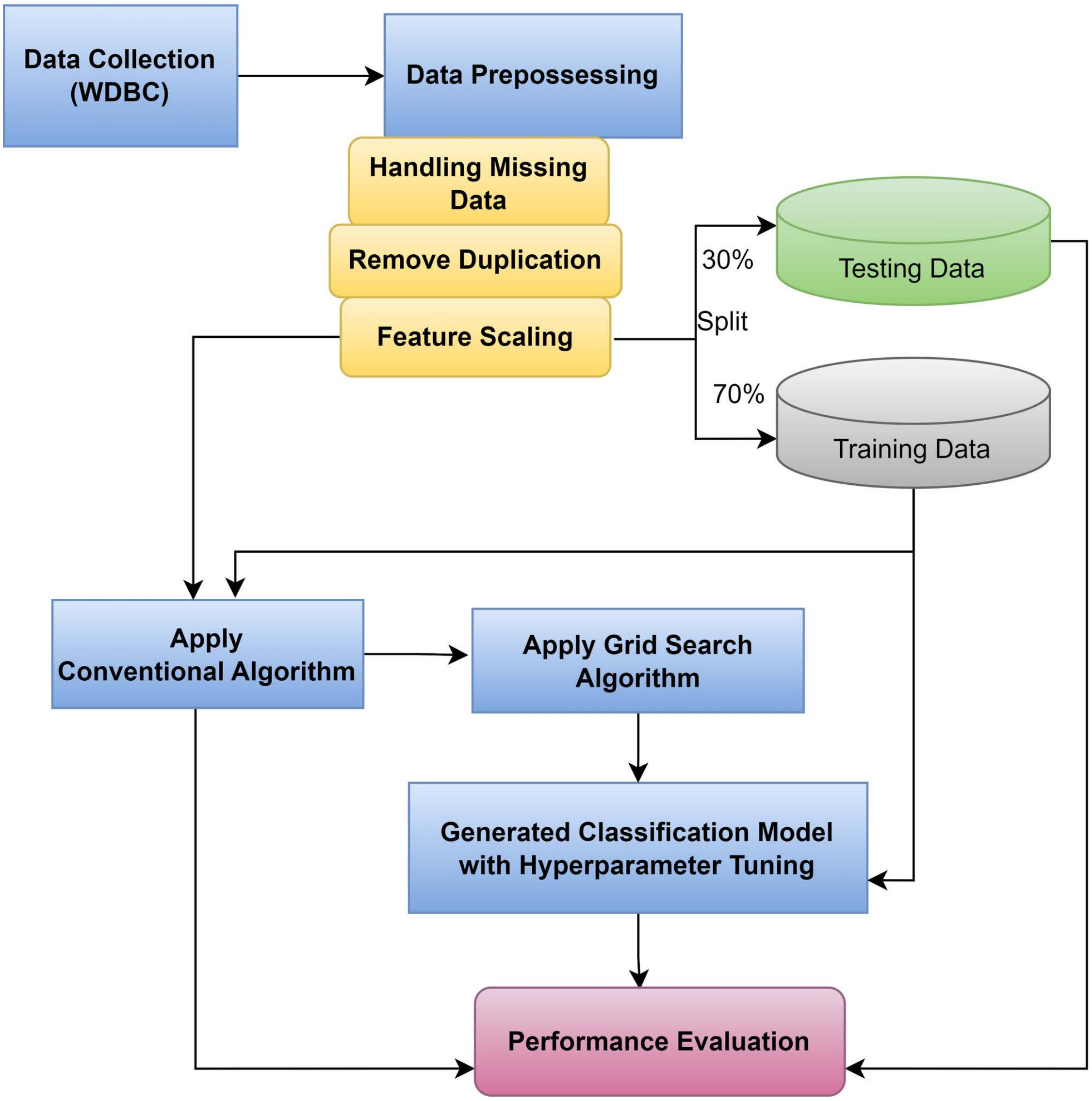

Figure 1. Model for research system.

