A Comparative Study of Breast Cancer Detection using Machine Learning Algorithms.

Md. Taslim Hosen Shanto[1]
Department of Computer Science and Engineering
Jashore University of Science and Technology
Jashore 7408, Bangladesh
taslim.cse06@gmail.com

Md. Moradul Siddique[2]
Department of Computer Science and Engineering
Jashore University of Science and Technology
Jashore 7408, Bangladesh
moradul@just.edu.bd

Md. Delowar Hosen[2]
Department of Computer Science and Engineering
Jashore University of Science and Technology
Jashore 7408, Bangladesh
delowar.cse.just@gmail.com

Md. Shafiuzzaman[1*]
Department of Computer Science and Engineering
Jashore University of Science and Technology
Jashore 7408, Bangladesh
md.shafiuzzaman.cse@just.edu.bd

***Abstract*—Breast Cancer is a critical problem for women. Every year many women die of this issue. In this study, we have shown that comparison on the two breast cancer datasets named as Wisconsin Breast Cancer (Original) (WBC) dataset and Wisconsin Breast Cancer Diagnosis (WBCD) dataset of twelve machine learning algorithms: Naive Bayes (NB), Logistic Regression(LR), Decision Tree Classifier(DT), Support Vector Machine(SVM), Linear Discriminant Analysis(LDA), Voting Classifier(VC), KNeighborsClassifier(K-NN), AdaBoost Classifier(AD), Random Forest Classifier(RF), Stochastic Gradient Descent(SGD), Bagging Classifier(BC), Gradient Boosting Classifier(GB) to find out the best classifier of breast cancer detection. Finally, four classifiers achieved 100% accuracy on the Wisconsin Breast Cancer Diagnosis (WBCD) dataset and six classifiers achieved 100% accuracy on the Wisconsin Breast Cancer (Original) (WBC) dataset and also their sensitivity and specificity values 1.00. To measure the classifier performance we consider accuracy, precision, sensitivity, specificity, False Discovery Rate, and False Omission Rate.**

***Index Terms*—machine learning, breast cancer, classifier, breast cancer diagnosis.**

I. Introduction

Cancer is a long-term morbidity disease. In 2018, about 9.6 million people in the world are dying due to the cancer that informed World Health Organization (WHO) and about 70 percent of the deaths due to cancer happen in the developing countries [1]. In the world, this issue is considered the second reason of human death. Breast cancer is the most common type of cancer [2]. There are two types of breast cancer such as benign and malignant. Benign cancer represents the non-cancerous which is no threat of life but malignant cancer represents the cancerous that is threat of life [5]. Every year, almost 1.5 million women are diagnosed with breast cancer [3]. Approximately 29.9% of deaths from cancer in women are owing to breast cancer [4]. Hence, this is a potentially harmful for women and need more medical equipment and staff to diagnosis the breast cancer patient. Furthermore, it is challenging task to manipulate overall physical examination.

To overcome this challenging task many researchers proposed automated system to detect breast cancer such as MF Aslan et al. [12] proposed four different ML algorithms to detect breast cancer Artificial Neural Network (ANN), standard Extreme Learning Machine (ELM), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN). M.Hussain et al. [13] compared different SVM kernels for the detection of breast cancer and their system achieved around 96% accuracy. Breast cancer is diagnosed using a variety of procedures including physical syndromes, biopsy and radiography image [6]. The Biopsy method is used to ensure the sign of breast cancer. Mammography is the standard method to diagnose breast cancer along with surgical biopsy [7]. Radiology is the medical way that diagnose and treat diseases using clinical images. An effective of the screening process can believe radiologists' explanation [8]. However, radiologists may miss up to 30% of breast cancer based on the density of breasts [9].

Many researchers proposed different techniques to find out the best result of breast cancer detection on different breast cancer datasets. Whereas several studies show the comparison of different ML algorithms, but none of the study did not show the comparison with high efficiency. And the previous research comparison has a limited algorithm. As a result, we did not know that others ML algorithm how perform on those datasets. However, in this research we try eliminate this limitation and consider twelve ML algorithms.

In this study, we have shown a comparison study on two datasets with the high efficiency for breast cancer detection. To get high efficiency, we tuning parameter and show the reason why algorithms provide high accuracy.

The rest of the paper is planned as follows: Section II explains the Literature review.Section III describes the methodology, Section IV explained the results and discussion, and Section V drew the conclusion.

