Computer Systems Science & Engineering
DOI: 10.32604/csse.2023.033003
Article

Tech Science Press

Breast Cancer Diagnosis Using Feature Selection Approaches and Bayesian Optimization

Erkan Akkur[1], Fuat TURK[2,*] and Osman Erogul[1]

[1]Department of Biomedical Engineering, TOBB University of Economics and Technology, Ankara, 06560, Turkey
[2]Deparment of Computer Engineering, Karatekin University, Çankırı, 18100, Turkey
*Corresponding Author: Fuat TURK. Email: fuatturk@karatekin.edu.tr

Received: 06 June 2022; Accepted: 13 July 2022

Abstract: Breast cancer seriously affects many women. If breast cancer is detected at an early stage, it may be cured. This paper proposes a novel classification model based improved machine learning algorithms for diagnosis of breast cancer at its initial stage. It has been used by combining feature selection and Bayesian optimization approaches to build improved machine learning models. Support Vector Machine, K-Nearest Neighbor, Naive Bayes, Ensemble Learning and Decision Tree approaches were used as machine learning algorithms. All experiments were tested on two different datasets, which are Wisconsin Breast Cancer Dataset (WBCD) and Mammographic Breast Cancer Dataset (MBCD). Experiments were implemented to obtain the best classification process. Relief, Least Absolute Shrinkage and Selection Operator (LASSO) and Sequential Forward Selection were used to determine the most relevant features, respectively. The machine learning models were optimized with the help of Bayesian optimization approach to obtain optimal hyperparameter values. Experimental results showed the unified feature selection-hyperparameter optimization method improved the classification performance in all machine learning algorithms. Among the various experiments, LASSO-BO-SVM showed the highest accuracy, precision, recall and F1-score for two datasets (97.95%, 98.28%, 98.28%, 98.28% for MBCD and 98.95%, 97.17%, 100%, 98.56% for MBCD), yielding outperforming results compared to recent studies.

Keywords: Breast cancer; machine learning; Bayesian optimization; feature selection

1 Introduction

Breast cancer (BC) have been considered as the most diagnosed malignant disease among females in recent years [1]. Abnormal growths of some cell in the breast tissue can cause breast cancer. Benign and malign tumors are abnormally growing cells in the breast tissue. Benign tumor is noncancerous which can be treated with medicine or surgery. However, malignant tumor shows the characteristics of cancer. When malignant tumor is not treated, it can rapidly spread to organs and increase the mortality rates. Therefore, early diagnosis of BC is a necessary step to reduce the mortality rates [2]. Methods such as radiologic imaging and pathological examinations have been used for early diagnosis [3]. Mammography is the most preferred imaging tool for early diagnosis [4]. However, the detection of suspicious breast tumors on mammogram images can be challenging. When mammography detects a suspicious breast tumor, breast biopsy is generally performed [5]. Although breast biopsies provide a definite diagnosis, the statistics show that most biopsies are benign [6,7]. Thus, many studies have been carried out for efficient early diagnosis and to minimize the unnecessary biopsy rates in the literature. Due to the high successful classification rates, machine learning algorithms have been promising approach for the prediction of BC in recent years [8,9]. Machine learning (ML) approaches are a type of artificial intelligence that allow software applications to obtain higher accuracy of predicting outcomes. These algorithms use past values as input to predict new outputs [10]. Feature selection (FS) and hyperparameter optimization (HO) are two critical issues to improve the classification rates of ML. FS is the process of determining the suitable features to be used in machine learnings approaches. These methods decrease overfitting, improve the performance of classification process [11,12]. HO is a process in finding optimal hyperparameters for the training of ML. Such hyperparameters need to be set carefully to provide excellent classification performance. These hyperparameters are very important when creating and evaluating machine learning algorithms [13]. Although using ML to predict BC has been an active field of research, there are a few studies combining FS and HO [14–16]. In view of this, this manuscript proposes a new BC classification model based on improved ML algorithms integrating FS and HO. First, this paper is designed to explore the effects of FS and HO on ML algorithms. Second, the best combined algorithm identified in this manuscript can be used in future studies for efficient diagnosis of BC.

The summary of this article and its contribution to science is given below:

1. Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (K-NN) and Ensemble Learners (EL) were used to classify malign and benign breast lesions.
2. Three different feature selection methods namely, Relief (RF), LASSO and Sequential Forward Selection (SFS) were used to determine the most selective and discriminative features for effective identification of BC.
3. Bayesian optimization (BO) algorithm was utilized for optimizing the classification algorithms.
4. The statistical measures (accuracy, precision, recall and F1-Score) were used to measure the performance of suggested classification model which was implemented in MATLAB software.
5. A series of comparative analyses are performed on the Wisconsin Breast Cancer Dataset (WBCD) which was retrieved from the UCI machine learning repository and Mammographic Breast Cancer Dataset (MBCD) which has never been used before.

The rest of the study is presented as follow: The literature studies are summarized in Section 2. The general structure of methods and methods are given in Section 3. The results are demonstrated in Section 4. The results are discussed in Section 5 and Section 6 presents the conclusion.

2 Literature Survey

Numerous studies have been investigated to provide early diagnosis of BC in recent years. There are various ML algorithms that are being used for this purpose [17,18]. Studies of BC prediction, SVM, DT, NB, K-NN and EL algorithms were comparatively used [10]. Asri et al. [19] compared several classification algorithms such as K-NN, DT, NB and SVM. The experiments were conducted on WBCD. The study achieved the highest accuracy of 97,13% using SVM. Khan et al. [20] proposed comparative analysis based on machine learning models for BC prediction. The experiments were conducted on WBCD. Among the various classifiers, Logistic regression achieved the best accuracy of 98%.

