A Comparative Study on Supervised Learning Algorithms in Breast Cancer Detection

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.

Introduction

In the last century, it was thought that the most people die due to cancer. For that, it is considered the second most common reason of human death. In addition, cancer diagnostics are still out of reach for most people in the world. 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]. Among all the types of cancer, breast cancer is the most common type of cancer in the world [2]. It is estimated that global breast cancer cases will grow from 1.4 million in 2008 to over 2.1 million cases in 2030 [27]. 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]. There are two types of breast cancer such as benign and malignant. Benign represents the non-cancerous which has no threat to life but malignant represents the most cancerous and it has direct threat to life [5]. It is very important to identify cancer accurately as benign and malignant. Because cancer identification is the first stage of cancer diagnosis. If, for some reason there is a mistake in cancer identification, it means that the entire treatment of cancer will be affected by cancer identification.

However, cancer diagnosis requires dozens of medical equipment and staff to diagnose a breast cancer patient. Breast cancer can be diagnosed using a variety of procedures including physical syndromes, biopsy, and radiographic images [6]. The biopsy method is used to ensure the presence of breast cancer. However, biopsy methods are extremely dependent on a doctor's expertise. Mammography is the standard diagnostic method for breast cancer and surgical biopsy [7]. Though mammography does not provide 100% accurate results and sometimes finds something that is not cancer and it may miss some cancer. Radiology is the medical way that diagnoses and treats diseases using clinical images. However, the effectiveness of this process depends on radiologists' explanation [8] and radiologists may miss up to 30% of breast cancer based on the density of breasts [9]. So, the overall manual process of breast cancer diagnosis does not provide good results and at the same time, it is time-consuming and expensive. We used digital approaches to reduce the cost of breast cancer diagnosis and speed up the process. Machine learning methods are used in these digital approaches.

To overcome the manual process issues associated with a breast cancer diagnosis, computer scientists have contributed with several automated machine learning 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). Among of them, Extreme Learning Machine (ELM) achieved the highest accuracy (80%) with 0.0075s training time. However, other three methods achieved less accuracy with much training time. So, their overall performance is not so good to identify breast cancer. In another study, Potdar et al. [20] used three machine learning methods namely Artificial Neural Networks (ANN), K-Nearest Neighbor (KNN), and Bayesian Classifiers to classify breast cancer. They used 3-fold Cross Validation to eliminate the data imbalanced problem. In 3-fold Cross Validation, Artificial Neural Networks (ANN) provide the highest 97.4%. after that, the authors reach a conclusion that Artificial Neural Networks (ANN) are better for breast cancer classification than K-NN and Bayesian Classifiers. Another technique of breast cancer detection is Convolutional Neural Networks (CNN). Convolutional Neural Networks (CNN) is the best ML technique to classify the imaging problem. Convolutional Neural Networks (CNN) provide the best result when the image quality is so good. In the research, Y. J. Tan et al. [28] took the help of Convolutional Neural Networks (CNN) to identify breast cancer for the Mammogram Imaging. They used three version (version 1, 2, 3) of Convolutional Neural Networks (CNN). In this study achieved the highest results of 82.71% of version 3. However, performance in detecting breast cancer remains poor.

Every machine learning algorithm has some flaws. Therefore, We need to conduct a comparison study among them. Already computer scientists have contributed to several comparative studies. Such as 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] compare the machine learning model performance, applied two ML (SVM, ANN) models to the Wisconsin Breast Cancer (Original) dataset. For performance measures, they consider accuracy, precision, recall and ROC Area. In this study SVM obtained the best result with 96.997%. Moreover, 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 them, the MLP algorithm achieved the highest accuracy (99.04%). Additionally, Gayathri et al [21] represent another comparison study where Relevance Vector Machine (RVM) provides a low computational cost even though the variables are reduced compared with other machine learning algorithms that are used for breast cancer detection. However, the majority of those only showed their analysis on a single dataset, and none of those approaches performed a thorough investigation of ML algorithms, thus their comparative study was not fully completed. This is why we intend to conduct a comprehensive comparative study.