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.

id	diagnosis	radius_ mean	texture_ mean	perimeter_ mean	area_ mean	smoothness_ mean
906564	B	14.69	13.98	98.22	656.1	0.10310
85715	M	13.17	18.66	85.98	534.6	0.11580
891670	B	12.95	16.02	83.14	513.7	0.10050
874217	M	18.31	18.58	118.60	1041.0	0.08588
905680	M	15.13	29.81	96.71	719.5	0.08320

compactness_ mean	concavity_ mean	concave points_mean	symmetry_ mean	fractal_dimension_ mean
0.18360	0.14500	0.06300	0.2086	0.07406
0.12310	0.12260	0.07340	0.2128	0.06777
0.07943	0.06155	0.03370	0.1730	0.06470
0.08468	0.08169	0.05814	0.1621	0.05425
0.04605	0.04686	0.02739	0.1852	0.05294

radius_se	texture_se	perimeter_se	area_se	smoothness_se	compactness_se
0.5462	1.5110	4.795	49.45	0.009976	0.052440
0.2871	0.8937	1.897	24.25	0.006532	0.023360
0.2094	0.7636	1.231	17.67	0.008725	0.020030
0.2577	0.4757	1.817	28.92	0.002866	0.009181
0.4681	1.6270	3.043	45.38	0.006831	0.014270

concavity_se	concave points_se	symmetry_se	fractal_ dimension_se	radius_worst	texture_worst
0.05278	0.015800	0.02653	0.005444	16.46	18.34

Continued

0.02905	0.012150	0.01743	0.003643	15.67	27.95
0.02335	0.011320	0.02625	0.004726	13.74	19.93
0.01412	0.006719	0.01069	0.001087	21.31	26.36
0.02489	0.009087	0.03151	0.001750	17.26	36.91

perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst
114.10	809.2	0.1312	0.36350	0.3219	0.11080
102.80	759.4	0.1786	0.41660	0.5006	0.20880
88.81	585.4	0.1483	0.20680	0.2241	0.10560
139.20	1410.0	0.1234	0.24450	0.3538	0.15710
110.10	931.4	0.1148	0.09866	0.1547	0.06575

symmetry_worst	fractal_dimension_worst
0.2827	0.09208
0.3900	0.11790
0.3380	0.09584
0.3206	0.06938
0.3233	0.06165

3.2. Dataset 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. Here is a sample.

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. The feature scaling was performed using standard scaling.

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 **Figure 2**. We chose not to utilize a particular feature selection technique in this case since we obtained good results when compared to other feature selection strategies, such as correlation coefficient, and because the data in question pertains to medicine.

