A Comparative Study on Supervised Learning Algorithms in Breast Cancer Detection

Ivar Ekeland[1], Roger Temam[2] Jeffrey Dean, David Grove, Craig Chambers, Kim B. Bruce, and Elsa Bertino

[1] Princeton University, Princeton NJ 08544, USA,
1.Ekeland@princeton.edu,
WWW home page: http://users/~iekeland/web/welcome.html
[2] Université de Paris-Sud, Laboratoire d'Analyse Numérique, Bâtiment 425, F-91405 Orsay Cedex, France

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.

Keywords: machine learning, breast cancer, classifier, breast cancer diagnosis

1 Introduction

Cancer is a long-term morbidity disease. It is considered as the second most common reason of human death. According to the World Health Organization (WHO), about 9.6 million people died due to the cancer in 2018 and 70% of those happened in developing countries where cancer diagnosis facilities are still very expensive [1]. Breast cancer is the most common type of cancer [2]. There are two types of cancer such as benign and malignant. Benign cancer represents the non-cancerous which has no threat to life but malignant cancer represents the most cancerous and it has direct threat to 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, it is potentially most harmful disease for women. However, it requires dozens of medical equipment and staffs to diagnosis a breast cancer patient.

Breast cancer can be 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. However, effectiveness of this process depends radiologists' explanation [8] and radiologists may miss up to 30% of breast cancer based on the density of breasts [9].

To overcome the issues with breast cancer diagnosis, computer scientists have contributed with several automated methods. MF Aslan et al. [12] proposed four different Machine Learning (ML) algorithms to detect breast cancer, such as Artificial Neural Network (ANN), standard Extreme Learning Machine (ELM), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN). Another study, M.Hussain et al. [13] compared different SVM kernels for the detection of breast cancer and their system achieved around 96% accuracy. On the other hand, Bayrak et al.[19] to compare the machine learning model performance, they apply two ML (SVM, ANN) model on the Wisconsin Breast Cancer (Original) dataset. For performance measure, they consider accuracy, precision, recall and ROC Area. And Agarap et al. [16] proposed a comparison of six machine learning (ML) algorithms: GRU-SVM[4], Linear Regression, Multilayer Perceptron (MLP), Nearest Neighbor (NN) search, Softmax Regression, and Support Vector Machine (SVM on the Wisconsin Diagnosis Breast Cancer (WDBC) datasets. Among of them MLP algorithm achieved the highest accuracy is 99.04%. Moreover, Potdar et al.[20] claim that Artificial Neural Networks (ANN) is better for breast cancer classification than K-NN and Bayesian Classifiers that provides 97.4% accuracy . And Gayathri et al [21] represent another comparison study where Relevance vector machine(RVM) provides Low computational cost even the variables are reduced compared with other machine learning algorithms that are used for breast cancer detection.

However, none of those approaches did extensive analysis of ML algorithms. Besides most of those showed their analysis on a single dataset. In this study, we have used 12 ML algorithms in two different datasets.

The rest of the paper is arranged as follows: Section II explains the literature review, section III describes the methodology of the comparative study, section IV shows the comparative results and discussions, and finally, section V draws the conclusion.

2 Literature Review

As known, an optimized model always provides very competitive with the first and accurate solution. One of the most powerful techniques in the classification field is machine learning. Numerous studies have been conducted to apply machine learning in the medical field to classify medical datasets. Recently, to

determine the better k value, applied a grid search algorithm. Thus optimized the K-NN model to detect breast cancer. Further, they compared the results of the optimized model to the default hyper-parameter model, with the optimized model coming out ahead with a score of 94.35% [18]. Another approach obtained an accuracy of 98.57% (SVM) and 97.14% (K-NN) respectively using Decision Tree-Support Vector Machine (DT-SVM) in WEKA, which built a hybrid classifier model for predicting breast cancer and acquired an accuracy of 91% as well as a low error rate of 2.58% [11]. Likewise, blood data has been used to determine breast cancer in another experiment. In the experiment, the dataset was taken from UCI ML Repository and 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); Based on the approaches obtained an average accuracy of 73.5% [12]. Likewise, based on performance metrics (such as accuracy, recall, and precision), they compared the effectiveness and efficiency of five nonlinear ML algorithms on the Wisconsin Breast Cancer Diagnostic (WBCD) dataset [14]. Moreover, Sharma et al.[17] compared the performances among the algorithms (RF, K-NN, and NB) on the WDBC dataset. Similarly, Mohammed et al. [15] used two dataset (WBC and WBCD) and presented an approach that compare the performance of ML techniques. To assess the efficiency of the classifiers, they considered accuracy, standard deviation, ROC curve and so on. Especially, a resample filter was applied to increase the performance.