In the literature, many feature selection methods used to select optimal features. Three techniques are used in FS: filter, wrapper and embedded. [11]. Filter approaches use the statistical functions to choose and rank the feature subsets. Recently, Relief, Minimum-redundancy maximum relevancy and Fisher have been suggested as the most convenient filter methods due to simple ranking strategies of these algorithms [21]. Tian et al. [22] compared ten filter feature selection methods for BC prediction. They used the digital database for screening mammography where Relief with 6 features (AUC:0.855) showed good classification rates. Wrapper method uses a specific selection criterion to determine the quality of various feature subsets. Sequential forward and backward feature selection [23] and metaheuristic algorithms such as genetic algorithm [24] and particle swarm optimization [25] have been extensively used as wrapper algorithm for BC prediction. Dhanya et al. [26] compared the FS and ML algorithms for BC prediction. The experiments were conducted on WBCD. The experiments showed SFS-NB showed the highest accuracy with 98.24%. Embedded feature selection utilized a learning algorithm in order to select features [11]. One representative embedded feature selection method is the LASSO [27]. Albaldawi et al. [27] proposed hybrid model ANOVA-LASSO methods and classification algorithms for microarray data classification. Three different classification algorithms like Linear Support Vector Classifier (LSVC), Random Forest (RF) and Multilayer Perceptron Classifier (MLP) were used. The proposed model demonstrated accuracy of 100% when using all classifiers.

Machine learning contains many hyperparameters and these hyperparameters need to be set automatically to optimize the performance. HO is an approach that chooses a set optimal hyperparameters for a machine learning algorithm. In the literature, grid search, random search, and Bayesian optimization have been used to automatically set hyperparameters in machine learning [28]. Compared to hyperparameter optimization models, BO uses less time with smaller evaluations to find the best hyperparameter values. Kumar et al. [29] proposed an efficient model for BC using Bayesian hyperparameter tuned RF. The study used WBCD and Wisconsin Prognostic Breast Cancer Dataset (WPBC). Their proposed approach achieved an accuracy of 96.4% for WBCD and 97.9% for WPBC. Mate et al. [30] presented a hybrid model that combined FS and BO with machine learning for BC prediction. Extra Tree Classifier algorithms was the best classification method with accuracy of 96,2%. Bensaoucha et al. [31] compared several classification algorithms for BC prediction. The hyperparameters of the algorithms were determined using BO approach. SVM showed the highest accuracy with 96.52% for WDBC dataset.

3 Materials and Methods

The suggested classification model is illustrated for prediction of BC in Fig. 1.

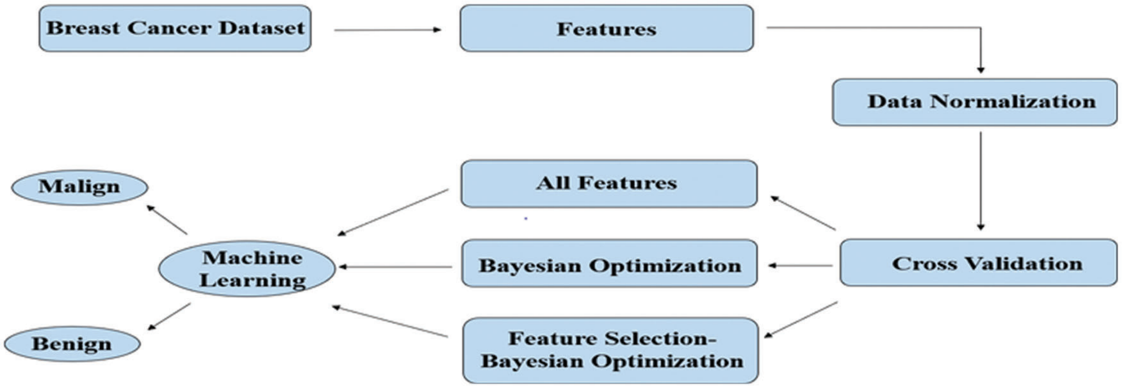

Figure 1: Overall workflow of proposed classification model

The working principle of suggestion model is shown Algorithm 1.

Algorithm 1

```
1. Loading the breast cancer dataset
2. Defining features of dataset
3. Data normalization of features
4. Data splitting using Cross-Validation
5. Classification in different three strategies
  5.1. All features are directly given to as input ML Algorithms
  5.2. All features are given to as input ML Algorithms with BO
      5.2.1. BO procedure steps are shown in below
            a. Constructing a surrogate probability model representing the objective function
            b. Defining of the best hyperparameters using the surrogate probability model
            c. Applying of hyperparameters to the objective function
            d. Updating of surrogate probability model
            e. Repeating the above steps until the maximum iteration is reached.
  5.3. Selecting features are given to as input to ML with FS-BO method
      5.3.1. FS-BO method procedure steps are shown in below.
            a. Selecting the most discriminating features using FS methods.
            b. Applying 5.2.1 steps again.
6. Discriminating of malign-benign breast lesions.
```

3.1 Description of the Dataset

The first dataset is the WBCD that consists of 569 instances with 32 features. The dataset contains 32 features (ID Number, Diagnosis and 30 input features). The features extracted from images of cell nuclei. Each instance is labeled as benign and malign. There are 356 benign and 213 instances [32]. The feature details are shown in Tab. 1.

The second dataset is the MBCD which includes a total of 195 breast tumors (116 images (59%) for malign, 79 (41%) images for benign). This dataset was a retrospective study, and it was retrieved from Ankara Training and Research Hospital. This retrospective study was approved by the Institutional Ethics Committee of Ankara Training and Research Hospital (319/E-20). All patients who underwent digital mammography between April 2015 and April 2020 were retrieved from the Picture Archiving and Communication System (PACS). All patients underwent mammography using IMS Giotto (Bologna-Italy). Patient consent was obtained on the condition that all data were anonymized. The mammogram images were subject to segmentation process to determine the region of interest (ROI) which represents breast tumors. The process of extraction of ROI is shown in Fig. 2. First, the mammogram images were retrieved (Fig. 2a). Then, the ROIs were defined manually by a physician with green contour (Fig. 2b). Finally, using the gray level thresholding and morphological operations, the ROIs were extracted from the image (Fig. 2c).