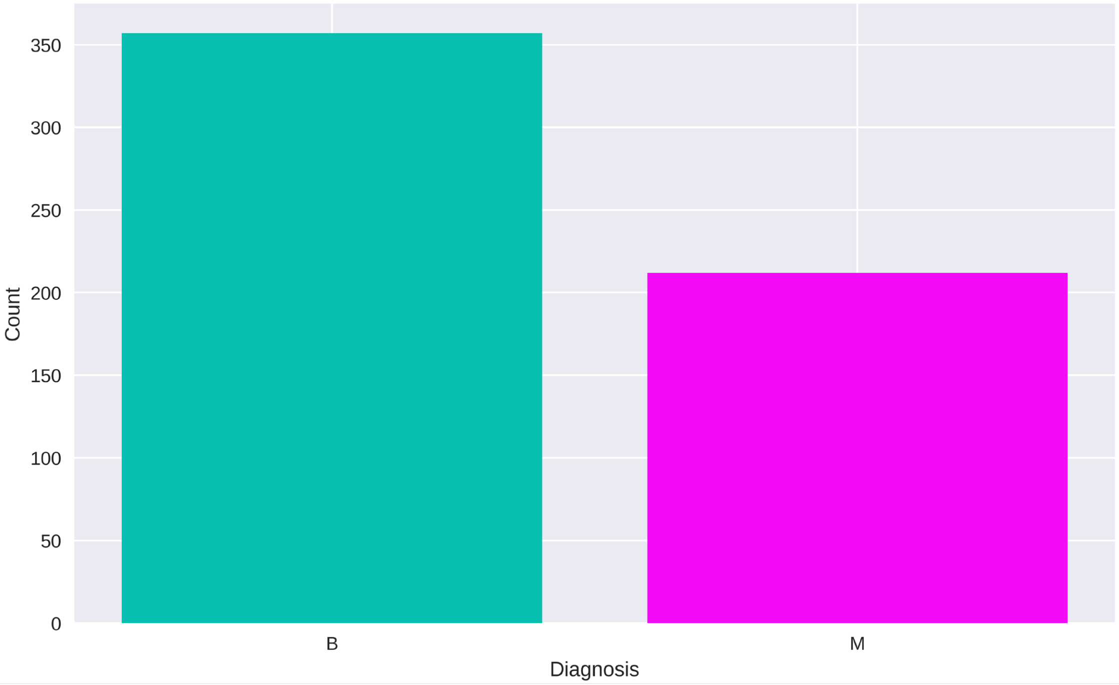

Figure 2. Benign vs malignant cells.

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 [11]. 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 [24]. 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 [25]. 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, [26] which is used for classification, regression, and other applications. 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 [24]. 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 [15]. 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 Results

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, F1 score and AUC & ROC curve for each method utilized, a confusion matrix (**Figure 3**) made up of TP, FP, TN, and FN is constructed for the actual and predicted results. The interpretation of the terms is listed below.

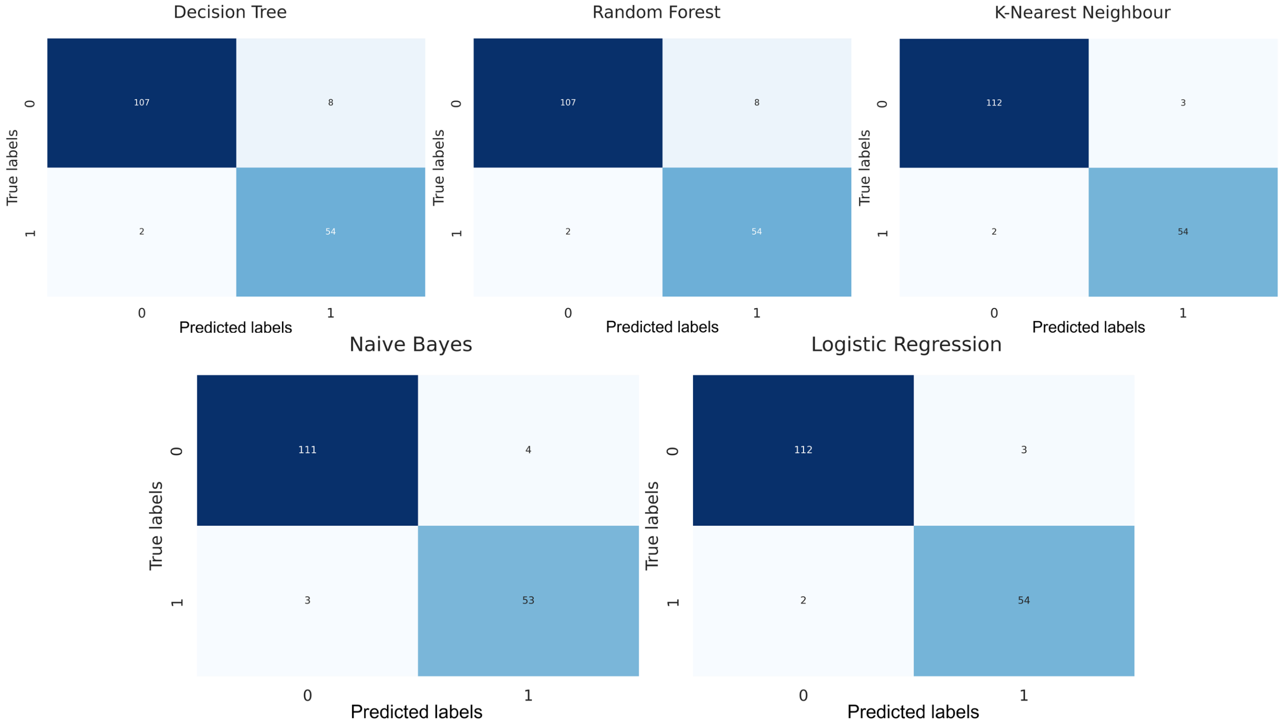

Figure 3. Confusion matrix after tuning.