II. Literature Review

Machine learning is one of the most powerful techniques in the classification field. Numerous research has been conducted to apply machine learning in the medical field to classify medical datasets. In this section, we represent previous research on breast cancer detection shortly. MM Islam et al. [10] proposed a novel approach that used two classification algorithms named Support Vector Machine (SVM) and K-Nearest Neighbors (K-NN) for the detection of breast cancer. They used 10-fold cross validation to get accurate results. The approach obtained an accuracy of 98.57% (SVM) and 97.14% (K-NN) respectively using Decision Tree-Support Vector Machine (DT-SVM) in WEKA that is built a hybrid classifier model for predicting breast cancer and acquired an accuracy of 91% besides a low error rate of 2.58% [11]. MF Aslan et al. [12] used blood analysis data to predict breast cancer. They used four different ML algorithms such as Artificial Neural Network (ANN), standard Extreme Learning Machine (ELM), Support Vector Machine (SVM), and K-Nearest Neighbor (K-NN). A study used Breast Cancer Coimbra dataset was taken from UCI ML Repository. In this technique processing many attributes such as glucose, insulin, age, body mass index (BMI), leptin, adiponectin, homeostasis model assessment (HOMA), and chemokine monocyte chemoattractant protein 1 (MCP1). The approach obtained an average accuracy of 73.5%. The performance of SVM depends on the different kernels. M. Hussain et al. [13] compared the different SVM kernel performance for the detection of breast cancer. An automated system provides accurate results and reduces the error rate. Al Bataineh et al. [14] proposed an automated system to compare the performance of five nonlinear machine learning algorithms on the Wisconsin Breast Cancer Diagnostic (WBCD) dataset. To evaluate the performance of these algorithms consider the accuracy, recall and precision. Siham A. Mohammed et al. [15] Proposed an approach that compares the performance of three machine learning techniques on the two breast cancer dataset. Their used dataset are Wisconsin Breast Cancer (WBC) and Breast Cancer dataset. To assess the efficiency of the classifier, consider accuracy, standard deviation, Roc curve, true positive, and false positive. They used a resample filter to increase the performance of the classifier. Agarap et al. [16] proposed a comparison of six machine learning (ML) algorithms on the Wisconsin DiagnosticBreast Cancer (WDBC) dataset. Using digitized images of FNA tests on a breast mass computed some features. Their approach obtained 99.04% test accuracy. Sharma et al. [17] comparison among Random Forest (RF), K-Nearest-Neighbor (K-NN), and Naïve Bayes (NB) ML algorithms using the Wisconsin Diagnosis Breast Cancer dataset. Also, compare the performance of the key parameters such as accuracy, and precision. An optimized model always provides the first and accurate solution. Assegie et al. [18] used a grid search algorithm to find the optimal k values. Then optimized K-NN model used to detect breast cancer. Finally, they compared the result between optimized model and default hyper-parameter model where optimized model obtained the highest result is 94.35%.

III. Methodology

In this section, we analyze our experiment steps how process the data, visualize the data, split the data for train-test and model fitting. And give the flowchart of our study. The given Figure 1 represent our experiment steps. According to the figure we describe each following steps.