Table 1: Features of WBCD

No	Features	No	Features
1	Radius mean	16	Compactness severity
2	Texture mean	17	Concavity severity
3	Perimeter mean	18	Concave points severity
4	Area mean	19	Symmetry severity
5	Smoothness mean	20	Fractal simension severity
6	Compactness mean	21	Radius worst
7	Concavity mean	22	Texture worst
8	Concave points mean	23	Perimeter worst
9	Symmetry mean	24	Area worst
10	Fractal dimension mean	25	Smoothness worst
11	Radius severity	26	Compactness worst
12	Texture severity	27	Concavity worst
13	Perimeter severity	28	Concave points worst
14	Area severity	29	Symmetry worst
15	Smoothness severity	30	Fractal dimension worst

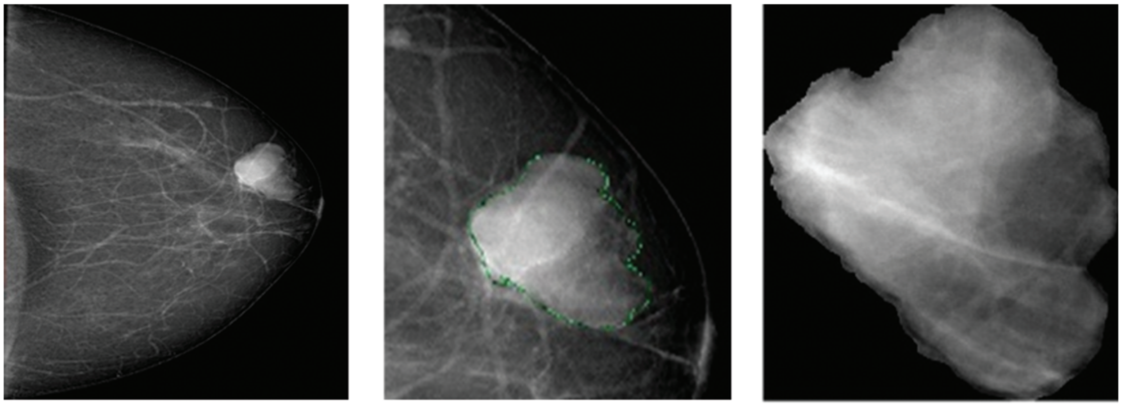

Figure 2: (a) Original mammogram image, (b) marking of ROI by a physician with green contours, (c) extracted ROI.

A total of 54 shape and texture features were calculated for each ROI. Intensity, Grey-level Co-occurrence Matrix (GLCM) and Gray Level Run Matrix (GLRM) were used to generate texture features. 16 shape features (F1–F16), 15 intensity-based features (F17–F31), 13 GLCM (F32–F43) and 11 GLRM (F44–F54) features were calculated [33–36]. The features are demonstrated in Tab. 2. These, Mean Absolute Deviation (MAD), Inter Quartile Range (IQR), Root Mean Square (RMS), Short Run Emphasis (SRE), Long Run Emphasis (LRE), Gray Level Non-Uniformity (GLNU), Run Length Nonuniformity (RLN), Run Percentage (RP), Low Gray-Level Run Emphasis (LGRE), High Gray-Level Run Emphasis (HGRE), Short Run Low Gray-Level Emphasis (SRLGE), Short Run High Gray-Level Emphasis

(SRHGE), Long Run Low Gray-Level Run Emphasis (LGRE), Long Run Low Gray-Level Run Emphasis (LGHE) can be expressed.

Table 2: Features of MBCD

No	Features	No	Features	No	Features
1	Area	19	Variance	37	Sum of mean
2	Perimeter	20	Smoothness	38	Sum of variance
3	Max. Radius	21	Skewness	39	Sum of entropy
4	Min. Radius	22	Kurtosis	40	Difference variance
5	Euler Number	23	MAD	41	Difference entropy
6	Eccentricity	24	Minimum	42	Information measure of correlation 1
7	Solidity	25	Maximum	43	Information measure of correlation 2
8	Entropy	26	10_{th} Percentile	44	SRE
9	Equiv. Diameter	27	90_{th} Percentile	45	LRE
10	Elongatedness	28	IQR	46	GLNU
11	Circulation 1	29	Range	47	RLN
12	Circulation 2	30	RMS	48	RP
13	Compactness	31	Median	49	LGRE
14	Dispersion	32	Contrast	50	HGRE
15	Thinness ratio	33	Correlation	51	SRLGE
16	Shape index	34	Energy	52	SRHGE
17	Mean	35	Homogeneity	53	LRLGE
18	Std. Deviation	36	Sum of Square	54	LRHGE

3.2 Data Normalization

Data normalization is a preprocessing technique that aims to identify numeric values in the datasets within a fixed range. In this study, z-score normalization method was used. Z-score is a technique that represents the number of standard deviations away from the mean [37].

3.3 Feature Selection

FS is used to eliminate redundant and irrelevant features. Removing the irrelevant features improves the machine learning classification performance and reduces the computational cost of modeling [11]. RF is a filter approach that weighs feature according to their relationships. This algorithm uses similar methodology as in K-NN to determine the weights. The most important features get high weights while the remaining features get small weights. All features were ranked according to this measure. Here we selected 0 as a threshold value. If the weights of the features were higher than 0, the feature was selected, otherwise the feature is ignored [38,39]. LASSO is an embedded feature selection approach where the selected feature depends on modifying the absolute coefficient values. These features coefficient values set to zero and zero coefficient features are removed from the feature set. The selected features have high coefficient [40]. SFS is a wrapper-based feature selection method it adds features from the dataset sequentially to an empty candidate set. This process continues until further addition features does not reduce the criterion. The candidate feature subsets were evaluated by using 10-fold cross validation [41].

3.4 Cross Validation

Cross validation is a resampling model that divides into two groups; training and testing. 10-fold cross validation is used for evaluation of models. 90% of data were used for training, while 10% of data were used for testing purpose [42].

