Effectiveness evaluation of machine learning algorithms for breast cancer prediction

Abdulrahman Ahmed Jasim[1], Ahmed Adeeb Jalal[1], Nabaa Mohammad Abdulateef[2], Noor Ali Talib[2]
[1]Department Computer Engineering, College of Engineering, Al-Iraqia University, Baghdad, Iraq
[2]Department Network Engineering, College of Engineering, Al-Iraqia University, Baghdad, Iraq

Article Info

Article history:

Received Jan 14, 2022
Revised Mar 4, 2022
Accepted Apr 8, 2022

Keywords:

Breast cancer disease
Supervised machine learning classifiers
Machine learning

ABSTRACT

Breast cancer is becoming a global epidemic, affecting predominantly women. As a result, the number of people diagnosed with breast cancer is increasing every day. As a result, it is critical to have certain early detection methods in place that can assist patients in recognizing this condition at an early stage. Therefore, they might begin taking their medication to prevent the sickness from killing them. Different prediction approaches for early diagnosis of such diseases have been created in the machine learning fields. Those algorithms employ a variety of computational classifiers and claim to achieve satisfactory results in a few areas. However, no research was reached to determine which computationally sophisticated approach is more effective in detecting breast cancer. As a result, it is necessary to select the most effective strategy from the available options. This paper makes a contribution to the performance evaluation of 12 alternative classification strategies on datasets of breast cancer. The right explanations for the classifiers' dominance were investigated.

Corresponding Author:

Abdulrahman Ahmed Jasim
Department of Computer Engineering, College of Engineering, Al-Iraqia University
Baghdad, Iraq
Email: abdulrahman.alsalmany@aliraqia.edu.iq

1. INTRODUCTION

Cancer has become the leading cause of mortality all over the world. Cancer is such a deadly disease that an approximated 10.0 million people died from it in 2020 [1]. To put it another way, cancer causes approximately one in every half-dozen death all over the world. Also, cancer is a general term that refers to various diseases that might affect any part of the body. The quick creation of abnormal cells in the body which develop beyond their natural boundaries and that could subsequently penetrate neighboring parts of the body and spread to other organs is a simple and strong defining trait of cancer. This process is known as metastasizing. Cancer fatalities are frequently caused by metastases [2]. To put it another way, a tumor might be classified as malignant or benign. In addition, a malignant tumor is a cancer type that spreads through the lymph system and blood to other organs and tissues. Lung cancer, colorectal prostate cancer, skin cancer, breast cancer, and stomach cancer are the most prevalent cancers in humans. As stated by the World Health Organization’s (WHO's) International Agency for Research on Cancer (IARC), 2.3 million incidences of breast cancer disease (BCD) were reported in 2020 [1]. Breast cancer is not just a female-only disease; men are also susceptible to it. However, statistics show that females have a higher risk of breast cancer than males.

Female breast cancer has been named the sixth leading cause of death in women (6.6%, 627000 deaths) [3]. As people get older, their chances of developing breast cancer increase dramatically. Early diagnosis of cancer saves lives and reduces treatment costs, according to the renowned great saying, "prevention is better than cure." As a result, many studies worldwide are making extraordinary attempts to combat the disease through developing detection and prediction technologies for effective therapy. Machine learning (ML) approaches are of high importance in disease prediction in this regard. The best method for the labelled data is classification, which is a supervised ML method [4], [5]. Therefore, depending on the test results of patients, the classification approaches might be used for predicting the disease. Various attempts have previously been made to benchmark the precision of classifier results on a variety of disease datasets, but more analysis regarding the classifier performance evaluation on the BCD dataset is still needed. Now, it is clear that BCD prediction is considered as a two-class problem, with malignant and benign classes. Similarly, apart from the class label, the BCD dataset is specified to be quantitative and includes continuous values. This study aims to give a clear picture of the best classifier model among 12 candidates (logistic regression [6], support vector machines (SVM) [7], k-nearest neighbor (k-NN) [8], random forest [9], multi-layer perceptron (MLP) [10], Gaussian Naive Bayes (NB) [11], decision tree [12], MLP regression [13], perceptron [14], linear recognition [15], extreme gradient boosting (XGboosting) [16], and gradient boosting [17]) that might be utilized for predicting the most accurate results with the use of the Breast Cancer Wisconsin (diagnostic) dataset. This work utilized four characteristics to achieve a robust evaluation: accuracy, recall, precision, and F1-score [18], [19].

Various studies have concentrated on breast cancer because it is regarded as one of the major dangerous diseases that have rapidly spread worldwide. Different research was carried out, with varying outcomes that have improved over time. We'll go over a few examples of such research in detail, focusing on the datasets and ML approaches they employed, as well as the accuracy of their findings.

Alghunaim and Al-Baity [20] used 3 classification techniques, which are decision tree (DT), SVM, and random forest (RF), to develop 9 models which aid in breast cancer prediction. They used 3 scenarios with the use of diabetes mellitus (DM), gastric emptying (GE), and GE and DM combined to see which of the 3 forms of data might yield the greatest result with regard to error rate and accuracy. According to the testing results, the scaled SVM classifier in the spark environment outperformed the other classifiers in terms of error rate and accuracy using the GE dataset. A total of 7 supervised ML approaches were evaluated in [21], with regard to precision, accuracy, recall/sensitivity/true positive (TP) rate, specificity, negative predictive value, false-positive rate (FPR), F1-score, rate of misclassification, and receiver operating characteristic (ROC) curve, for the purpose of finding the best model for BCD prediction. Results demonstrate that KNN is considered the best performer on the data-set of the BCD, with a 97% accuracy. Although NB performed similarly to KNN, its precision was not as high as KNN's. With a 94% classification accuracy, the DT classifier was the worst performer. RF, SVM, logistic regression, and ANN all performed in the middle of the NB and DT performance ranges.