In this study, we have used Twelve reputed ML algorithms in two different datasets. The Twelve reputed ML algorithms namely 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). Hereby, widely verified these two datasets and did numerous research on them. For that, we choose these two datasets. The two datasets namely Wisconsin Breast Cancer (Original) (WBC) and Wisconsin Breast Cancer Diagnosis (WBCD). Both the dataset was collected from the Kaggle. The dataset is split into 90% and 10% for training and testing, respectively. We used default parameter of all algorithms except random state. Among all the classifier DT, KNN, RF and GB achieved the 100% on the WBC dataset and also LR, SVM, RF, VC, SGD, and BC achieved the 100% on the WBCD dataset. Others classifier obtained the 94% above accuracy on both datasets. Finally, we compared the performance of each ML algorithm on two 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.

Literature Review

Medical science has greatly benefited from the application of machine learning and data mining in technology. Numerous studies have been conducted to apply machine learning in the medical field to classify medical datasets. This section provides a brief overview of several studies conducted in the field of breast cancer detection.

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. The Wisconsin breast cancer diagnosis dataset that was used in their implementation was taken from the UCI machine learning repository. For accurate results, 10-fold cross-validation was performed. The approach's accuracy was 98.57% for SVM and 97.14% for K-NN, respectively, and their model's specificity for the testing phase was 95.65% for SVM and 92.31% for K-NN. However, their proposed model used a single dataset and only two machine learning techniques.

In another study, K Sivakami et al. [11] presents a disease status prediction strategy. This strategy is divided into two parts. 1. Information Treatment and Option Extraction, and 2. Decision Tree-Support Vector Machine (DT-SVM) Hybrid Model for predictions. They used Weka Software tools for data preparation, data analysis, and result comparison. In this study, three classification techniques are compared, and DT-SVM (91%) outperforms Instance-based Learning (IBL), Sequential Minimal Optimization (SMO), and Nave-based classifiers. They also used the Wisconsin breast cancer diagnosis dataset from the UCI machine learning repository. Furthermore, they only used single dataset, and their prediction results are poor. In addition, their dataset distribution (60% training and 40 testing) is not perfect because the sample dataset is limited.

Similarly, AA Bataineh et al. [14] 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: Multilayer Perceptron (MLP), K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Gaussian Nave Bayes (NB), and Support Vector Machines (SVM). The Multilayer Perceptron (MLP) algorithm achieved the highest accuracy of the five nonlinear ML algorithms, with 99.12%. Despite having more proposed algorithms than the previous two and offering higher performance, the dataset problem persists.

Moreover, Mohammed et al. [15] proposed an approach that improves the accuracy and enhances the performance of three different classifiers: Decision Tree (J48), Naïve Bayes (NB), and Sequential Minimal Optimization (SMO). They conducted two different datasets as Wisconsin Breast Cancer (WBC) and the Breast Cancer dataset. The authors of this paper mainly focus to deal on imbalanced data. Data imbalance is a big problem in the classification field. The resampling techniques are used to deal with imbalanced data. In addition, they employed the 10-fold crossed validation method to solve the Data Unbalance issue. In particular, a resample filter was used to improve classifier performance. SMO outperformed the other two classifiers in terms of accuracy among the three. To assess the efficiency of the classifiers, they considered accuracy, standard deviation, ROC curve, and so on.

In addition, S Sharma et al. [17] compare the performance of three machine learning algorithms (ML) on the Wisconsin Diagnosis Breast Cancer dataset: Random Forest, K-NN(K-Nearest-Neighbor), and Naive Bayes. To compare the model performances, they have considered accuracy and precision. The K-Nearest-Neighbor(K-NN) (95.90%) outperforms Random Forest and Naïve Bayes. Furthermore, K-Nearest-Neighbor (K-NN) had the highest precision (98.27%) and f1-Score (94.20%). This study also made use of fewer ML algorithms and a single dataset.

Some researchers optimize the model to improve performance. To detect breast cancer, Assegie et al. [18] proposed an optimized K-Nearest Neighbor (K-NN) model. They used grid search techniques to find the best value of k for the K-NN model. This study also compares the effect of the hyper-parameter tuning model to the effect of the default hyper-parameter model. Hyper-parameter tuning has a significant impact on the performance of the KNN model. The optimized hyper-parameter tuning model then achieved 94.35%, while the default hyper-parameter achieved 90.10%. However, the model's performance is insufficient, and it only used a single dataset.