3.5 Machine Learning Algorithms

In this study, DT, NB, SVM, K-NN and EL algorithms were used for classification. In the DT approach, simple decision rules are used to estimate the value of target. This simple decision rules are extracted from the data. DT is generally used classification and regression process [43]. NB algorithm tries to estimate the class of the data by using "Bayes Theorem". NB algorithm uses a series of probability principles to determine the class of the data [44]. SVM builds a line or a hyperplane that splits the data into classes. It provides a good accuracy value while performing prediction of large datasets [45]. K-NN uses feature similarity to estimate the values of new sample points [46]. EL is a machine learning model that combines two or more configuration algorithms such as bagging and boosting techniques. Bagging and boosted approaches are generally based on decision tree learners. Bagging model aims to achieve better success by using and combining a set of classifiers. Boosting algorithm aims to obtain strong classifiers from the weak classifiers with a low training error [47].

3.6 Hyperparameter Optimization

Hyperparameters are very important effect on performance of ML because they directly affect the training process. For example, box-constraint, kernel parameter and kernel scale are very important for SVM. Moreover, maximum number of splits affect the performance of decision tree. These hyperparameters are essential to be set to obtain excellent result. HO provides automation of the selection of hyperparameter values [48].

In this manuscript, BO was used to select automatically hyperparameters for machine learning algorithm. This algorithm is an effective approach for parameter search and is a black-box optimization technique. Algorithm builds a probabilistic model by setting a prior probability distribution over the function being optimized. Then, it combines with sample information to obtain a posterior function [49,50]. Hyperparameters and search ranges of ML are demonstrated in Tab. 3.

Table 3: The hyperparameters of machine learning algorithms

Algorithm	Hyperparameters	Search range
DT	'Maximum number of splits'	[1–568]
	'Split Criterion'	Gini's diversity index, Maximum deviance reduction
NB	'Distribution names'	Gaussian, Kernel
	'Kernel Type'	Gaussian, Box. Epanechniko, Triangle
SVM	'Kernel Function'	Gaussian, Linear, Quadratic, Cubic
	'Kernel Scale'	[0.001–1000]
	'Box Constraint Level'	[0.001–1000]
	'Standardize data'	True/False

(Continued)

Table 3 (continued)

Algorithm	Hyperparameters	Search range
EL	'Ensemble Method'	Bag, GentleBoost, LogitBoost, AdaBoost, RUSBoost
	'Number of Learners'	[10–500]
	'Learning Rate'	[0.001–1]
	'Maximum number of splits'	[1–568]
K-NN	'Number of Neighbors'	[1–285]
	'Distance Metric'	City Block, Chebyshev, Correlation, Euclidian, Hamming, Jaccard, Mahalonobis, Minkowski, Spearmen

3.7 Performance Evaluation Metrics

Confusion matrix is used to visualize the success of ML models. True positive (TP), True negative (TN), False positive (FP) and False negative (FN) values need to be identified to measure the confusion matrix. TP means that malign cases are properly recognized as malign. TN means that benign cases are properly identified as benign. FP means that benign cases are mistakenly identified as malign. FN means that benign cases are mistakenly recognized as malign. Accuracy is the proportion of the cases correctly identified to entire cases. Precision is calculated by dividing true positive by overall positives. Recall is defined that the percentages of true positive among the real positive cases. F1-Score demonstrates the harmonic mean of precision and recall values [51].

$$\mathrm{Accuracy} = ((\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})) * 100 \tag{1}$$

$$\mathrm{Recall} = ((\mathrm{TP})/(\mathrm{TP} + \mathrm{FN})) * 100 \tag{2}$$

$$\mathrm{Precision} = ((\mathrm{TP})/(\mathrm{TP} + \mathrm{FP})) * 100 \tag{3}$$

$$\mathrm{F1} - \mathrm{Score} = 2 * (\mathrm{Precision} * \mathrm{Recall})/(\mathrm{Precision} + \mathrm{Recall})) \tag{4}$$

4 Experimental Results

Different experiments were implemented on machine learning methods to achieve the best classification rates for BC datasets. Experiments first started by using the functions developed by the MATLAB 2020a program. Then, by using all the features, MATLAB Statistics and Machine Learning Toolbox program [52] was used and the classification process was implemented by optimizing the features of machine learning methods with Bayesian Optimization technique. Finally, distinctive features were determined in the datasets using RF, LASSO and SFS methods, and the classification process was implemented by optimizing the features of machine learning methods with Bayesian Hyper Optimization technique. To apply FS, relief, lasso and sequential functions developed by MATLAB were used [53]. The selecting features for the BC datasets are demonstrated in Tab. 4. After applying FS, the selecting features were 16, 8 and 3 for MBCD, and they were 12, 10 and 4 for RF, LASSO and SFS, for WBCD, respectively.

Table 4: Selected features after feature selection methods

Dataset	Method	Number	Selected features
WBCD	RF	16	2, 4, 6, 7, 10, 12, 14, 19, 23, 27, 33, 40, 43, 46 ,48, 49, 51
	LASSO	8	2, 7, 16, 17, 19, 31, 35, 47
	SFS	3	7, 32, 48
MBCD	RF	12	1, 2, 7, 11, 13, 14, 17, 19, 21, 22, 23, 29
	LASSO	10	1, 2, 8, 11, 15, 18, 21, 22, 25, 30
	SFS	4	1, 21, 22, 23

The classification results based on accuracy for WBCD and MBCD are shown in Tabs. 5 and 6. The results are presented as ALL, BO, RF-BO, LASSO-BO and SFS-BO in ML algorithms in Tabs. 5 and 6. When the results in the Tab. 5 are analyzed, it is seen that RF-BO method showed the highest accuracy rates of 96,66% in NB. The LASSO-BO method resulted in better accuracy rates 95,43%, 98,95% and 98,24% in both DT, SVM and EL algorithms. In K-NN algorithm, the SFS-BO method showed the highest accuracy rate of 98.06%.