Zhang *et al.* [22] describe an unsupervised feature learning framework for identifying various traits from gene expression profiles through combining a principal component analysis (PCA) technique and an autoencoder NN. As the foundation for the collected characteristics, an ensemble classifier based on the AdaBoost algorithm (PCA-AE-Ada) was built. Throughout the studies, they created an additional classifier that used the same classifier learning technique PCA-Ada as the suggested approach, with the only variation being the training inputs. On many independent breast cancer datasets, the suggested approach's area under the receiver operating characteristic curve index, Matthew's correlation coefficient index, accuracy, and other parameters of the evaluation have been tested and put to comparison against the representative gene signature-based algorithms, such as the base-line technique. Experiments show that the suggested technique, which employs deep learning (DL) approaches, and outperforms others. Two of the most common ML approaches were employed to classify the Wisconsin Breast Cancer (original) dataset in [23], while each approach's classification performance was put to comparison with values of precision, accuracy, ROC area, and recall. The SVM approach produced the greatest results with the maximum accuracy.

Naji *et al.* [24] proposed five ML approaches on the breast Cancer Wisconsin Diagnostic dataset: RF, SVM, logistic regression, DT (C4.5), and KNN. After getting the results, the authors performed a performance comparison and evaluation between the 3 classifiers. The major goal of this study was to use ML approaches for predicting and diagnosing breast cancer, and to determine which ones were the most efficient in terms of accuracy, confusion matrix, and precision. SVM was found to outperform all other classifiers and reach the best accuracy of 97.2%. A total of six supervised ML methods are presented in [25], including k-NN, DT, logistic regression, RF, and SVM with radial basis function (RBF) kernel. DL with the use of Adam gradient descent learning has also been used since it combines the advantages of the adaptive gradient approach with root mean square propagation. Each of the models has a distinctive hyper-parametric change that improves accuracy both within the model and when comparing it to other models. DL produces the most precise results with the least amount of loss. DL with Adam gradient descent learning has an accuracy rate of 98.24%.

Arya and Saha [26] suggested gated attentive DL models stacked with RF classifiers to improve breast cancer prognostic prediction using informative features and multi-modal data. It's a bi-phase model: phase one generates stacked features using a sigmoid gated attention convolutional neural network (CNN) and phase two delivers the stacked features to the RF classifier. A comparison of the proposed and other current approaches reveals significant improvements in the estimation of survival of breast cancer patients, with a 5.1% increase in sensitivity values.

Yang *et al.* [27] acquired 287 stages I breast cancer cases and divided them into two groups: test (N=90) and training (N=197). A total of four accessible microarray datasets yielded 14 potential genes. After choosing a superior algorithm, a prediction model has been created utilizing these 14 candidate genes and 3 reference genes whose expressions have been tested with the use of TaqMan probe-based quantitative polymerase chain reaction. When put to comparison with SVM, RF, k-NN algorithms, and the NB algorithm exhibited a greater predictive value (P less than 0.05). This 17-gene model had a strong positive connection with PCR (odds ratio, 8.914, 95% confidence interval, 4.43–17.934, and P 0.001). With the use of this approach, the enrolled patients have been divided into insensitive (INS) and sensitive (SE) groups. The INS and SE groups had significantly different polymerase chain reaction (PCR) rates (42.3% vs. 7.6%, P 0.001). This prediction model's specificity and sensitivity were 62.0% and 84.5%, respectively. Panel gene expression with tens of important genes applied in an ML model provides predictive potential for the chemo-sensitivity of breast cancer rather than the entire transcriptome-based approaches. Our paper examines the behaviour of twelve-candidate classifiers (logistic regression, SVM, RF, Gaussian NB, KNN, MLP regression, MLP, DT, perceptron, linear recognition, XGboosting, and gradient boosting) for the prediction of BCD. In future samples, we want to see which classifiers are the most accurate at predicting breast cancer.

2. PROPOSED METHOD

2.1. Description and distribution of the dataset

The Breast Cancer Wisconsin (diagnostic) dataset was downloaded from the Kaggle machine learning and data science community website (https://www.kaggle.com/uciml/breast-cancer-wisconsin-data) for this empirical investigation. There are a total of 569 instances in this collection, each of which has 33 attributes. There are 212 malignant and 357 benign samples. Table 1 contains a summary of the dataset's features. With regard to each cell nucleus, a total of 10 real-valued characteristics are evaluated: i) texture (standard deviation of grey-scale values), ii radius (average of the distance values from the center to the points on the perimeter), iii) area, iv) perimeter, v) smoothness (local variation in the lengths of the radius), vi) compactness (perimeter2/area-1), vii) concavity (contour's concave portions' severity), viii) concave points (number of the contour's concave portions), ix) asymmetry, and x) fractal dimension ("coast-line approximation"-1).