For the detection of breast cancer, MM Islam [10] suggested two ML techniques such as SVM and K-NN and to optimize effectiveness, they used 10-fold cross validation. Another work by M. Hussain compared the different SVM kernel performances to detect breast cancer, providing accurate results and reducing the error rate [13]. To get better features used digitized images of Fine Needle Aspiration (FNA) tests on the WBCD dataset simultaneously compared among the employed three ML algorithms which was obtained 99.04% test accuracy [10].

3 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.

3.1 Environment setup

To the compared outcomes of the ML techniques basis on the two breast cancer datasets, we applied twelve ML techniques that are implemented in a local computer. The configuration of the computer was Intel Celeron with 8GB RAM. We used open source platform as Scikit-learn for ML library. An Integrated development environment named as Jupyter Notebook is used to run the program.

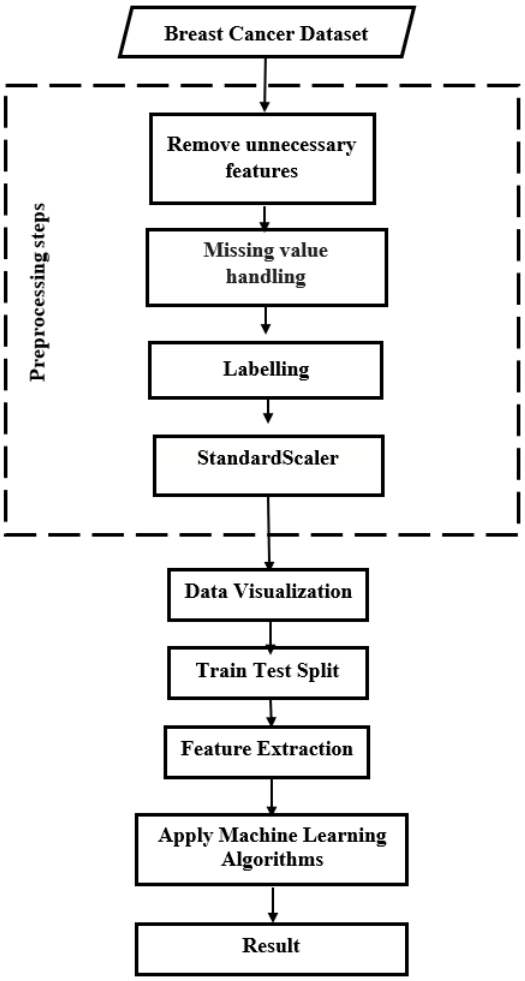

Fig. 1. A typical workflow diagram of our experiments.

3.2 Dataset Description

In this study, we used two datasets as Wisconsin Breast Cancer (Original) (WBC) and Wisconsin Breast Cancer Diagnosis (WBCD). Both datasets collected from Kaggle. Wisconsin Breast Cancer (Original) (WBC) dataset contains 699 samples with 11 features. They are two classes as Malignant and Benign that is denotes 4 and 2 respectably. Following the figure 2 show the statistics of WBC datasets. The datasets have 16 missing values that represent the question mark("?") symbol.

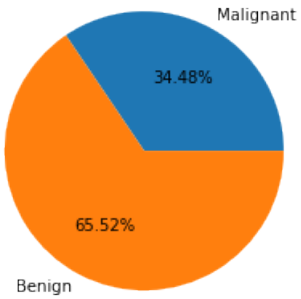

Fig. 2. Statistics of Wisconsin Breast Cancer (Original) (WBC) dataset.

And the Wisconsin Breast Cancer Diagnosis (WBCD) dataset contains 569 samples with 32 features where two classes are denotes as 'M'(Malignant) and 'B'(Benign). And following the figure 3 show the statistics of WBCD datasets.

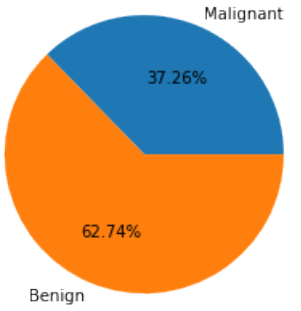

Fig. 3. Statistics of Wisconsin Breast Cancer Diagnosis (WBCD) dataset.

3.3 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.