Table 5: The classification results of machine learning algorithms for WBCD

Method	DT	NB	SVM	K-NN	EL
ALL	92,1	93,5	95,96	95,43	95,08
BO	92,8	94,38	96,84	95,78	95,61
RF-BO	94,03	**96,66**	98,77	96,48	96,30
LASSO-BO	**95,43**	95,43	**98,95**	97,19	**98,24**
SFS-BO	94,55	95,08	97,19	**98,06**	97,01

Table 6: The classification results of machine learning algorithms for MBCD

Method	DT	NB	SVM	K-NN	EL
ALL	91,28	90,26	90,77	85,13	89,23
BO	92,31	91,28	92,82	90,26	91,79
RF-BO	**95,38**	93,85	95,9	94,36	95,41
LASSO-BO	94,36	94,36	**97,95**	95,38	**97,43**
SFS-BO	93,85	**95,9**	96,41	**96,92**	95,9

As observed in Tab. 6, the LASSO-BO method was the best performer in both SVM and EL (97,95% and 97,43%). The SFS-BO method gave the highest classification results with accuracy rates of 95,9% and 96,92% in NB and K-NN algorithms. The RF-BO method showed the highest classification results with accuracy rates of 95,38% in DT algorithm.

5 Discussion

For efficient early diagnosis of BC, a classification model based on improved machine learning algorithms presented in this study. The proposed classification model was tested on two different BC datasets. Initially, all features were given directly as input to machine learning algorithms. Then, machine learning methods were optimized with the help of Bayesian optimization method to improve the classification performance and all features were given to ML as input. Finally, we combined RF, LASSO and SFS and Bayesian optimization approach to improve ML and guarantee the performance efficiency of ML. In Figs. 3 and 4, how the performance of each ML was changed all the experiments are demonstrated for WBCD and MBCD. The classification results in the Figs. 3 and 4 are demonstrated in terms of all features and the FS-BO with the best classification rates. Based on the results in Figs. 3 and 4 when machine learning algorithms were optimized with help of Bayesian optimization, the optimization process were slightly increased the accuracy rates of all machine learning algorithms. However, when the combination of BO and FS method was used, the accuracy rates of all ML algorithms were significant increased. In the light of this information, it can be said that FS-BO approach significantly increased the classification rates of machine learning algorithms.

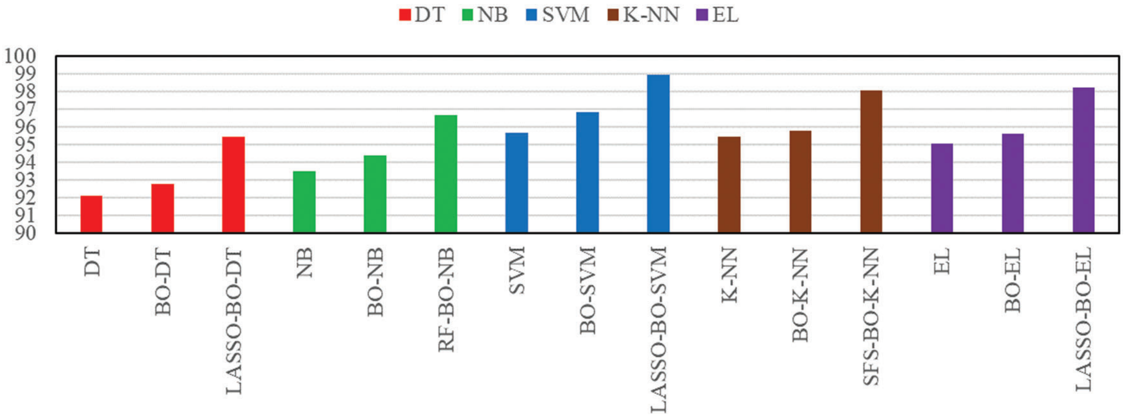

Figure 3: The accuracy values of for WBCD

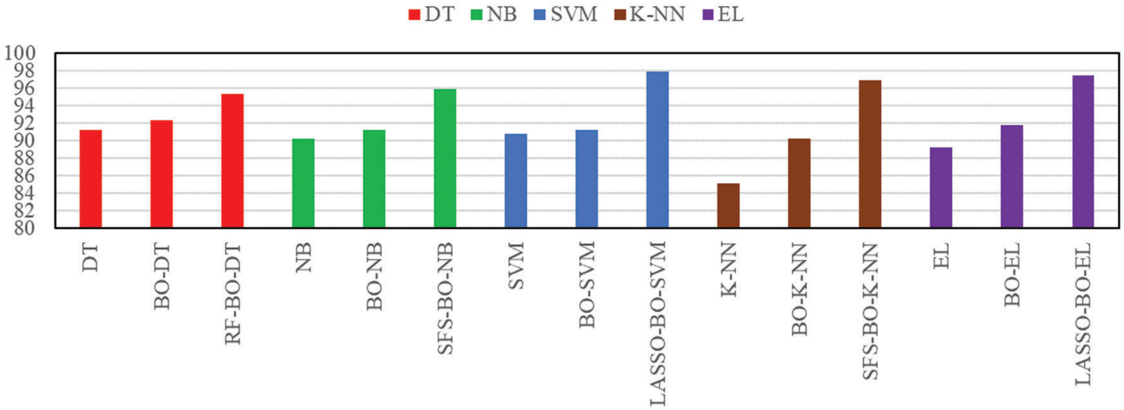

Figure 4: The accuracy values for MBCD

The Tab. 7 depicts the best classification rates of each algorithm for WBCD. The performance like accuracy, precision, recall and F1-score is compared with various classifiers in Tab. 7. Here, LASSO-DT, RF-BO-NB, LASSO-BO-SVM, SFS-BO-K-NN methods are compared in Tab. 7. As a result of comparisons, LASSO-BO-SVM method showed higher accuracy of 98,95%, precision of 97,17%, recall of 100% and F1-score of 98,57% than other methods.