For every one of the images, the standard error, mean, and "worst" or largest (average of the 3 largest values) regarding such features have been calculated, yielding 30 features. For instance, field 3 represents mean radius, field 13 represents radius SE, and field 23 represents the worst radius. All feature values have 4 significant digits recoded. There are no missing attribute values.

Table 1. Dataset features information

– ID	– radius_worst	– fractal_dimension_mean
– radius_mean	– perimeter_worst	– texture_se
– perimeter_mean	– smoothness_worst	– area_se
– smoothness_mean	– concavity_wors	– compactness_se
– concavity_mean	– symmetry_worst	– concave points_se
– symmetry_mean	– Unnamed	– fractal_dimension_se
– radius_se	– Diagnosis	– texture_worst
– perimeter_se	– texture_mean	– area_worst
– smoothness_se	– area_mean	– compactness_worst
– concavity_se	– compactness_mean	– concave points worst
– symmetry_ se	– concave points mean	– fractal_dimension_se

As demonstrated in Figure 1, we give a determination of whether the variables in the dataset have any correlation in this dataset. After that, as can be seen in Figure 2, we plot the diagnosis result to see if it is malignant=(1) or benign=(0). After that, as indicated in the figures, we display the primary features that are significant in evaluating whether a tumor is malignant or benign: texture mean Figure 3, perimeter mean Figure 4, smoothness mean Figure 5, compactness mean Figure 6, and symmetry mean Figure 7.

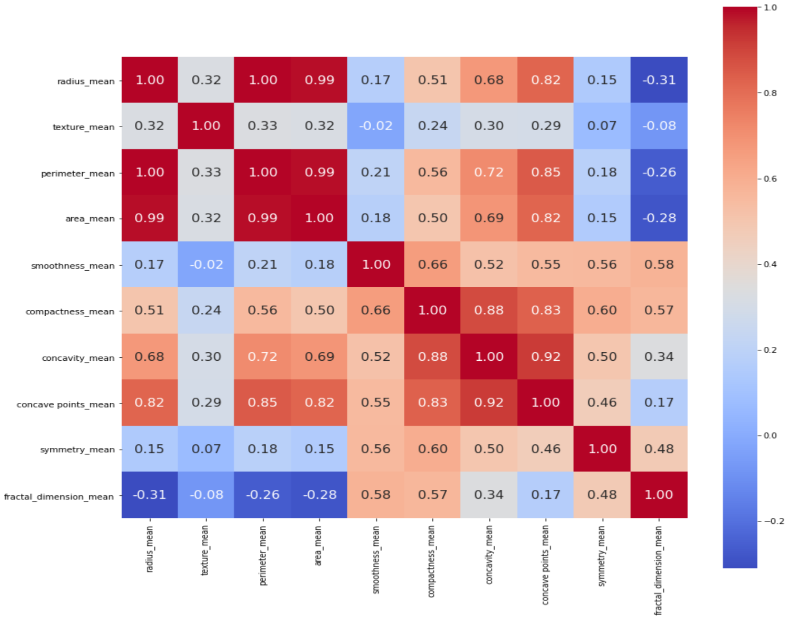
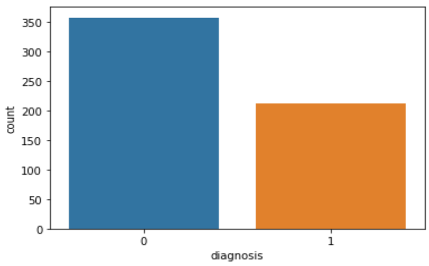


Figure 1. Description of the dataset



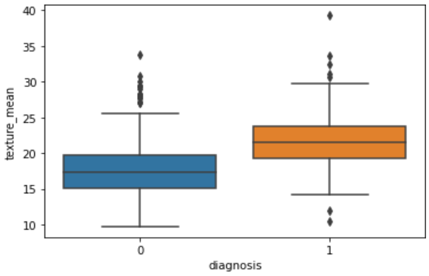
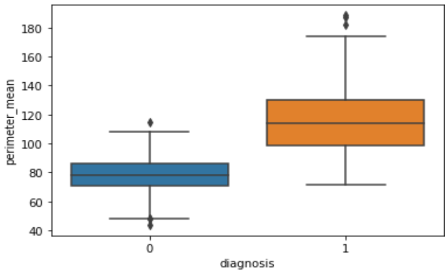


Figure 2. Diagnosis

Figure 3. Texture mean



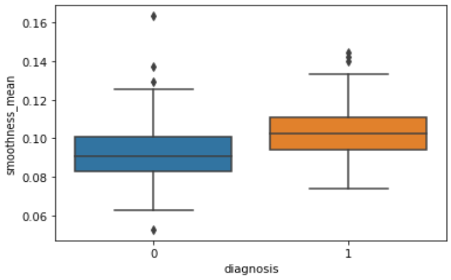
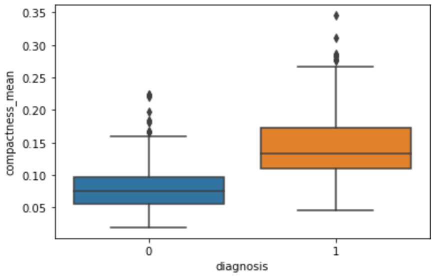


Figure 4. Perimeter mean

Figure 5. Smoothness mean



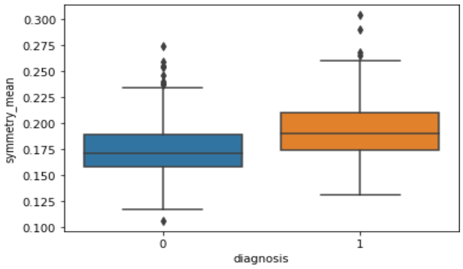


Figure 6. Compactness mean

Figure 7. Symmetry mean

2.2. Description of classification methods

A brief summary of each classification model's method was provided below to provide basic information regarding 12 classifiers. A decision tree classifier can be defined as one of the flowchart structures in which every node represents a test over an attribute value, every one of the branches represents a result of test activity, and the tree's leaf nodes represent classes. It's worth noting that no two root-to-leaf paths must contain identical discrete attributes.