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).

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FN}} \times 100 \tag{1}$$

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).

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \times 100 \tag{2}$$

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).

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100 \tag{3}$$

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).

$$\text{F1 Score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall}+\text{Precision}} \tag{4}$$

4.5. AUC & ROC Curve

The ROC curve is a graphical representation of the True Positive Rate (TPR) plotted against the False Positive Rate (FPR) for different threshold values of the model's predicted probabilities. AUC is a metric that quantifies the area under the ROC curve. It has a value ranging from 0 to 1, where 0.5 indicates a random classifier, and 1 represents a perfect classifier. The performance of the tuned model is illustrated in **Figure 4** using the AUC and ROC curve.

The results shown in **Table 3** & **Table 4** 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 Figure 5 shows a graphical representation for better understanding.

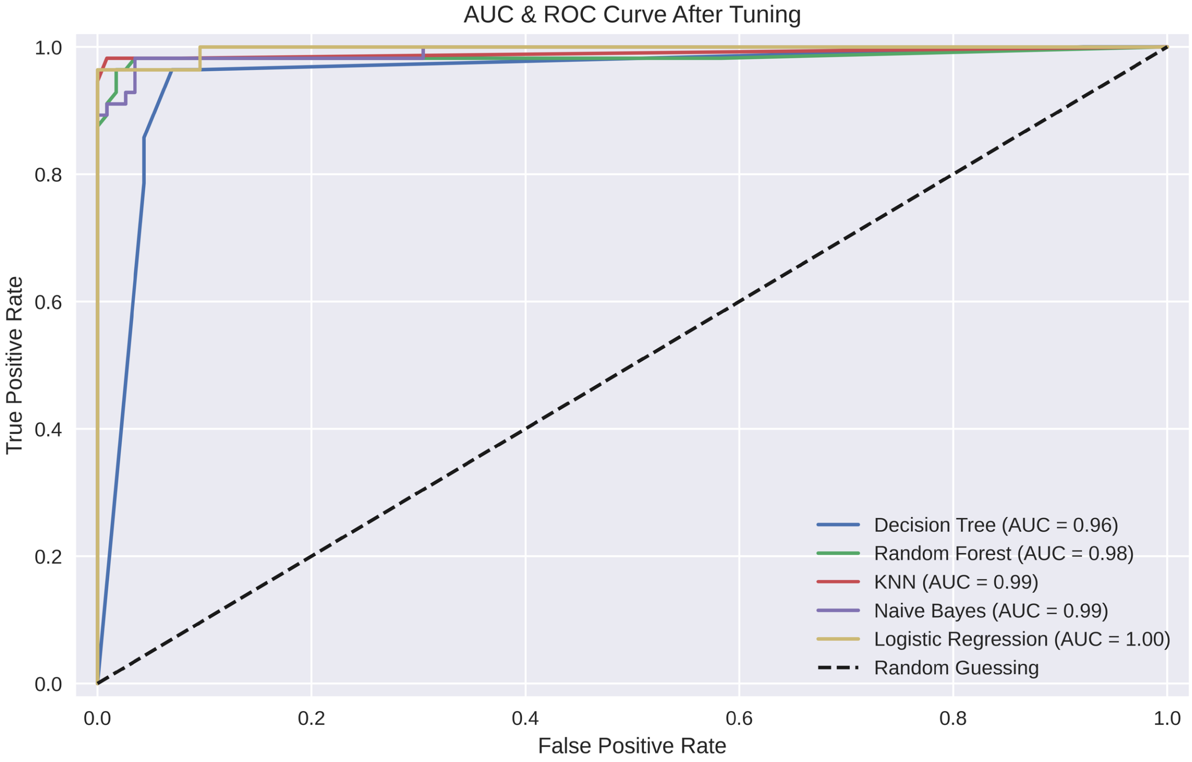

Figure 4. AUC and ROC curve after tuning.