3.4 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.

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

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

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

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

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

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

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

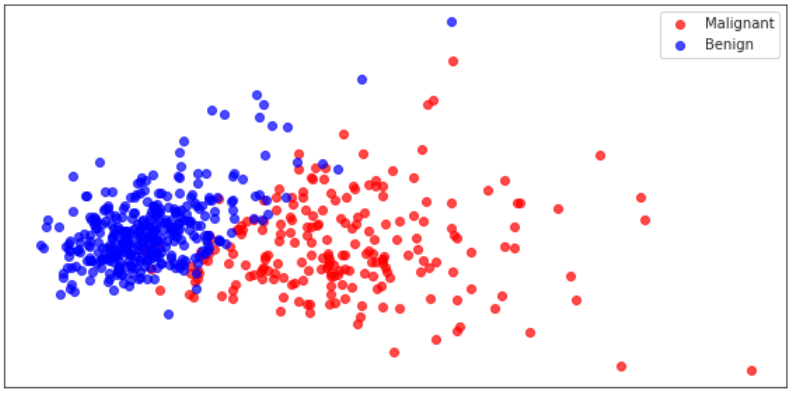

Fig. 4. Represent two classes are almost separated on the WBCD datasets.

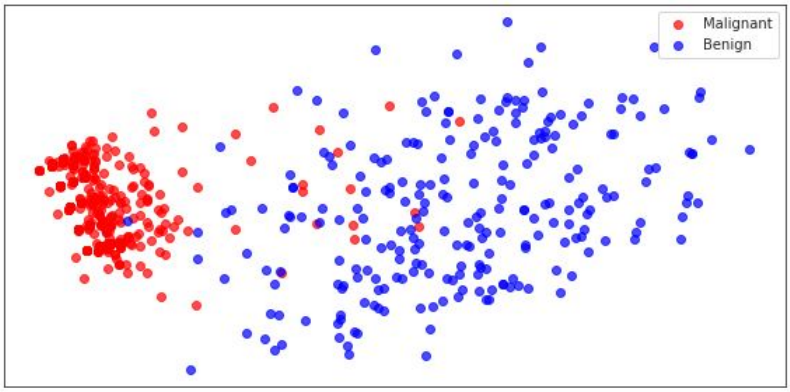

Fig. 5. Represent two classes are almost separated on the WBC datasets.

3.5 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 4 and Figure 5, we see that the two classes are almost separated. As a result, machine learning algorithm is easily classified into two separate categories and achieved high accuracy.

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 6 and 7, we have seen that there are no correlated features.

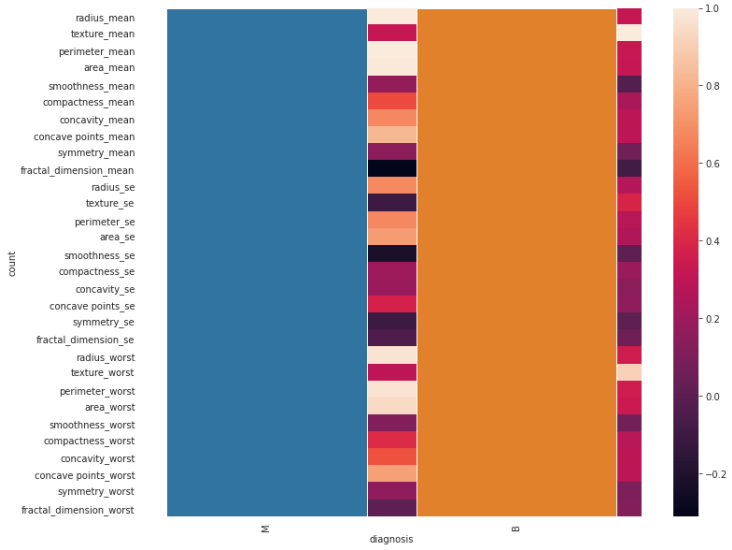

Fig. 6. Heat map for checking correlated features on the WBCD dataset.