On the other hand, MF Aslan et al. [12] employed four machine learning algorithms for the early identification of breast cancer: the Artificial Neural Network (ANN), the conventional Extreme Learning Machine (ELM), the Support Vector Machine (SVM), and the K-Nearest Neighbor algorithm (k-NN). Their objective is to process the results based on the routine blood analysis and determine how well these techniques work to find breast cancer. They considered age, body mass index (BMI), glucose, insulin, homeostasis model assessment (HOMA), leptin, adiponectin, resistin, and chemokine monocyte chemoattractant protein 1 (MCP1) attributes across the entire blood analysis dataset to complete this study. The blood analysis dataset was collected from papers M Patrício et al. [28]. They also tune the model's hyper-parameter for better results. Finally, standard Extreme Learning Machine (ELM) methods obtained the best results (80%) with the shortest training time (0.0075s) compared to the other three. Despite the fact that this study used a different dataset, model performance issues persist.

In order to identify the subtype of breast cancer, AA Bataineh et al. [7] present a performance comparison study. Five nonlinear machine learning methods are compared in this study: Gaussian Nave Bayes (NB), Classification and Regression Trees (CART), Multilayer Perceptron (MLP), and Support Vector Machines (SVM). Wisconsin Breast Cancer Diagnostic is their conducted dataset (WBCD). Finally, they evaluate the model performance with respect to the effectiveness and efficiency of each algorithm in terms of precision, recall, and accuracy. The Multilayer Perceptron (MLP) model outperformed the others in terms of accuracy (96.70%), precision (100%), and recall (97%). They used 10-fold cross-validation for more accurate results.

Table: Comparison of Previous Study

Authors	Models	Dataset	Results (Best)	Weakness
MM Islam et al. [10]	1. Support Vector Machine (SVM) 2. K-Nearest Neighbors (K-NN)	Wisconsin breast cancer diagnosis (WBCD) dataset	SVM (98.57%)	1. Single dataset Only two 2. Machine learning techniques.
K Sivakami et al. [11]	1. DT-SVM 2. Instance-based Learning (IBL) 3. Sequential Minimal Optimization (SMO) 4. Nave-based classifiers.	Wisconsin breast cancer diagnosis (WBCD) dataset	DT-SVM (91%)	1. Single dataset 2. Results are poor 3. Dataset distribution (60% training and 40 testing) is not perfect

				4. Fewer ML algorithms
AA Bataineh et al. [14]	1. Multilayer Perceptron (MLP) 2. K-Nearest Neighbors (KNN) 3. Classification and Regression Trees (CART) 4. Gaussian Nave Bayes (NB) 5. Support Vector Machines (SVM)	Wisconsin breast cancer diagnosis (WBCD) dataset	MLP (99.12%)	1. Single dataset
Mohammed et al. [15]	1. Decision Tree (J48) 2. Naïve Bayes (NB) 3. Sequential Minimal Optimization (SMO)	1. Wisconsin Breast Cancer (WBC) 2. Breast Cancer dataset	1. For Wisconsin Breast Cancer (WBC) SMO (99.56%) 2. For Breast Cancer dataset J48 (98.20%)	1. Fewer ML algorithms
S Sharma et al. [17]	1. Random Forest 2. K-NN(K-Nearest-Neighbor) 3. Naive Bayes	Wisconsin breast cancer diagnosis (WBCD) dataset	K-NN (95.90%)	1. Fewer ML algorithms 2. Single dataset
Assegie et al. [18]	1. K-Nearest Neighbor (K-NN)	Wisconsin breast cancer (WBC) dataset	K-NN (94.35%)	1. Model performance is insufficient 2. Single dataset
MF Aslan et al. [12]	1. Artificial Neural Network (ANN) 2. conventional Extreme 3. Learning Machine (ELM) 4. Support Vector Machine (SVM) 5. K-Nearest Neighbor algorithm (k-NN).	Blood analysis Dataset	ELM (80%)	1. Model performance is insufficient
AA Bataineh et al. [7]	1. Gaussian Nave Bayes (NB) 2. Classification and Regression Trees (CART) 3. Multilayer Perceptron (MLP)	Wisconsin breast cancer diagnosis (WBCD) dataset	MLP (96.70%)	1. Low performance 2. Single dataset

	4. Support Vector Machines (SVM)			

We can observe from the summary above that several studies employed few machine learning (ML) approaches, had poor performance, and only used one or two datasets. In this study, we attempt to address these problems and demonstrate how these ML approaches compare to one another. From the prior research, we chose two datasets and twelve machine learning approaches to conduct this study.