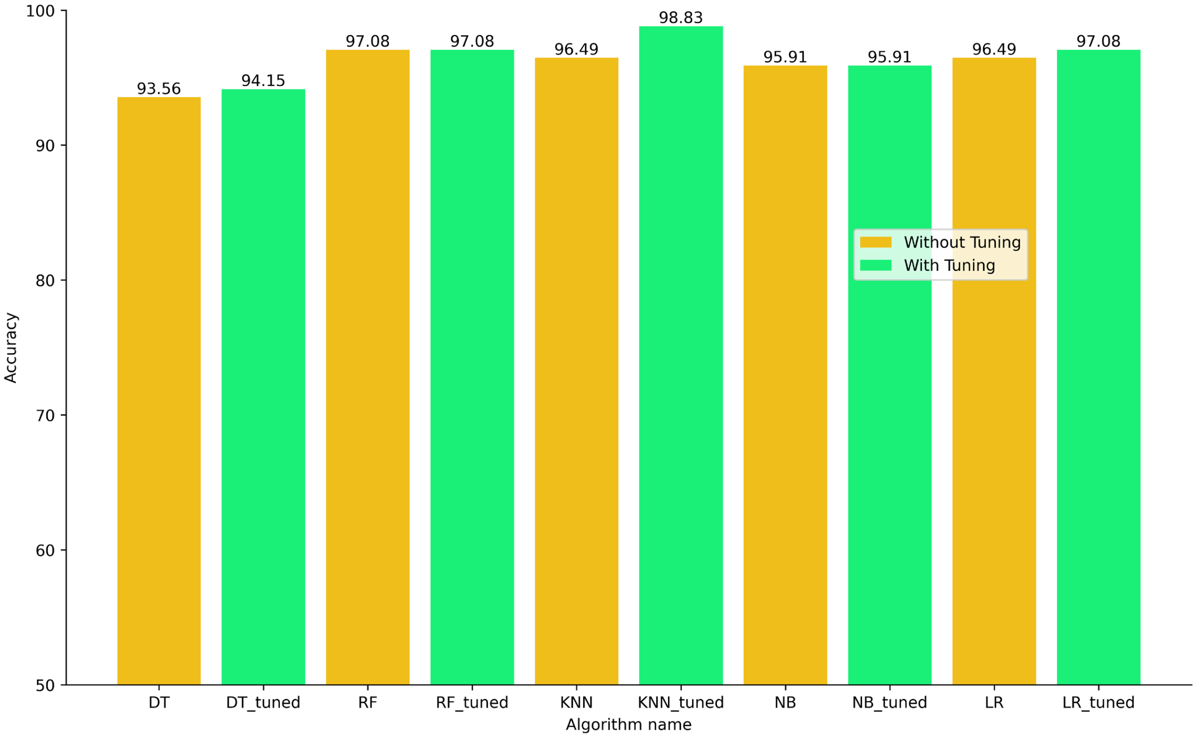

Figure 5. Result analysis on accuracy.

Table 3. Performance evaluation without hyperparameter tuning.

Algorithm Names	Accuracy	Precisions	Recall	F1 Score
Decision Tree	93.56%	94%	94%	94%
Random Forest	97.08%	97%	97%	97%
K Nearest Neighbour	96.49%	96%	96%	96%
Naive Bayes	95.91%	96%	96%	96%
Logistic Regression	96.49%	96%	96%	96%

Table 4. Performance evaluation with hyperparameter tuning.

Algorithm Names	Accuracy	Precisions	Recall	F1 Score
Decision Tree	94.15%	95%	94%	94%
Random Forest	97.08%	97%	97%	97%
K Nearest Neighbour	98.83%	99%	99%	99%
Naive Bayes	95.91%	96%	96%	96%
Logistic Regression	97.08%	97%	97%	97%

Table 5 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 5. According to this Figure 6 shows a graphical representation for better understanding.