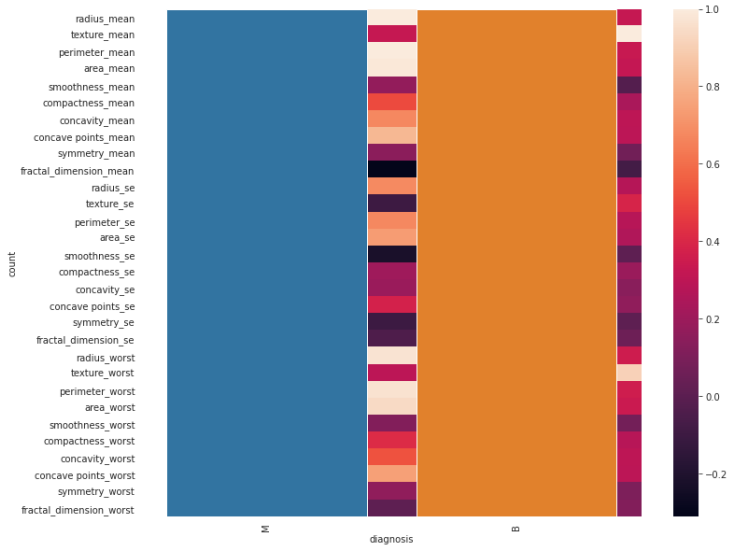

Fig. 7. Heat map for checking correlated features on the WBC dataset

4 Explanation of 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, 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. In the experiment, 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 1 and Table 2.

Table 1. Represent the six classifiers performance on WBCD

Parameters	**NB**	**LR**	**DT**	**SVM**	**LDA**	**KNN**
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

From the Table 1 & 2, it is clearly show that 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

Table 2. Represent the another six classifiers performance on WBCD

Parameters	**AB**	**RF**	**VC**	**SGD**	**BC**	**GB**
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 3. Represent the six classifiers performance on WBC

Parameters	NB	LR	DT	SVM	LDA	KNN
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 4. Represent the another six classifiers performance on WBC

Parameters	AB	RF	VC	SGD	BC	GB
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

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 is the F1-measures, Specificity, Sensitivity which represent Table 3 & 4. 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 respectively.

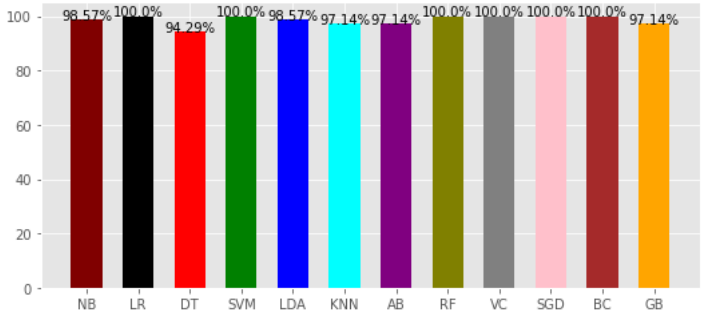

Fig. 8. Results obtained from WBC dataset.

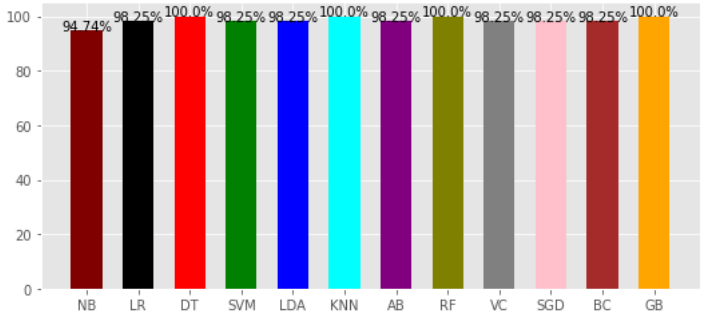

Fig. 9. Result obtained from WBCD dataset.

Following the figure shown the comparison of twelve ML algorithms. From the Figure 8 show that six ML algorithms obtained the 100% accuracy on the WBC dataset and figure 9 four Algorithms obtained the 100% accuracy on the WBCD dataset.

5 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 obtained the highest efficiency of other algorithms and to provide an efficient model.

References

[1] https://www.who.int/en/news-room/fact-sheets/detail/cancer. Last Access: 06.02.2022.

[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.

[19] Bayrak, Ebru Aydındag, Pınar Kırcı, and Tolga Ensari. "Comparison of machine learning methods for breast cancer diagnosis." In 2019 Scientific meeting on electrical-electronics biomedical engineering and computer science (EBBT), pp. 1-3. Ieee, 2019.

[20] Potdar, Kedar, and Rishab Kinnerkar. "A comparative study of machine learning algorithms applied to predictive breast cancer data." International Journal of Science and Research 5, no. 9 (2016): 1550-1553.

[21] Gayathri, B. M., and C. P. Sumathi. "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer." In 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-5. IEEE, 2016.