Table 7: The best classification results of FS-BO-ML for WBCD

Methods	Accuracy	Precision	Recall	F1-Score
LASSO-BO-DT	95,43	94,33	93,46	93,9
RF-BO-NB	96,66	97,17	94,06	95,59
LASSO-BO-SVM	**98,95**	**97,17**	**100**	**98,57**
SFS-BO-K-NN	98,06	95,28	99	97,35
LASSO-BO-EL	98,24	98,11	97,08	97,65

The best classification rates for MBCD are shown in Tab. 8. RF-BO-DT, SFS-BO-NB, LASSO-BO-SVM, SFS-BO-K-NN and LASSO-BO-EL methods are compared in Tab. 8. As observed in Tab. 8, LASSO-BO-SVM method gave the highest performance (accuracy of 97,95%, precision of 98,28%, recall of 98,28% and F1-score of 98,28%).

Table 8: The best classification results of FS-BO-ML for MBCD

Methods	Accuracy	Precision	Recall	F1-Score
RF-BO-DT	95,38	96,55	95,72	96,14
SFS-BO-NB	95,9	95,69	97,37	96,52
LASSO-BO-SVM	**97,95**	**98,28**	**98,28**	**98,28**
SFS-BO-K-NN	96,92	98,28	96,61	97,44
LASSO-BO-EL	97,43	98,28	98,24	97,85

The confusion of matrices of LASSO-BO-SVM methods for WBCD and MBCD are showed in Fig. 5. While “0” represents benign status, “1” represents malign status.

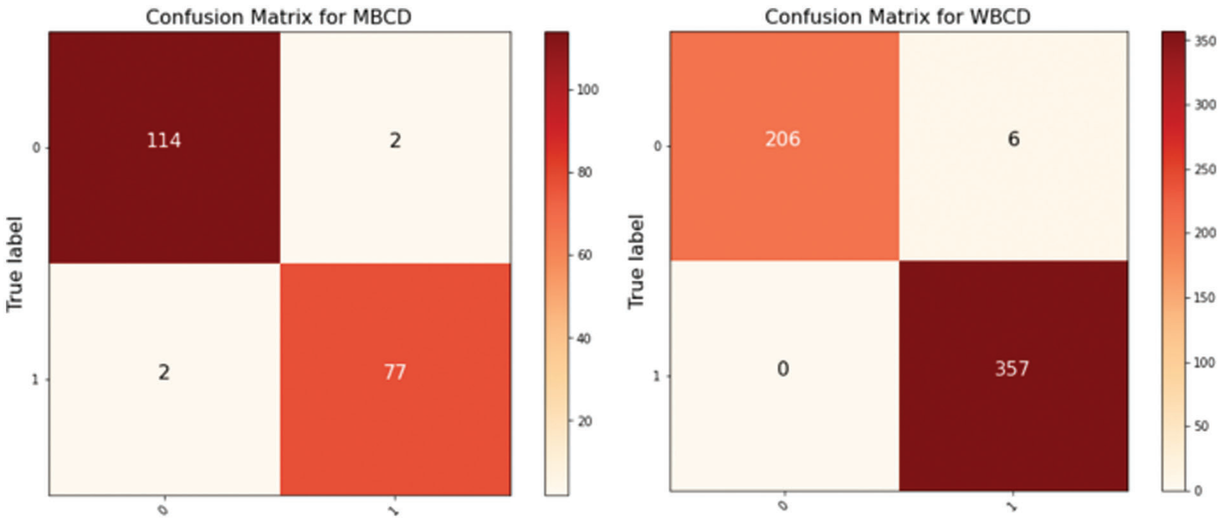

Figure 5: The confusion matrices of WBCD and MBCD

The proposed method (LASSO-BO-SVM) are compared to six recent studies with using WBCD in Tab. 9. It was noticed that LASSO-BO-SVM method achieved higher accuracy values compared to previous works. LASSO-BHO-SVM method may be used an effective approach in the early diagnosis of BC with the high classification rates.

Table 9: Comparison of accuracy of LASSO-BO-SVM with other existing approaches

Reference	Methods	Dataset	Accuracy
Mate et al. [30]	BO-ETC	WBCD	96.52%
Kumar et al. [29]	BO-RF Classifier	WBCD	96.14%
Thawkar et. [54]	BOA-ALO-ANN	WBCD	98.16%
Asri et al. [19]	SVM	WBCD	97.13%
Bensaoucha et.al. [31]	BO-SVM	WBCD	96.52%
Khandezemin et al. [55]	LR-GMDH	WBCD	97.9%
Proposed Method	**LASSO-BO-SVM**	**WBCD**	**98.95%**
		MBCD	**97.95%**

Note: LR: Logistic Regression, GMDH: Group Method Data Handling, RF: Random Forest, BOA: Butterfly Optimization Algorithm, ALO: Ant Lion Optimizer, ANN: Artificial Neural Network, ETC: Extra Tree Classifier.

6 Conclusion and Future Works

Recent years witnessed many studies toward the diagnosis of BC in its initial stage. Although much effort has been directed to this field, it is still very challenging for researchers to choose the right method for an effective diagnostic model. The study proposes a novel classification model based on improved ML algorithms combining RF, LASSO and SFS methods and Bayesian optimization for efficient diagnosis of BC. Among the many variations, the LASSO-BO-SVM method depicted the highest accuracy, sensitivity, precision and F1-score for BC datasets. With these high classification rates, LASSO-BO-SVM technique has a potential to help radiologists for making more accurate BC diagnosis decisions. In the future works, we will use deep learning models for early diagnosis of BC and compare machine learning algorithms and deep learning models.

Acknowledgement: The authors would like to acknowledge the Department of Radiology of Ankara Training and Research Hospital, Ankara for their support in this work.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] E. H. Houssein, M. M. Emam, A. A. Ali and P. N. Suganthan, "Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review," *Expert Systems with Applications*, vol. 167, no. 24, pp. 114161, 2021.

[3] A. R. Vaka, B. Soni and S. Reddy, "Breast cancer detection by leveraging machine learning," *ICT Express*, vol. 6, no. 4, pp. 320–324, 2020.

[4] P. Zubor, P. Kubatka, K. Kajo, Z. Dankova, H. Polacek *et al.,* "Why the gold standard approach by mammography demands extension by multiomics? Application of liquid biopsy miRNA profiles to breast cancer disease management," *International Journal of Molecular Sciences*, vol. 20, no. 12, pp. 2878, 2019.