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.

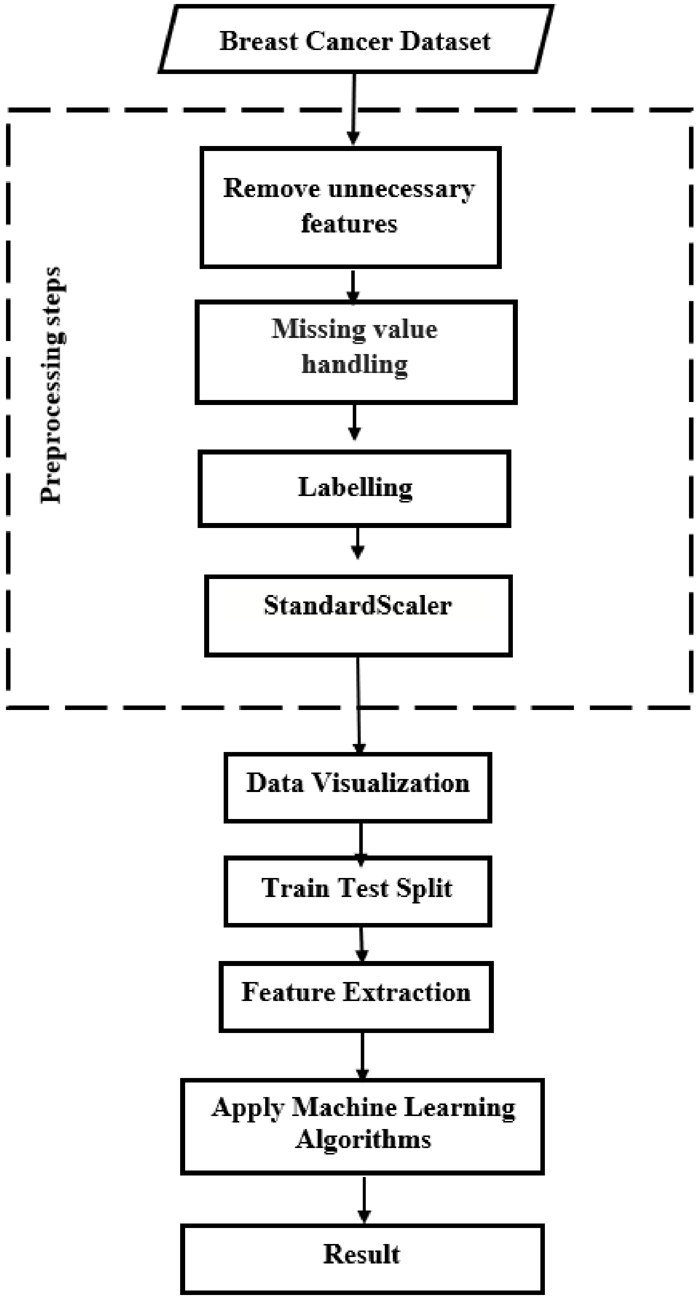


Fig. 1. A typical workflow diagram of our experiments.

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.

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.

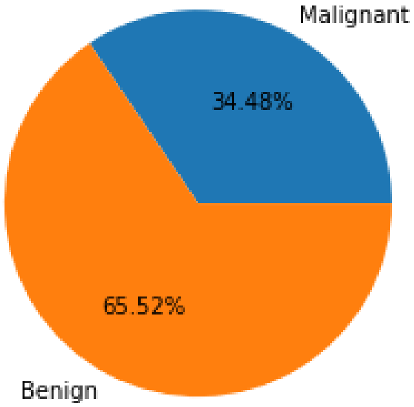


Fig. 2. Statistics of Wisconsin Breast Cancer (Original) (WBC) dataset.