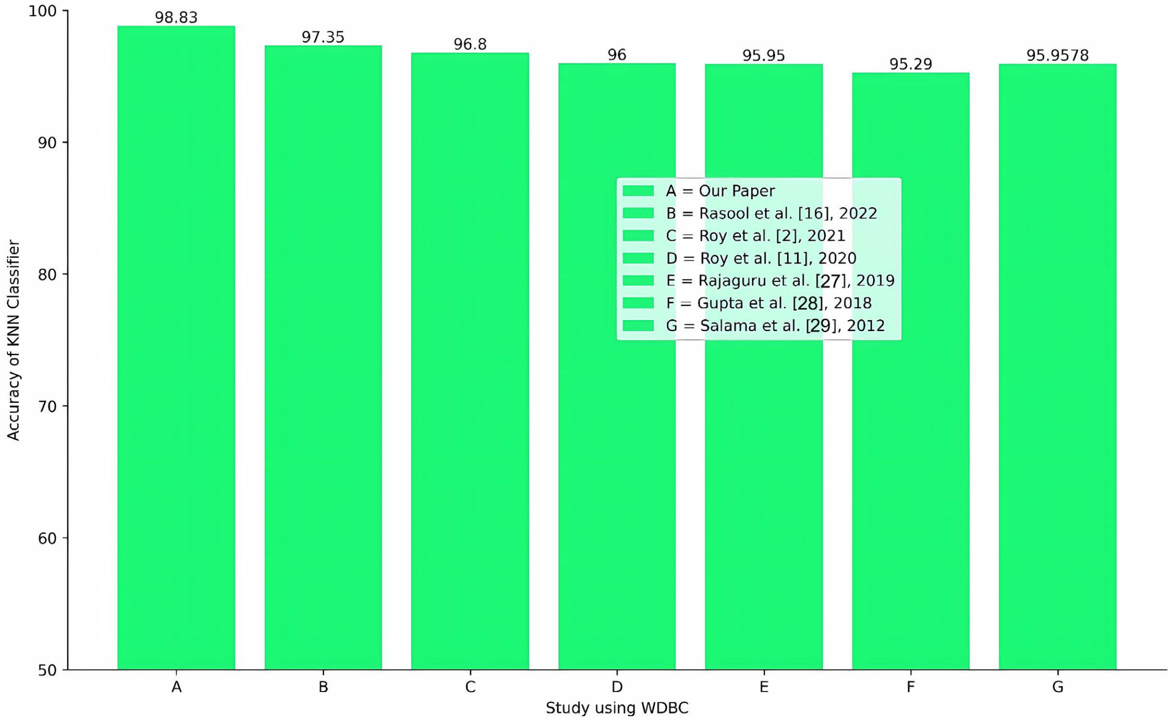

Figure 6. Result comparison with existing work.

Table 5. Result comparison with existing work.

Study using WDBC	Accuracy of KNN Classifier
Our paper	98.83%
Rasool *et al.* [16], 2022	97.35%
Roy *et al.* [2], 2021	96.8%
Roy *et al.* [11], 2020	96%
Rajaguru*et al.* [27], 2019	95.95%
Gupta *et al.* [28], 2018	95.29%
Salama *et al.* [29], 2012	95.9578% (IBK)

5. Conclusion

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.

Acknowledgements

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

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Begum, S.A., Mahmud, T., Rahman, T., Zannat, J., Khatun, F., Nahar, K., Towhida, M., Joarder, M., Harun, A. and Sharmin, F. (2019) Knowledge, Attitude and Practice of Bangladeshi Women towards Breast Cancer: A Cross Sectional Study. *Mymensingh Medical Journal*, **28**, 96-104.
https://pubmed.ncbi.nlm.nih.gov/30755557/

[2] Roy, S., Gawande, R., Nawghare, A. and Mistry, S. (2021) Comparative Study of Machine Learning Algorithms for Detecting Breast Cancer. *International Journal of Computer Science Trends and Technology* (*IJCST*), **9**, 103-111.