The k-NN model of prediction has been characterized as a lazy learning (no learning) approach-based predictive mechanism that produces predictions depending on the k-NN provided to it. In the case when the prediction of any instance is requested, the whole process of prediction is completed. The Euclidian distance [28] is frequently used to determine closeness.

SVMs are substantially more successful when the dataset has a high number of characteristics. A hyperplane is created with the use of important training tuples to define the data segregation in higher dimensional space, while such training tuples have been treated as support vectors. In brief, the SVMs are based upon the concept of "margin." In addition, a hyperplane will always split two data labels that are on opposite sides of it. The goal is to maximize the margin to create a large enough likely gap between the instances and segregate the hyperplane on each of the sides. The SVMs may be defined as in (1).

$$d(XT) = \sum_{i=1}^{1}. YixiXi|XT + b0 \tag{1}$$

Instead of predictions, the logistic regression model produces probability approximations. With regard to the problem of binary classification, this model is appropriate. The probability of any event occurring has been handled as a linear function of a collection of input features in the model. For determining the actual class label, the logistic regression model calculates p for a linear combination of independent factors. In (2) is a representation of the estimated regression model.

$$P = \frac{e^{\beta 0+\beta 1x1}}{1+e^{\beta 0+\beta 1x1}} \tag{2}$$

The random forest process works by first constructing many DTs and combining them to find steady and strong predictions. A random forest might handle both regression and classification problems efficiently. The perceptron is a parameterized function that takes a real-valued vector as input and creates a Boolean output, as presented in [29]. The output is particularly obtained through thresholding a linear function regarding the input: the Perceptron's parameters are the coefficients of the linear function. Gauss distributions [30] are used to express the likelihoods regarding the features that have been conditioned on classes, a common technique for handling continuous attributes in NB classification. As a result, a Gaussian probability density function (PDF) is used to define each property as in (3):

$$Xi \sim N(\mu, \sigma 2) \tag{3}$$

The Gaussian PDF is shaped like a bell and is specified via the equation:

$$N(\mu, \sigma 2)(X) = \frac{1}{\sqrt{2\pi\sigma\text{^}2}} e^{-\frac{(x-\mu)\text{^}2}{2\sigma\text{^}2}} \tag{4}$$

In which μ represents the mean and σ2 represents the variance. In NB, the parameters required are in the order of O(nk), in which n represent the number of attributes and k represent the number of the classes. Particularly, it is required to specify a normal distribution P(Xi|C) ~N (μ, σ2) for every one of the continuous attributes. The parameters of those normal distributions obtained by (5) and (6):

$$\mu xi|C = c = \frac{1}{Nc}\sum_{i=1}^{Nc} xi \tag{5}$$

$$\sigma 2Xi|C = c = \frac{1}{Nc}\sum_{i=1}^{Nc} xi - \mu\text{^}2 \tag{6}$$

In which Nc represent the number of examples where C=c and N representing the number of the total examples that are utilized for training. Estimating P (C=c) for all classes is easy with the use of the relative frequencies such that:

$$P(C = c) = \frac{Nc}{N} \tag{7}$$

The most frequent and widely used feedforward NN is the MLP network. MLP networks' key processing elements are neurons. In addition, neurons in MLP networks are coupled in a 1-directional manner through connections known as 'weights' [31]. In (8) is used for calculating the MLP network's output:

$$Sj \ = \ \sum_{i=1}^{n}(Wij, Xi) - \theta j, \text{ j= 1,2,…h} \qquad (8)$$

In which Wij represents the weight connecting ith node (in the input layer) to jth- node (in the hidden layer), θj denotes jth bias node (in the hidden layer), and Xi represents input to ith node (in the input layer). The output of every one of the hidden nodes has been estimated with the use of the sigmoid function. The multilayer perceptron regressor (MLPR) technique is a regression-process-specific application of ANNs. It uses the Waikato environment for knowledge analysis (WEKA) optimization class for training a multi-layer perceptron with 1 hidden layer for the minimization of the loss function that has been chosen. MLPR uses logistic functions as activation functions for all of the units except the output one, and it also uses standardization for rescaling the target attribute. Small, regularly distributed random values are used for initializing all network parameters.