[5] S. Al-Mahmood, J. Sapieznyski, O. B. Garbuzenko and T. Minko, "Metastatic and triple negative breast cancer: Challenges and treatment options," *Drug Delivery and Translational Research*, vol. 8, no. 5, pp. 1483–1507, 2018.

[6] J. Chhatwal, O. Alagoz and E. S. Burnside, "Optimal breast biopsy decision-making based on mammographic features and demographic factors," *Operations Research*, vol. 58, no. 6, pp. 1577–1591, 2010.

[7] W. A. Berg, "Reducing unnecessary biopsy and follow-up of benign cystic breast lesions," *Radiology*, vol. 295, no. 1, pp. 52–53, 2020.

[8] R. Sharma, J. B. Sharma, R. Maheshwari and P. Agarwal, "Thermogram adaptive efficient method for breast cancer detection using fractional derivative mask and hybrid feature set in the IoT environment," *Computer Modeling in Engineering & Sciences*, vol. 130, no. 2, pp. 923–947, 2022.

[9] G. Jayandhi, J. L. Jasmine and S. M. Joans, "Mammogram learning system for breast diagnosis using deep learning SVM," *Computer Systems Science and Engineering*, vol. 40, no. 2, pp. 491–503, 2022.

[10] N. Fatima, L. Liu, S. Hong and H. Ahmed, "Prediction of breast cancer, comparatively review of machine learning techniques and their analysis," *IEEE Access*, vol. 8, pp. 150360–150376, 2020.

[11] J. Miao and L. Niu, "A survey on feature selection," *Procedia Computer Science*, vol. 91, no. 4, pp. 919–926, 2016.

[12] O. M. Alyasiri, Y. N. Cheah, A. K. Abasi and O. M. Al-Janabi, "Wrapper and hybrid feature selection methods using metaheuristic algorithms for English text classification: A systematic review," *IEEE Access*, vol. 10, pp. 39833–39852, 2022.

[13] M. Claesen and B. De Moor, "Hyperparameter search in machine learning," *pre-print, arXiv:1502.02127*, 2015. [Online]. Available: https://arxiv.org/abs/1502.02127.

[14] S. F. M. Radzi, M. K. A. Karim, M. I. Saripan, M. A. A. Rahman, I. N. C. Isa *et al.,* "Hyperparameter tuning and pipeline optimization via grid search method and tree-based AutoML in breast cancer prediction," *Journal of Personalized Medicine*, vol. 11, no. 10, pp. 978, 2021.

[15] L. Gao and Y. Ding, "Disease prediction via Bayesian hyperparameter optimization and ensemble learning," *BMC Research Notes*, vol. 13, no. 1, pp. 205, 2020.

[16] S. Ibrahim, S. Nazir. and S. A. Velastin, "Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis," *Journal of Imaging*, vol. 7, no. 11, pp. 225, 2021.

[17] O. Bardhi and B. Garcia Zapirain, "Machine learning techniques applied to electronic healthcare records to predict cancer patient survivability," *Computers, Materials & Continua*, vol. 68, no. 2, pp. 1595–1613, 2021.

[18] Y. E. Almalki1, A. Shaf, T. Ali, M. Aamir, S. K. Alduraibi *et al.,* "Breast cancer detection in Saudi Arabian women using hybrid machine learning on mammographic images," *Computers, Materials & Continua*, vol. 72, no. 3, pp. 4833–4851, 2022.

[19] H. Asri, H. Mousannif, H. Al Moatassime and T. Noel, "Using machine learning algorithms for breast risk prediction and diagnosis," *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016.

[20] M. M. Khan, S. Islam, S. Sarkar, F. I. Ayaz, M. K. Ananda *et al.,* "Machine learning based comparative analysis for breast cancer prediction," *Journal of Healthcare Engineering*, vol. 2022, pp. 4365855, 2022.

[21] Y. Wang, X. Gao, X. Ru, P. Sun and J. Wang, "A hybrid feature selection algorithm and its application in bioinformatics," *PeerJ Computer Science*, vol. 8, pp. e933, 2022.

[22] C. Tian, J. Lv and X. F. Xu, "Evaluation of feature selection methods for mammographic breast cancer diagnosis in a unified framework," *BioMed Research International*, vol. 2021, no. 3, pp. 6079163, 2021.

[23] N. Naveed, H. T. Madhloom and M. S. Husain, "Breast cancer diagnosis using wrapper-based feature selection and artificial neural network," *Applied Computer Science*, vol. 17, no. 3, pp. 19–30, 2021.

[24] M. Abd-elnaby, M. Alfonse and M. Roushdy, "A hybrid mutual information-LASSO-genetic algorithm selection approach for classifying breast cancer," in *Digital Transformation Technology*, vol. 224. Singapore: Springer, pp. 547–560, 2022.

[25] J. O. Afoloyan, M. O. Adebiyi, M. O. Arowolo, C. Chakraborty and A. A. Adebiyi, “Breast cancer using particle swarm optimization and decision tree machine learning technique,” in *Intelligent Healthcare*. Singapore: Springer, pp. 61–83, 2022.

[26] R. Dhanya, I. R. Paul, S. S. Akula, M. Sivakumar and J. J. Nair, “A comparative study for breast cancer prediction using machine learning and feature selection,” in *2019 IEEE Int. Conf. on Intelligent Computing and Control Systems (ICCS)*, Madurai, India, pp. 1049–1055, 2019.

[27] W. S. Abdaldawi and R. M. Almuttari, “Hybrid ANOVA and LASSO methods for feature selection and Linear Support Vector, Multilayer Perceptron and Random Forest Classifiers based on spark environment for microarray data classification,” in *IOP Conf. Series: Materials Sciences and Engineering*, Baghdad, Iraq, vol. 1094, pp. 12107, 2021.