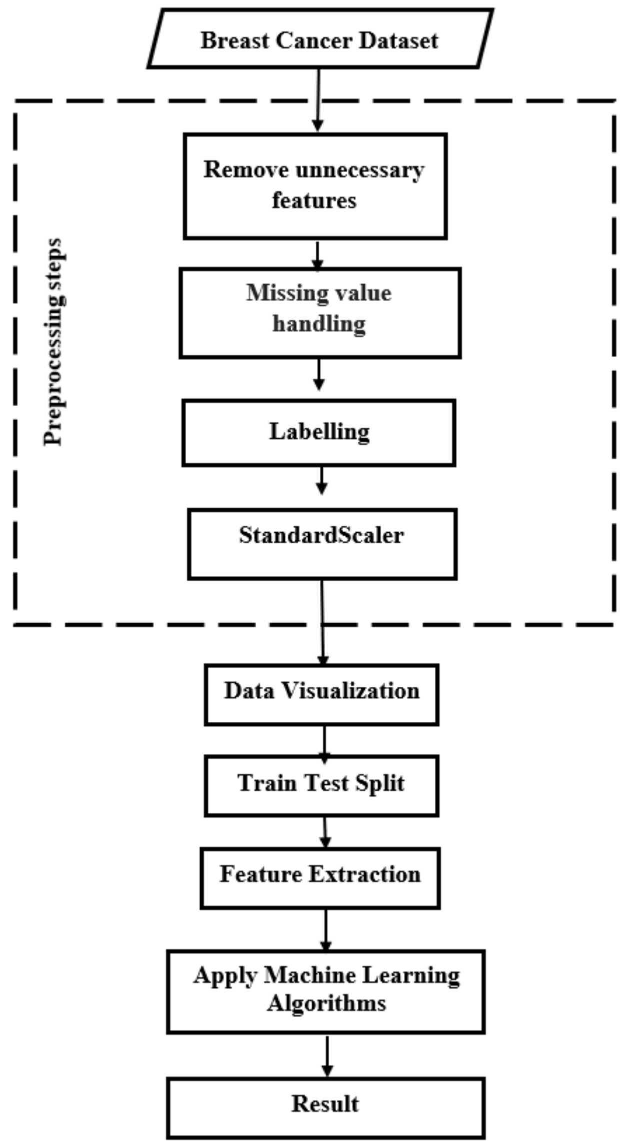

Fig. 1. A typical workflow diagram of our experiments.

A. Dataset Description

In this study, we used two datasets as Wisconsin Breast Cancer (Original) (WBC) dataset and Wisconsin Breast Cancer Diagnosis (WBCD) dataset. Both datasets collect from Kaggle. Wisconsin Breast Cancer (Original) (WBC) dataset contains 699 samples with 11 features. They are two classes 241 (34.5%) Malignant and 458 (65.5%) of Benign. The datasets have 16 missing values that represent the question mark(?) symbol. And the Wisconsin Breast Cancer Diagnosis (WBCD) dataset contains 569 samples with 32 features where 357 (62.74%) Benign and 212 (37.26%) Malignant.

B. Data Preprocessing

Data preprocessing is one of the crucial phases in any machine learning-based application. In the preprocessing stage, at first drop the "Sample code number" feature from the WBC dataset and id, Unnamed: 32 from the WBCD dataset. The WBC data missing value ("?") was replaced by the mode using the mode () function. The mode is a statistical term that returns the value which is most frequently occurred enter a dataset. And to labeling and for Standard Scaling used LabelEncoder () and StandardScaler () functions.

C. Performance Metrics

The efficiency of machine learning algorithms is assessed using a set of performance measures. To evaluate the parameter, TP, FP, TN, and FN are used to create a confusion matrix for the actual and predicted classes. The meanings of the terms are listed below.

- TP stands for True Positive (Correctly Classified)
- TN stands for True Negative (Incorrectly Classified)
- FP stands for False Positive (Correctly misclassified)
- FN stands for False Negative (Incorrectly misclassified)

The following formulas are used to evaluate the proposed system's performance.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

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

$$Sensitivity\ or Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

$$FalseDiscoveryRate = \frac{FP}{FP + TP} \tag{6}$$

$$FalseOmissionRate = \frac{FN}{FN + TN} \tag{7}$$

D. Data Visualization

Data visualization is the most important part of any machine learning application. Through the data visualization, we find out the characteristics of data and how to correlate features to features. There are two classes as Benign and malignant. The Figure 2 and Figure 3, we see that the two classes are clearly separated. That is pretty good, any machine learning algorithm is easily classified into two separate categories and got high accuracy.

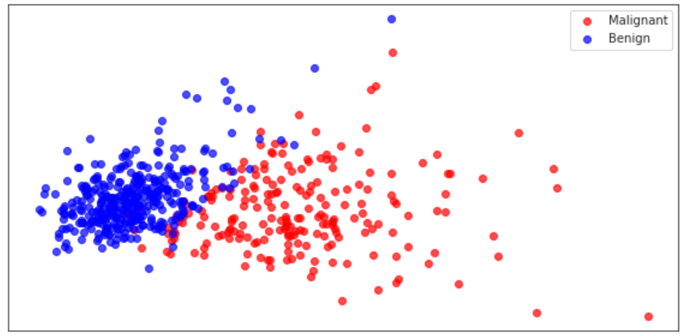

Fig. 2. Represent two classes are separated on the WBCD datasets.