[3] Chaurasiya, S. and Rajak, R. (2022) Comparative Analysis of Machine Learning Algorithms in Breast Cancer Classification. (Preprint)
https://doi.org/10.21203/rs.3.rs-1772158/v1

[4] Kim, I., Lee, K., Lee, S., Park, Y. and Lee, K. (2021) A Predictive Model for Pathologic Complete Response in Breast Cancer Patients Treated with Neoadjuvant Chemotherapy Using Machine Learning. *Advances in Breast Cancer Research*, **10**, 141-155.
https://doi.org/10.4236/abcr.2021.104012
https://www.scirp.org/journal/paperinformation.aspx?paperid=111495

[5] Bazazeh, D. and Shubair, R. (2016) Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis. *Proceedings of* 2016 5*th International Conference on Electronic Devices*, *Systems and Applications* (*ICEDSA*), Ras Al Khaimah, 6-8 December 2016, 1-4.
https://doi.org/10.1109/ICEDSA.2016.7818560
https://ieeexplore.ieee.org/abstract/document/7818560

[6] Assegie, T.A. (2021) An Optimized K-Nearest Neighbor Based Breast Cancer Detection. *Journal of Robotics and Control* (*JRC*), **2**, 115-118.
https://doi.org/10.18196/jrc.2363

[7] Mashudi, N.A., Rossli, S.A., Ahmad, N. and Noor, N.M. (2021) Comparison on Some Machine Learning Techniques in Breast Cancer Classification. *Proceedings of* 2020 *IEEE-EMBS Conference on Biomedical Engineering and Sciences* (*IECBES*), Langkawi Island, 1-3 March 2021, 499-504.
https://ieeexplore.ieee.org/abstract/document/9398837
https://doi.org/10.1109/IECBES48179.2021.9398837

[8] Gupta, P. and Garg, S. (2020) Breast Cancer Prediction Using Varying Parameters of Machine Learning Models. *Procedia Computer Science*, **171**, 593-601.
https://doi.org/10.1016/j.procs.2020.04.064

[9] Ara, S., Das, A. and Dey, A. (2021) Malignant and Benign Breast Cancer Classification Using Machine Learning Algorithms. *Proceedings of* 2021 *International Conference on Artificial Intelligence* (*ICAI*), Islamabad, 5-7 April 2021, 97-101.
https://ieeexplore.ieee.org/abstract/document/9445249
https://doi.org/10.1109/ICAI52203.2021.9445249

[10] Amrane, M., Oukid, S., Gagaoua, I. and Ensari, T. (2018) Breast Cancer Classification Using Machine Learning. *Proceedings of* 2018 *Electric Electronics*, *Computer Science, Biomedical Engineerings*' *Meeting* (*EBBT*), Istanbul, 18-19 April 2018, 1-4.
https://ieeexplore.ieee.org/abstract/document/8391453
https://doi.org/10.1109/EBBT.2018.8391453

[11] Roy, B.R., Pal, M., Das, S. and Huq, A. (2020) Comparative Study of Machine Learning Approaches on Diagnosing Breast Cancer for Two Different Dataset. *Proceedings of* 2020 2*nd International Conference on Advanced Information and Communication Technology* (*ICAICT*), Dhaka, 28-29 November 2020, 29-34.
https://ieeexplore.ieee.org/abstract/document/9333507
https://doi.org/10.1109/ICAICT51780.2020.9333507

[12] El-Shair, Z.A., Sánchez-Pérez, L.A. and Rawashdeh, S.A. (2020) Comparative Study of Machine Learning Algorithms Using a Breast Cancer Dataset. *Proceedings of* 2020 *IEEE International Conference on Electro Information Technology* (*EIT*), Chicago, 31 July 2020-1 August 2020, 500-508.
https://ieeexplore.ieee.org/abstract/document/9208315
https://doi.org/10.1109/EIT48999.2020.9208315

[13] Bataineh, A.A. (2019) A Comparative Analysis of Nonlinear Machine Learning Algorithms for Breast Cancer Detection. *International Journal of Machine Learning*

and Computing, **9**, 248-254. https://doi.org/10.18178/ijmlc.2019.9.3.794

[14] Li, Y. and Chen, Z. (2018) Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction. *Applied and Computational Mathematics*, **7**, 212-216. https://doi.org/10.11648/j.acm.20180704.15

[15] Hashi, E.K. and Zaman, M.S.U. (2020) Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction. *Journal of Applied Science & Process Engineering*, **7**, 631-647. https://publisher.unimas.my/ojs/index.php/JASPE/article/view/2639 https://doi.org/10.33736/jaspe.2639.2020