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.

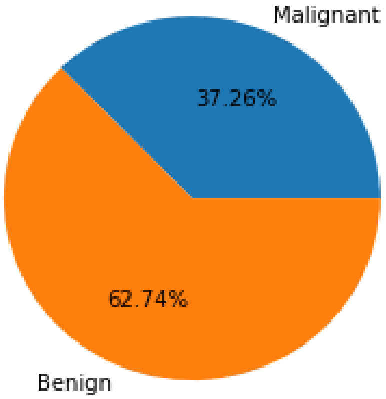


Fig. 3. Statistics of Wisconsin Breast Cancer Diagnosis (WBCD) dataset.

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.

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.

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

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

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

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

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

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

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

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.

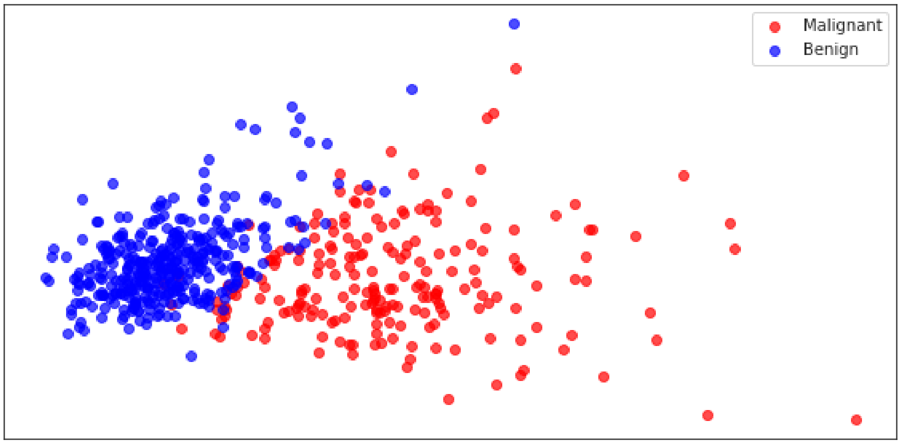


Fig. 4. Represent two classes are almost separated on the WBCD datasets.

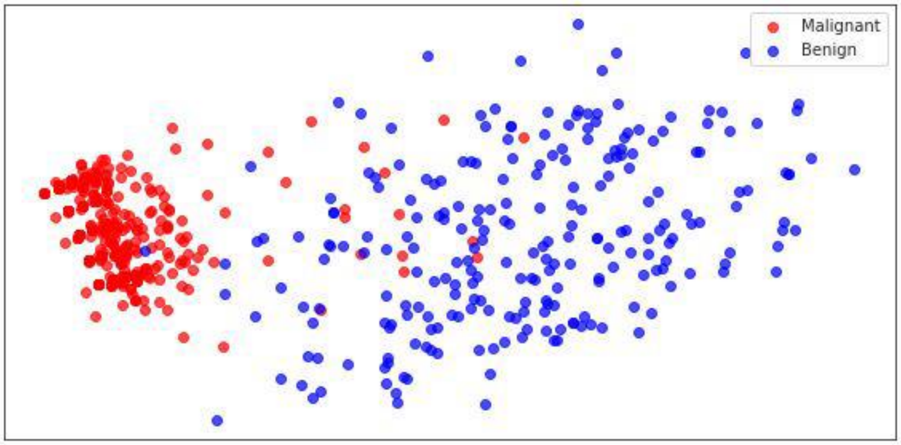


Fig. 5. Represent two classes are almost separated on the WBC datasets.