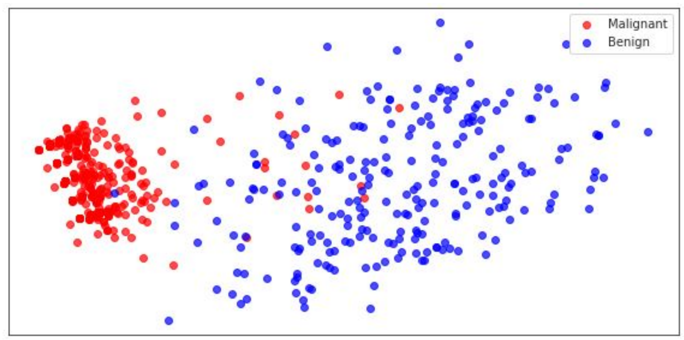

Fig. 3. Represent two classes are separated on the WBC datasets.

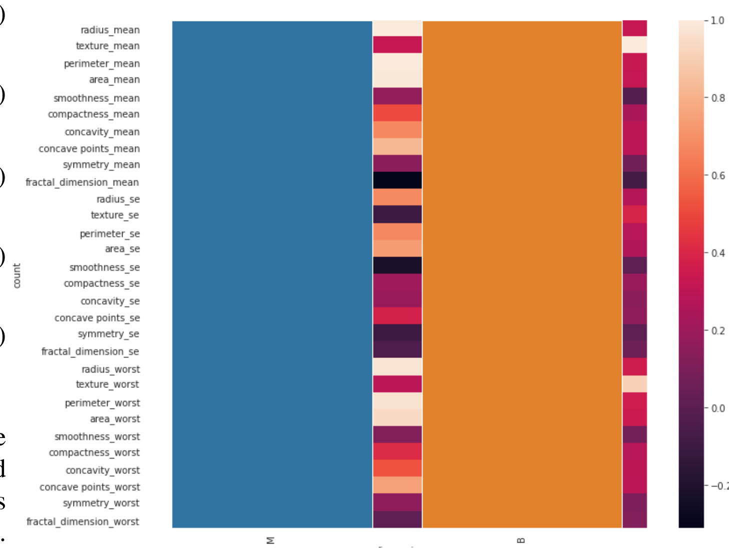

Fig. 4. Heat map for checking correlated features on the WBCD dataset.

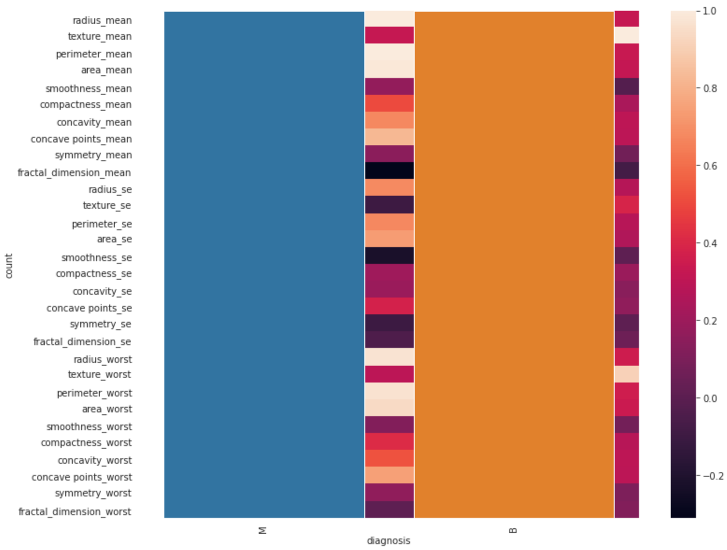

Fig. 5. Heat map for checking correlated features on the WBC dataset

The WBCD dataset has 32 features and the WBC dataset has 11 features. Following the figure shows that the correlation among features on the two datasets.

Both the Figures 4 and 5, we have seen that there are no correlated features.

IV. Explained the Results and Discussion

We have performed a comparison-based study of various machine learning algorithms using two breast cancer datasets. Therefore, in our comparison study describes twelve machine learning algorithms such as Naive Bayes (NB), Logistic Regression (LR), Decision Tree Classifier(DT), Support Vector Machine(SVM), Linear Discriminant Analysis(LDA), Voting Classifier(VC), KNeighbors Classifier(K-NN), AdaBoost Classifier(AD), Random Forest Classifier(RF), Stochastic Gradient Descent(SGD), Bagging Classifier(BC), Gradient Boosting Classifier(GB). In the case of WBCD dataset, we considered 512 samples (90%) for training and 57 samples (10%) for testing, as well as 489 samples (90%) for training and 210 samples (10%) for testing on the WBC dataset. We consider the same weight for all the features on both datasets and also consider default parameters except random state. The random state was considered as 52 and 101 for the dataset WBCD and WBC respectivly. Because for those random state some classifiers provide highest accuracy. Furthermore, to determine the performance of all classifiers have been assessed by evaluating in terms of accuracy, precision, F1-measures, specificity, sensitivity, False Discovery Rate (FDR), False Omission Rate (FOR). The outcomes of our experiments for Breast Cancer detection on the WBCD dataset are represented in Table I and Table II.