[28] J. Wu, X. Y. Chen, L. D. Xiang, H. Lei and S. H. Deng, “Hyperparameter optimization for machine learning models based on Bayesian optimization,” *Journal of Electronic Science and Technology*, vol. 17, no. 1, pp. 26–40, 2019.

[29] P. Kumar and G. G. Nair, “An efficient classification framework for breast cancer using hyperparameter tuned Random Forest and Bayesian Optimization,” *Biomedical Signal Processing and Control*, vol. 68, pp. 102681, 2021.

[30] Y. Mate and N. Somai, “Hybrid feature selection and Bayesian optimization with machine learning for breast cancer prediction,” in *2021 7th Int. Conf. on Advanced Computing and Communication Systems (ICCS)*, Coimbatore, India, pp. 612–619, 2021.

[31] S. Bensaoucha, “Breast cancer diagnosis using optimized machine learning algorithms,” in *2021 Int. Conf. on Recent Advances in Mathematics and Informatics (ICRAMI)*, Tebessa, Algeria, pp. 1–6, 2021.

[32] W. H. Wolberg, W. N. Street and O. L. Mangasarian, “Breast cancer Wisconsin (diagnostic) data set,” in *UCI Machine Learning Repository*, 1992. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic).

[33] B. Surendiran and A. Vadivel, “Mammogram mass classification using various geometric shape and margin features for early detection of breast cancer,” *International Journal of Medical Engineering and Informatics*, vol. 4, no. 1, pp. 36–54, 2012.

[34] A. Vadivel and B. Surendiran, “A fuzzy rule-based approach for characterization of mammogram masses into BI-RADS shape categories,” *Computers in Biology and Medicine*, vol. 43, no. 4, pp. 259–267, 2013.

[35] P. Klezcek, GLCM_Features (glcm), 2017. Available: https://www.mathworks.com/matlabcentral/fileexchange/56661-glcm_features-glcm.

[36] X. Wei, Gray level run length matrix toolbox v1.0, software, Beijing Aeronautical Technology Research Center, 2007. Available: https://www.mathworks.com/matlabcentral/fileexchange/17482-gray-level-run-length-matrix-toolbox.

[37] S. Kotsiantis, D. Kanellopoulos and P. E. Pintelas, “Data preprocessing for supervised learning,” *International Journal of Computer and Information Engineering*, vol. 1, no. 12, pp. 4091–4096, 2007.

[38] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson and J. H. Moore, “ReliefF-based feature selection: Introduction and review,” *Journal of Biomedical Informatics*, vol. 85, no. 1, pp. 189–203, 2018.

[39] M. Robnik-Sikonja and I. Konenko, “Theoretical and empirical analysis of ReliefF and RReliefF,” *Machine Learning*, vol. 53, no. 1, pp. 23–69, 2003.

[40] R. Muthukrishnan and R. Rohini, “LASSO: A feature selection technique in predictive modeling for machine learning,” in *2016 IEEE Int. Conf. on Advances in Computer Applications (ICACA)*, Coimbatore, India, pp. 18–20, 2016.

[41] R. Aggrawal and S. Pal, “Sequential feature selection and machine learning algorithm-based patient’s death events prediction and diagnosis in heart disease,” *SN Computer Science*, vol. 1, no. 6, pp. 1–16, 2020.

[42] Z. Nematzedeh, R. Ibrahim and A. Selamet, “Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques,” in *2015 IEEE 10th Asian Control Conf. (ASCC)*, Kota Kinabalu, Malaysia, pp. 1–6, 2015.

[43] S. B. Kotsiantis, “Decision trees: A recent overview,” *Artificial Intelligence Review*, vol. 39, no. 4, pp. 261–283, 2013.

[44] H. Zhang, “Exploring conditions for the optimality of Naive Bayes,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 19, no. 2, pp. 183–198, 2005.

[45] S. Peng, Q. Hu, Y. Chen and J. Dang, “Improved support vector machine algorithm for heterogeneous data,” *Pattern Recognition*, vol. 48, no. 6, pp. 2072–2083, 2015.

[46] Z. Zhang, “Introduction to machine learning: K-nearest neighbors,” *Annals of Translational Medicine*, vol. 4, no. 11, pp. 218, 2016.

[47] T. N. Rincy and R. Gupta, “Ensemble learning techniques and its efficiency in machine learning: A survey,” in *2nd Int. Conf. on Data, Engineering and Applications (IDEA)*, Bhopal, India, pp. 1–6, 2020.

[48] L. Yang and A. Shami, “On hyperparameter optimization of machine learning algorithms: Theory and practice,” *Neurocomputing*, vol. 415, no. 1, pp. 295–316, 2020.

[49] J. Wu, X. Y. Chen, H. Zhang, L. D. Xiong, H. Lei *et al.,* “Hyperparameter optimization for machine learning models based on Bayesian optimization,” *Journal of Electronic Science and Technology*, vol. 17, no. 1, pp. 26–40, 2020.

[50] J. Snoek, H. Larochelle and R. P. Adams, “Practical Bayesian optimization of machine learning algorithms,” *Advances in Neural Information Processing Systems*, vol. 2, pp. 2951–2959, 2012.

[51] Y. Liu, Y. Zhou, S. Wen and C. Tang, “A strategy on selecting performance metrics for classifier evaluation,” *International Journal of Mobile Computing and Multimedia Communications*, vol. 6, no. 4, pp. 20–35, 2014.

[52] MATLAB and Statistics Toolbox Release 2020a, The MathWorks, Inc., Natick, Massachusetts, United States, 2022.

[53] Introduction to Feature Selection, 2022. [Online]. Available: https://www.mathworks.com/help/stats/feature-selection.html.

[54] S. Thawkar, S. Sharma, M. Khanna and L. K. Singh, “Breast cancer prediction using a hybrid method based on Butterfly Optimization and Ant Lion Optimizer,” *Computers in Biology and Medicine*, vol. 139, no. 3, pp. 104968, 2021.

[55] Z. Khandezamin, M. Naderan and M. J. Rashti, “Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier,” *Journal of Biomedical Informatics*, vol. 111, no. 144, pp. 103591, 2020.