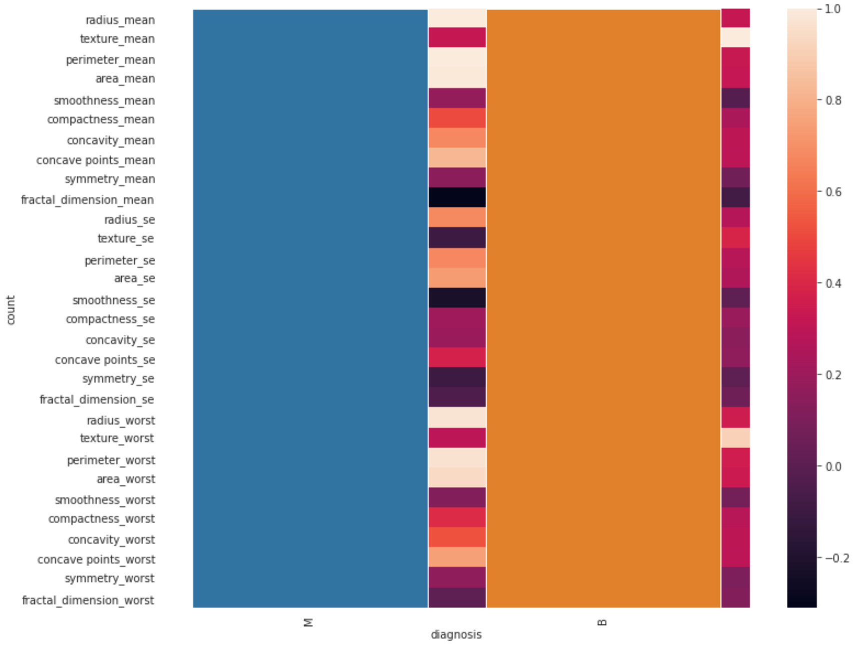


Fig. 6. Heat map for checking correlated features on the WBCD dataset.

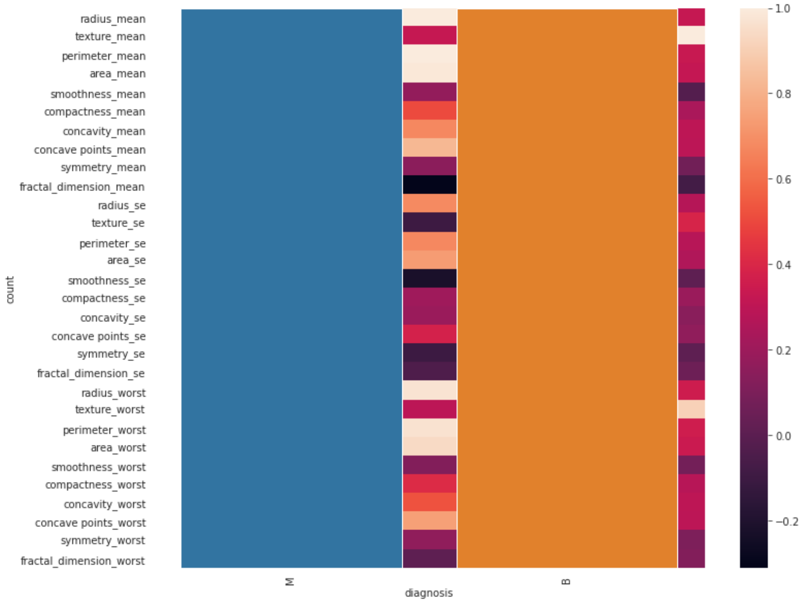
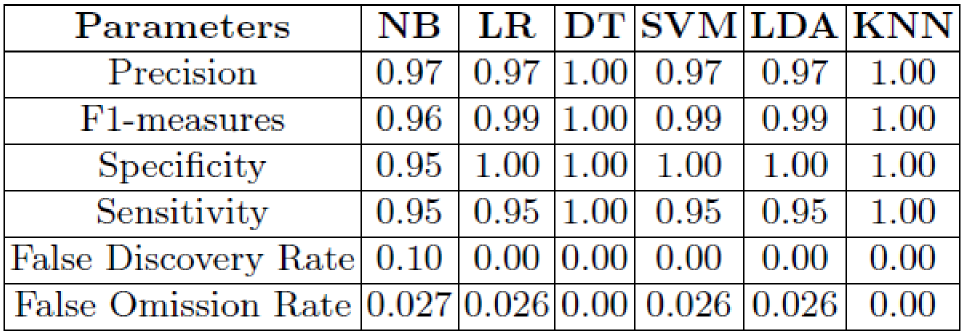


Fig. 7. Heat map for checking correlated features on the WBC dataset.

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 respectively. 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 is represented in Table 1 and Table 2.