The most common one amongst the prediction models for the determination of associations between the variables is linear regression. The notion is linear, despite the fact that the data is multivariate or univariate. Simple linear regression and multiple linear regression are two types of linear regression. In (9) describes the linear regression.

$$Y=x\beta+\varepsilon \qquad (9)$$

Chen and Guestrin [32], suggested XGBoost in 2016. It was identified as an advanced estimator with ultra-high performance in both regression and classification, and it presents significant advantages over typical gradient boosting algorithms. In contrast to gradient boosted decision trees (GBDT), the loss function in the XGBoost has included regularization for preventing overfitting:

$$\mathcal{L}k\big(f(xi)\big) = \sum_{i=1}^{n}\Psi\big(yi, Fk(xi)\big) + \sum_{k=1}^{k}\Omega(fk) \qquad (10)$$

In which, FK (xi) represent the prediction on ith sample at the Kth boost, Ψ (∗) represent a loss function that evaluates differences between the actual and the prediction labels. Ω (fk) represent the term of regularization and could be represented as:

$$\Omega(f) = \Upsilon T + 1/2\lambda||W||2 \qquad (11)$$

In the term of the regularization, γ represents the parameter of complexity as well as complexity of the leaves. T represents the number of the leaves, λ represents the parameter of the penalty, and II ω II 2 represents output of every one of the leaf nodes. In addition to that, different from GBDT, XG-boost adopts a 2nd-order Taylor series as objective function.

$$\mathcal{L}k = \sum_{j=1}^{T}[(\sum i \in IjGi)wj + 1/2\ (\sum i \in Ij\ hi + \lambda)w^{\wedge}2] + \mathcal{Y}T \qquad (12)$$

In which hi and gi represent second- and first-order gradient statistics on loss function, respectively. Assuming that Ij represent the sample set of leaf j.

$$\mathcal{L}k = \sum_{j=1}^{T}[(\sum i \in IjGi)wj + 1/2\ (\sum i \in Ij\ hi + \lambda)w^{\wedge}2] + \mathcal{Y}T \qquad (13)$$

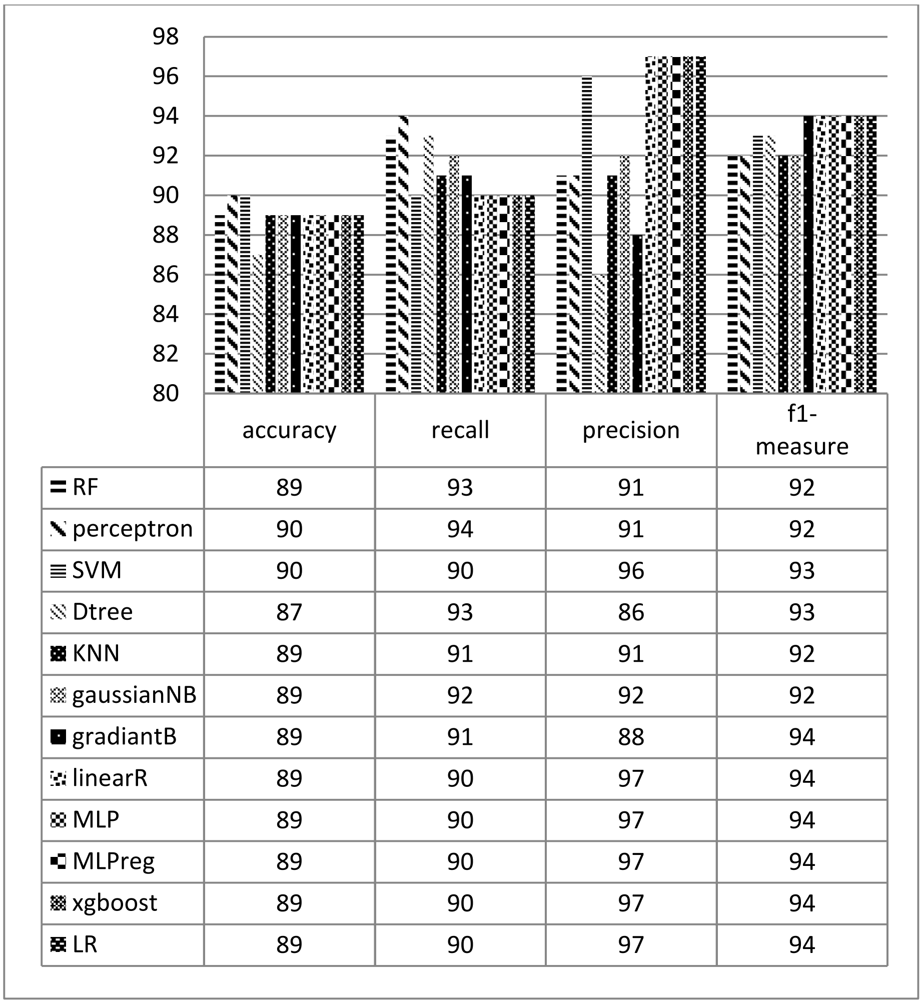
Finally, the objective function is turned into the minimum of a quadratic function determination problem. XGBoost utilizes learning rate, maximum tree depth, boosting numbers, and subsampling for tackling the over-fitting problem, similar to GBDT. Gradient boosting can be defined as an approach that is used as part of an ensemble. This approach integrates various predictors in a sequential manner with certain shrinking. Each one of the iterations regarding the randomly-selected training set is tested against the base model in gradient boosting. Through randomly subsampling the training data, the accuracy and speed of gradient boosting for execution could be increased. As an ensemble technique, gradient boosting can be characterized in as:

$$Y=\mu + \sum_{m=1}^{M} vh(y;X) + e, \qquad (14)$$

To successfully estimate any model's performance, specific performance measures must be developed that may be utilized to assess the goodness of any classifier under evaluation. To evaluate the usefulness of classifiers in this paper, four different performance measures were applied to achieve a robust evaluation: accuracy, recall, precision, and F1-score.