We have seen that Table I & II the four (DT, KNN, RF and GB) classifiers achieved the 100% accuracy with 0 false discovering and emission rate that's mean the classifier correctly classify every testing instance. Consequently, others seven algorithms (LR, SVM, LDA, AB, VC, SGD and BC) ensure 98% accuracy with the 1 Specificity. And NB classifier got 94.74% where false discovery rate 0.10.

On the other hand, six classifiers provide the 100% accuracy on the WBC datasets with 0 false discovering rate and 1 F1-measures, Specificity, Sensitivity which represent Table III & IV. Moreover, the other two algorithms (NB, LDA) delivered 98.57% accuracy. Even though KNN, AB, GB provide the 97.14 % accuracy and DT classifier achieved 94.29% with 0.93 and 0.90 Sensitivity.

TABLE I
Represent the six classifiers performance on WBCD

Parameters	NB	LR	DT	SVM	LDA	KNN
Accuracy(%)	94.74	98.25	100	98.25	98.25	100
Precision	0.97	0.97	1.00	0.97	0.97	1.00
F1-measures	0.96	0.99	1.00	0.99	0.99	1.00
Specificity	0.95	1.00	1.00	1.00	1.00	1.00
Sensitivity	0.95	0.95	1.00	0.95	0.95	1.00
False Discovery Rate	0.10	0.00	0.00	0.00	0.00	0.00
False Omission Rate	0.027	0.026	0.00	0.026	0.026	0.00

TABLE II
Represent the another six classifiers performance on WBCD

Parameters	AB	RF	VC	SGD	BC	GB
Accuracy(%)	98.25	100	98.25	98.25	98.25	100
Precision	0.97	1.00	0.97	0.97	0.97	1.00
F1-measures	0.99	1.00	0.99	0.99	0.99	1.00
Specificity	1.00	1.00	1.00	1.00	1.00	1.00
Sensitivity	0.95	1.00	0.95	0.95	0.95	1.00
False Discovery Rate	0.00	0.00	0.00	0.00	0.00	0.00
False Omission Rate	0.026	0.00	0.026	0.026	0.026	0.00

TABLE III
Represent the six classifiers performance on WBC

Parameters	NB	LR	DT	SVM	LDA	KNN
Accuracy(%)	98.57	100	94.29	100	98.57	97.14
Precision	1.00	1.00	0.93	1.00	0.98	0.95
F1-measures	0.99	1.00	0.95	1.00	0.99	0.98
Specificity	0.98	1.00	0.98	1.00	1.00	1.00
Sensitivity	1.00	1.00	0.90	1.00	0.97	0.93
False Discovery Rate	0.033	0.00	0.037	0.00	0.00	0.00
False Omission Rate	0.00	0.00	0.069	0.00	0.024	0.047

TABLE IV
Represent the another six classifiers performance on WBC

Parameters	AB	RF	VC	SGD	BC	GB
Accuracy(%)	97.14	100	100	100	100	97.14
Precision	0.95	1.00	1.00	1.00	1.00	0.95
F1-measures	0.98	1.00	1.00	1.00	1.00	0.95
Specificity	1.00	1.00	1.00	1.00	1.00	.98
Sensitivity	0.93	1.00	1.00	1.00	1.00	0.97
False Discovery Rate	0.00	0.00	0.00	0.00	0.00	0.034
False Omission Rate	0.047	0.00	0.00	0.00	0.00	0.024

V. Conclusion

Machine learning algorithms have been used for different applications in medical sectors and as well as an effective tool for guiding clinicians in making decisions based on available data and producing medical expert machines. In medical fields, breast cancer prediction is much significance. The aim of this paper was to compare several classifier models that could predict breast cancer using twelve machine learning algorithms. Hereby, the two datasets Wisconsin Breast Cancer (Original)(WBC) and Wisconsin Breast Cancer Diagnosis (WBCD) achieved the highest performance in terms of all performance metrics. In future work, researchers can be tuning the parameter to get the highest accuracy of other algorithms and to provide an efficient model.