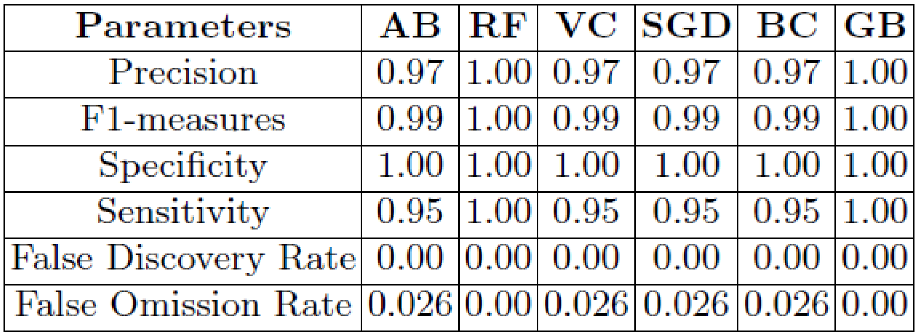
Table 1. Represent the six classifiers performance on WBCD



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

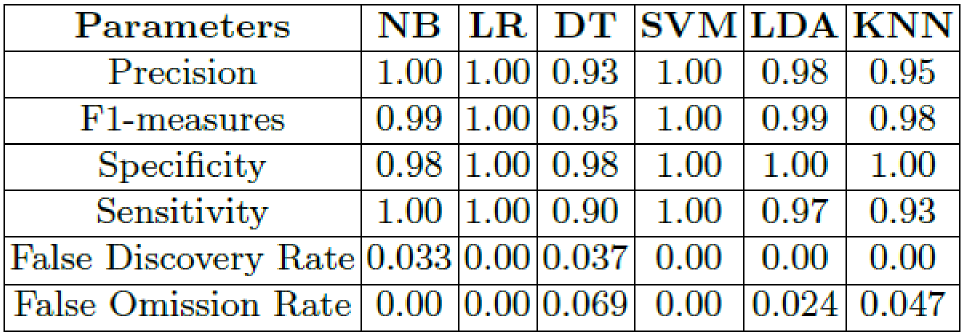
From the Table 1 & 2, it is clearly show that the four (DT, KNN, RF and GB) classifiers achieved the 100% accuracy with 0 false discovering and emission rate that's mean the classifier correctly classify every testing instance. Consequently, others seven algorithms (LR, SVM, LDA, AB, VC, SGD and BC) ensure 98% accuracy with the 1 Specificity. And NB classifier got 94.74% where false discovery rate 0.10.

Table 2. Represent another six classifiers performance on WBCD



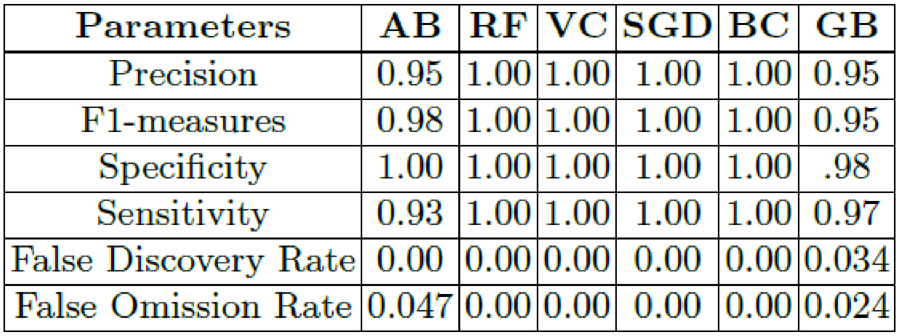
Parameters	AB	RF	VC	SGD	BC	GB
Precision	0.97	1.00	0.97	0.97	0.97	1.00
F1-measures	0.99	1.00	0.99	0.99	0.99	1.00
Specificity	1.00	1.00	1.00	1.00	1.00	1.00
Sensitivity	0.95	1.00	0.95	0.95	0.95	1.00
False Discovery Rate	0.00	0.00	0.00	0.00	0.00	0.00
False Omission Rate	0.026	0.00	0.026	0.026	0.026	0.00

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 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.