[16] Rasool, A., Bunterngchit, C., Tiejian, L., Islam, M.R., Qu, Q. and Jiang, Q. (2022) Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis. *International Journal of Environmental Research and Public Health*, **19**, Article 3211. https://www.mdpi.com/1660-4601/19/6/3211 https://doi.org/10.3390/ijerph19063211

[17] Merouane, E. and Said, A. (2022) Prediction of Metastatic Relapse in Breast Cancer using Machine Learning Classifiers. *International Journal of Advanced Computer Science and Applications*, **13**, 176-181. https://doi.org/10.14569/IJACSA.2022.0130222

[18] Aswathy, M.A. and Jagannath, M. (2021) An SVM Approach towards Breast Cancer Classification from H&E-Stained Histopathology Images Based on Integrated Features. *Medical & Biological Engineering & Computing*, **59**, 1773-1783. https://link.springer.com/article/10.1007/s11517-021-02403-0 https://doi.org/10.1007/s11517-021-02403-0

[19] Sengar, P.P., Gaikwad, M.J. and Nagdive, A.S. (2020) Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction. *Proceedings of* 2020 *Third International Conference on Smart Systems and Inventive Technology* (*ICSSIT*), Tirunelveli, 20-22 August 2020, 796-801. https://doi.org/10.1109/ICSSIT48917.2020.9214267 https://ieeexplore.ieee.org/abstract/document/9214267

[20] Gayathri, B.M. and Sumathi, C.P. (2016) Comparative Study of Relevance Vector Machine with Various Machine Learning Techniques Used for Detecting Breast Cancer. *Proceedings of* 2016 *IEEE International Conference on Computational Intelligence and Computing Research* (*ICCIC*), Chennai, 15-17 December 2016, 1-5. https://ieeexplore.ieee.org/abstract/document/7919576 https://doi.org/10.1109/ICCIC.2016.7919576

[21] Kumar, A. and Poonkodi, M. (2019) Comparative Study of Different Machine Learning Models for Breast Cancer Diagnosis. In: Chattopadhyay, J., Singh, R. and Bhattacherjee, V., Eds., *Innovations in Soft Computing and Information Technology*, Springer, The Gateway, 17-25. https://doi.org/10.1007/978-981-13-3185-5_3

[22] Sharma, A., Kulshrestha, S. and Daniel, S. (2017) Machine Learning Approaches for Breast Cancer Diagnosis and Prognosis. *Proceedings of* 2017 *International Conference on Soft Computing and Its Engineering Applications* (*icSoftComp*), Changa, 1-2 December 2017, 1-5. https://ieeexplore.ieee.org/abstract/document/8280082 https://doi.org/10.1109/ICSOFTCOMP.2017.8280082

[23] Zheng, B., Yoon, S.W. and Lam, S.S. (2013) Breast Cancer Diagnosis Based on Feature Extraction Using a Hybrid of K-Means and Support Vector Machine Algorithms. *Expert Systems with Applications*, **41**, 1476-1482. https://doi.org/10.1016/j.eswa.2013.08.044

[24] Javapoint (2023) K-Nearest Neighbor (KNN) Algorithm for Machine Learning.

https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

[25] XORIANT (2023) Decision Trees for Classification: A Machine Learning Algorithm. https://www.xoriant.com/blog/decision-trees-for-classification-a-machine-learning-algorithm/

[26] Wikipedia (2023) Random Forest. https://en.wikipedia.org/wiki/Random_forest

[27] Rajaguru, H. and Chakravarthy, S.R. (2019) Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer. *Asian Pacific Journal of Cancer Prevention*, **20**, 3777-3781.

[28] Gupta, M. and Gupta, B. (2018) A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques. *Proceedings of* 2018 *Second International Conference on Computing Methodologies and Communication* (*ICCMC*), Erode, 15-16 February 2018, 997-1002. https://doi.org/10.1109/ICCMC.2018.8487537

[29] Salama, G.I., Abdelhalim, M.B. and Zeid, M.A. (2012) Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers. *International Journal of Computer and Information Technology*, **1**, 36-43.