3. RESULTS AND DISCUSSION

The experiment used the Breast Cancer Wisconsin (diagnostic) dataset to diagnose BCD with the use of several classifiers with the following sampling types: tenfold cross-validation, stratified shuffle split, and random samples with a 75% training data-set size. As can be seen in Figure 8, the results from such classifiers: RF, k-NN, gradientB, Gaussian NB, MLP, linear regression, XGboost, MLP regression, and linear regression all had an 89% classification accuracy. While, SVM and perceptron had a higher classification accuracy of 90%, DT had the lowest accuracy of 87%. F1-score that presents combined result of precision and recall ratios, its higher value represents better classification capability of a model, the classifiers-gradientB, MLP, MLP regression, linear regression, XGboost and linear regression show equal F1-score value, i.e., 94%, while RF, perceptron, k-NN, and gaussian NB shows the F1-score value of 92%, and DT, SVM shows the value of 93%.



	accuracy	recall	precision	f1-measure
RF	89	93	91	92
perceptron	90	94	91	92
SVM	90	90	96	93
Dtree	87	93	86	93
KNN	89	91	91	92
gaussianNB	89	92	92	92
gradiantB	89	91	88	94
linearR	89	90	97	94
MLP	89	90	97	94
MLPreg	89	90	97	94
xgboost	89	90	97	94
LR	89	90	97	94

Figure 8. Comparison of performance evaluation

It must be noted that whereas the perceptron and SVM approaches provided higher sensitivity and accuracy in Figure 8, the linear regression generated the best results in terms of F1-score and precision, whereas the DT generated less sensitivity and accuracy. In addition, the KNN had a lower precision and F1-

score. In terms of support, k-NN scores the most, while the gradian boosting, DT scores the lowest. The SVM and perceptron classifiers have been trained on the dataset and obtained the highest performance classifier for the prediction of new cases among all classifiers, according to this comparison.

The performance of the proposed classifiers was evaluated using common validation metrics: accuracy, recall, precision, and F-measure. The separation of related and unrelated components is used in these measurements. The validation findings in Figure 8 show that the papers' categorisation was correctly labeled. The performance of the proposed classifiers using SVM, KNN, NB, and MLP algorithms was compared to the same algorithms with ensemble of filters (EoF) as well as the adaptive mutation enhanced elephant herding optimization+Kernel extreme learning machine (AMEHO+KELM) classifier based on Wisconsin diagnostic Breast Cancer (WDBC) and Wisconsin original Breast Cancer (WOBC) datasets [33]. As it is evident in Figure 9, the MLP classifier achieved outstanding performance in the resultant clusters, compared to the AMEHO+KELM classifier and other classifiers.

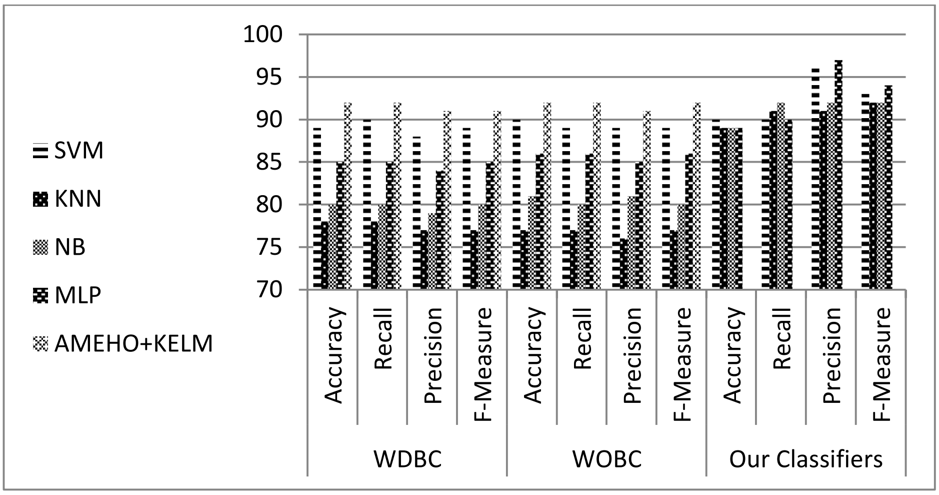


Figure 9. Comparison of performance evaluation with WDBC and WOBC datasets

4. CONCLUSION

Cancer can be defined as a disease that kills a lot of people. One of the various cancer types is the breast cancer. Early identification of cancer not just saves lives, yet also minimizes treatment costs. In the disease's prediction, a reliable prediction system is quite effective. A total of 12 alternative supervised ML algorithms were evaluated in this study (with regard to precision, accuracy, recall/sensitivity/TP rate, and F1-score) so as to discover the best model for BCD prediction. On the Breast Cancer Wisconsin (diagnostic) Dataset, the perceptron and SVM approaches provided higher sensitivity and accuracy, whereas linear regression generated the best results for precision and f1-score, whereas the DT generated less sensitivity and accuracy. In addition, the k-NN had a lower precision and f1-score. The SVM and perceptron classifiers have been trained on the dataset and reached the highest performance classifier for the prediction of new cases among all classifiers, according to this comparison. Perceptron and SVM are the most accurate predictors among the many ML approaches, with accuracy of 90%. Based on recall, accuracy, precision, f1-measure, and other factors, both perceptron and SVM were able to demonstrate their efficiency. The focus of future work will be on developing a better prediction model utilizing ensemble approaches and fine-tuning ensemble techniques for improving model performance.