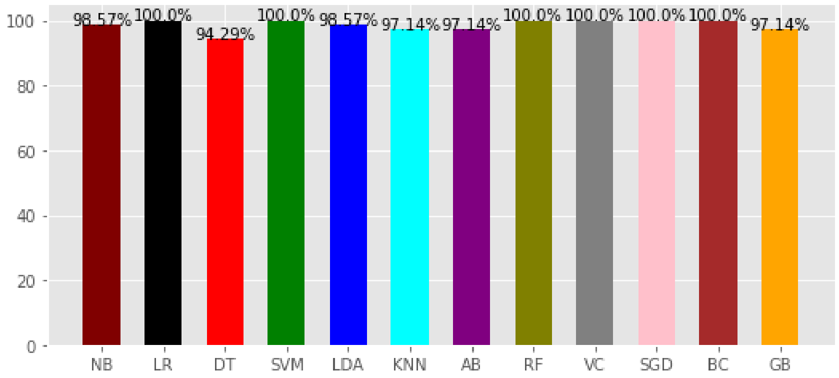


Fig. 8. Results obtained from WBC dataset.

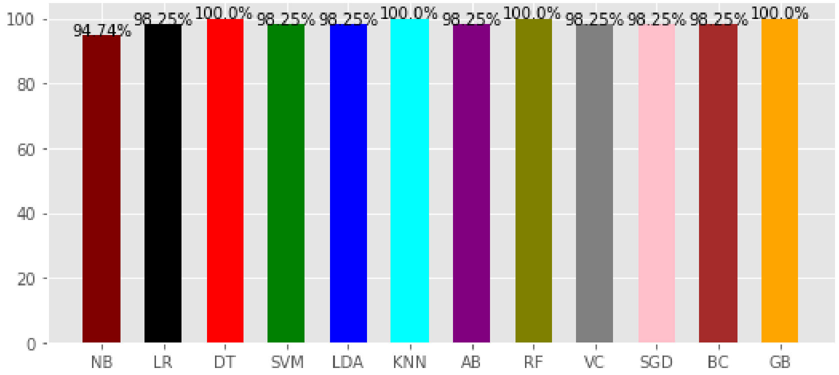


Fig. 9. Result obtained from WBCD dataset.

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.

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.
12 Md. Taslim et al.

[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.
A Comparative Study of Breast Cancer Detection 13

[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.

[22] Siegel, R.L.; Miller, K.D.; Fuchs, H.E.; Jemal, A. Cancer statistics. CA Cancer J. Clin. 2022. [CrossRef]

[23] Byrne, D.; Ohalloran, M.; Jones, E.; Glavin, M. A comparison of data-independent microwave beamforming algorithms for the early detection of breast cancer. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Minneapolis, MN, USA, 3–6 September 2009. [CrossRef]

[24] Samsuzzaman, M.; Islam, M.T.; Shovon, A.; Faruque, R.I.; Misran, N. A 16-modified antipodal Vivaldi Antenna Array for microwave-based breast tumor imaging applications. Microw. Opt. Technol. Lett. 2019, 61, 2110–2118. [CrossRef].

[25] Alqahtani, W.S.; Almufareh, N.A.; Domiaty, D.M.; Albasher, G.; Alduwish, M.A.; Alkhalaf, H.; Almuzzaini, B.; Al-Marshidy, S.S.; Alfraihi, R.; Elasbali, A.M.; et al. Epidemiology of cancer in Saudi Arabia thru 2010–2019: A systematic review with constrained meta-analysis. AIMS Public Health 2020, 7, 679. [PubMed]

[26] M. Brown, S. Goldie, G. Draisma, J. Harford, and J. Lipscomb, "Health service interventions for cancer control in developing countries," in Disease Control Priorities in Developing Countries, pp. 569–590, Oxford University Press, New York, NY, USA, 2nd edition, 2006.

[27] World Cancer Research Fund, "Breast cancer worldwide," http://www.wcrf.org/cancer facts/women-breast-cancer.php/.

[28] Tan, Y. J., K. S. Sim, and F. F. Ting. "Breast cancer detection using convolutional neural networks for mammogram imaging system." In 2017 International Conference on Robotics, Automation and Sciences (ICORAS), pp. 1-5. IEEE, 2017.