References

[1] https://www.who.int/en/news-room/fact-sheets/detail/cancer. Last Access: 25.01.2019.

[2] Sivakami, K., and Nadar Saraswathi. "Mining big data: breast cancer prediction using DT-SVM hybrid model." International Journal of Scientific Engineering and Applied Science (IJSEAS) 1, no. 5 (2015): 418-429.

[3] Mojrian, Sanaz, Gergo Pinter, Javad Hassannataj Joloudari, Imre Felde, Akos Szabo-Gali, Laszlo Nadai, and Amir Mosavi. "Hybrid machine learning model of extreme learning machine radial basis function for breast cancer detection and diagnosis; a multilayer fuzzy expert system." In 2020 RIVF International Conference on Computing and Communication Technologies (RIVF), pp. 1-7. IEEE, 2020.

[4] H. You and G. Rumbe, "Comparative study of classification techniques on breast cancer FNA biopsy data," International Journal of Artificial Intelligence and Interactive Multimedia, vol. 1, no. 3, pp. 6-13, 2010.

[5] Al Bataineh, Ali. "A comparative analysis of nonlinear machine learning algorithms for breast cancer detection." International Journal of Machine Learning and Computing 9, no. 3 (2019): 248-254.

[6] A. A. Ardakani, A. Gharbali, and A. Mohammadi, "Classification of breast tumors using sonographic texture analysis," J. Ultrasound Med., vol. 34, no. 2, pp. 225–231, 2015.

[7] Al Bataineh, Ali. "A comparative analysis of nonlinear machine learning algorithms for breast cancer detection." International Journal of Machine Learning and Computing 9, no. 3 (2019): 248-254.

[8] B. L. Sprague, E. F. Conant, T. Onega, M. P. Garcia, E. F. Beaber, S. D. Herschorn, C. D. Lehman, A. N. A. Tosteson, R. Lacson, M. D. Schnall, D. Kontos, J. S. Haas, D. L. Weaver, and W. E. Barlow, "Variation in Mammographic Breast Density Assessments among Radiologists in Clinical Practice: A Multicenter Observational Study," Ann. Intern. Med., vol. 165, no. 7, pp. 457–464, 2016.

[9] T. M. Kolb, J. Lichy, and J. H. Newhouse, "Comparison of the Performance of Screening Mammography, Physical Examination, and Breast US and Evaluation of Factors that Influence Them: An Analysis of 27,825 Patient Evaluations," Radiology, vol. 225, no. 1, pp. 165–175, 2002.

[10] Islam, Md Milon, Hasib Iqbal, Md Rezwanul Haque, and Md Kamrul Hasan. "Prediction of breast cancer using support vector machine and K-Nearest neighbors." In 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), pp. 226-229. IEEE, 2017.

[11] Sivakami, "Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model", 2015

[12] Aslan, Muhammet Fatih, Yunus Celik, Kadir Sabancı, and Akif Durdu. "Breast cancer diagnosis by different machine learning methods using blood analysis data." (2018).

[13] M. Hussain, S. K. Wajid, A. Elzaart, and M. Berbar, "A comparison of SVM kernel functions for breast cancer detection." IEEE Eighth International Conference Computer Graphics, Imaging and Visualization, pp. 145-150, 2011.

[14] Al Bataineh, Ali. "A comparative analysis of nonlinear machine learning algorithms for breast cancer detection." International Journal of Machine Learning and Computing 9, no. 3 (2019): 248-254.

[15] Mohammed, Siham A., Sadeq Darrab, Salah A. Noaman, and Gunter Saake. "Analysis of breast cancer detection using different machine learning techniques." In International Conference on Data Mining and Big Data, pp. 108-117. Springer, Singapore, 2020.

[16] Agarap, Abien Fred M. "On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset." In Proceedings of the 2nd international conference on machine learning and soft computing, pp. 5-9. 2018.

[17] Sharma, Shubham, Archit Aggarwal, and Tanupriya Choudhury. "Breast cancer detection using machine learning algorithms." In 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), pp. 114-118. IEEE, 2018.

[18] Assegie, Tsehay Admassu. "An optimized K-Nearest Neighbor based breast cancer detection." Journal of Robotics and Control (JRC) 2, no. 3 (2021): 115-118.