REFERENCES

[1] H. Sung *et al*., “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” *A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, February 2021, doi: 10.3322/caac.21660.

[2] F. Marazzi *et al*., “Diagnosis and Treatment of Bone Metastases in Breast Cancer: Radiotherapy, Local Approach and Systemic Therapy in a Guide for Clinicians,” *Cancers*, vol. 12, no. 9, p. 2390, August 2020, doi: 10.3390/cancers12092390.

[3] S. A. Rahman, A. Al–Marzouki, M. Otim, N. E. K. Khayat, R. Yousef, and P. Rahman, “Awareness about Breast Cancer and Breast Self-Examination among Female Students at the University of Sharjah: A Cross-Sectional Study,” *Asian Pacific Journal of Cancer Prevention*, vol. 20, no. 6, pp. 1901–1908, 2019, doi: 10.31557/APJCP.2019.20.6.1901.

[4] S. Chowdhury and M. P. Schoen, "Research Paper Classification using Supervised Machine Learning Techniques," *2020 Intermountain Engineering, Technology and Computing (IETC)*, 2020, pp. 1–6, doi: 10.1109/IETC47856.2020.9249211.

[5] A. A. Jasim, l. R. Hazim, and W. D. Abdullah, “Characteristics of Data Mining By Classification Educational Dataset to Improve Student's Evaluation,” *Journal of Engineering Science and Technology*, vol. 16, no. 4, pp. 2825–2844, 2021.

[6] M. F. Ak, “A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications,” *Healthcare*, vol. 8, no. 2, p. 111, April 2020, doi: 10.3390/healthcare8020111.

[7] M. Chen and Y. Jia, "Support Vector Machine Based Diagnosis of Breast Cancer," *2020 International Conference on Communications, Information System and Computer Engineering (CISCE)*, 2020, pp. 321–325, doi: 10.1109/CISCE50729.2020.00071.

[8] M. D. Bakthavachalam and S. A. A. Raj, “A Study Of Breast Cancer Analysis Using KNearest Neighbor With Different Distance Measures And Classification Rules Using Machine Learning,” *European Journal of Molecular & Clinical Medicine*, vol. 7, no. 3, pp. 4842–4851, 2020.

[9] J. Quist, L. Taylor, J. Staaf, and A. Grigoriadis, “Random Forest Modelling of High-Dimensional Mixed-Type Data for Breast Cancer Classification,” *Cancers*, vol. 13, no. 5, p. 991, February 2021, doi: 10.3390/cancers13050991.

[10] M. Desai and M. Shah, “An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN),” *Clinical eHealth*, vol. 4, pp. 1–11, 2021, doi: 10.1016/j.ceh.2020.11.002.

[11] Y. Xiong, M. Ye , and C. Wu, “Cancer Classification with a Cost-Sensitive Naive Bayes Stacking Ensemble,” *Computational and Mathematical Methods in Medicine*, vol. 2021, pp. 1–12, April 2021, doi: 10.1155/2021/5556992.

[12] M. P. Hendriks *et al.*, “Clinical decision trees support systematic evaluation of multidisciplinary team recommendations,” *Breast Cancer Research and Treatment*, vol. 183, no. 2, pp. 355–363, July 2020, doi: 10.1007/s10549-020-05769-1.

[13] B. Al-Shargabi, B. Al-Shami, and R. S. Alkhawaldeh, “Enhancing Multi-Layer Perceptron for Breast Cancer Prediction,” *International Journal of Advanced Science and Technology*, vol. 130, pp. 11–20, October 2019.

[14] A. López-Cortés *et al.*, “Prediction of breast cancer proteins involved in immunotherapy, metastasis, and RNA-binding using molecular descriptors and artificial neural networks,” *Scientific Reports*, vol. 10, no. 8515, pp. 1–13, May 2020, doi: 10.1038/s41598-020-65584-y.

[15] M. Mohammed, H. Mwambi, I. B. Mboya, M. K. Elbashir, and B. Omolo, “A stacking ensemble deep learning approach to cancer type classification based on TCGA data,” *Scientific Reports*, vol. 11, no. 15626, pp. 1–22, August 2021, doi: 10.1038/s41598-021-95128-x.

[16] X. Y. Liew, N. Hameed, and J. Clos, “An investigation of XGBoost-based algorithm for breast cancer classification,” *Machine Learning with Applications*, vol. 6, p. 100154, December 2021, doi: 10.1016/j.mlwa.2021.100154.

[17] A. Derangula, S. R. Edara, and P. K. Karri, “Feature Selection of Breast Cancer Data Using Gradient Boosting Techniques of Machine Learning,” *European Journal of Molecular & Clinical Medicine*, vol. 7, no. 2, pp. 3488–3504, 2020.

[18] J. Han, M. Kamber, and J. Pei, “*Data Mining: Concepts and Techniques (3rd Edition)*,” Elsevier, 2011.

[19] H. Bhavsar and A. Ganatra, “A Comparative Study of Training Algorithms for Supervised Machine Learning,” *International Journal of Soft Computing and Engineering*, vol. 2, no. 4, pp. 74–81, September 2012.

[20] S. Alghunaim and H. H. Al-Baity, "On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context," in *IEEE Access*, vol. 7, pp. 91535–91546, 2019, doi: 10.1109/ACCESS.2019.2927080.

[21] V. Kumar, “Evaluation of computationally intelligent techniques for breast cancer diagnosis,” *Neural Computing and Applications*, vol. 33, no. 8, pp. 3195–3208, 2021, doi: 10.1007/s00521-020-05204-y.

[22] D. Zhang, L. Zou, X. Zhou, and F. He, "Integrating Feature Selection and Feature Extraction Methods With Deep Learning to Predict Clinical Outcome of Breast Cancer," in *IEEE Access*, vol. 6, pp. 28936–28944, 2018, doi: 10.1109/ACCESS.2018.2837654.

[23] E. A. Bayrak, P. Kırcı, and T. Ensari, "Comparison of Machine Learning Methods for Breast Cancer Diagnosis," *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, 2019, pp. 1–3, doi: 10.1109/EBBT.2019.8741990.

[24] M. A. Naji, S. El Filali, K. Aarika, EL H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, ”Machine learning algorithms for breast cancer prediction and diagnosis,” *Procedia Computer Science*, vol. 191, pp. 487-492, 2021, doi: 10.1016/j.procs.2021.07.062.

[25] P. Gupta and S. Garg, “Breast Cancer Prediction using varying Parameters of Machine Learning Models,” *Procedia Computer Science*, vol. 171, pp. 593–601, 2020, doi: 10.1016/j.procs.2020.04.064.

[26] N. Arya and S. Saha, “Multi-modal advanced deep learning architectures for breast cancer survival prediction,” *Knowledge-Based Systems*, vol. 221, p. 106965, June 2021, doi: 10.1016/j.knosys.2021.106965.

[27] L. Yang *et al.*, “Prediction model of the response to neoadjuvant chemotherapy in breast cancers by a Naive Bayes algorithm,” *Computer Methods and Programs in Biomedicine*, vol. 192, p. 105458, August 2020, doi: 10.1016/j.cmpb.2020.105458.

[28] R. Ehsani and F. Drabløs, “Robust Distance Measures for kNN Classification of Cancer Data,” *Cancer Informatics*, vol. 19, pp. 1–9, October 2020, doi: 10.1177/1176935120965542.

[29] X. Wang and M. Benning, “Generalised Perceptron Learning,” *Proceedings of the 12th OPT Workshop on Optimization for Machine Learning*, 2020, doi: 10.48550/arXiv.2012.03642.

[30] J. S. Angarita-Zapata, G. Maestre-Gongora, and J. F. Calderín, “A Bibliometric Analysis and Benchmark of Machine Learning and AutoML in Crash Severity Prediction: The Case Study of Three Colombian Cities,” *sensors*, vol. 21, no. 24, p. 8401, December 2021, doi: 10.3390/s21248401.

[31] I. Aljarah, H. Faris, and S. Mirjalili, “Optimizing connection weights in neural networks using the whale optimization algorithm,” *Soft Computing*, vol. 22, no. 1, pp. 1–15, 2018, doi: 10.1007/s00500-016-2442-1.

[32] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2016, pp. 785–794, doi: 10.48550/arXiv.1603.02754.

[33] R. S. P. Priyaa and P. S. Vadivub, “A Novel Adaptive Mutation Enhanced Elephant Herding Optimization (Ameho) Based Feature Selection And Kernel Extreme Learning Machine (Kelm) Classifier For Breast Cancer Diagnosis,” *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 13, pp. 1198–1216, 2021.

BIOGRAPHIES OF AUTHORS

Abdulrahman Ahmed Jasim received the Engineer degree in Computer Engineering from Dijlah University, Iraq in 2012. He received the Master degree in Elecrtrical and Computer Engineering from Altinbas University, Turkey in 2018. Currently, he is a lecturer of Computer Engineering Department, College of Engineering, Al-Iraqia University, Iraq. Now, he’s a Phd student at Altinbas University, Turkey. His research interests include data mining, Machine learning, and Deep learning. He can be contacted at email: abdulrahman.alsalmany@aliraqia.edu.iq.

Ahmed Adeeb Jalal received the Engineer degree in Software Engineering from Al-Rafidain University College, Iraq in 2002. He received the Master degree in Computer Engineering from Yildiz Technical University, Turkey in 2016. Currently, he is a lecturer of Computer Engineering Department, College of Engineering, Al-Iraqia University, Iraq. His research interests include data mining, hybrid recommendation systems design, and web applications. He can be contacted at email: ahmedadeeb@aliraqia.edu.iq.

Nabaa Mohammad Abdulateef received the Engineer degree in Network Engineering from Al-Iraqia University, Iraq in 2021. Currently, she is a data center engineer at Earthlink company, Iraq. Her research interests include programming scripts in security networks field and improving algorithms using data mining. She can be contacted at email: Nabaama99@gmail.com.

Noor Ali Talib received the Engineer degree in Network Engineering from Al-iraqia University, Iraq in 2021. Currently, she is working in the field of Information and Communication Technologies at Dijlah University, Iraq. Her research interests include information technologies, network communications, and data mining. She can be contacted at email: 99nooraali@gmail.com.