Original Article

Developing the breast cancer risk prediction system using hybrid machine learning algorithms

Mohammad R. Afrash[1], Azadeh Bayani[1], Mostafa Shanbehzadeh[2], Mohammadkarim Bahadori[3], Hadi Kazemi-Arpanahi[4,5]

Access this article online
Quick Response Code:

Website:
www.jehp.net

DOI:
10.4103/jehp.jehp_42_22

[1]Department of Health Information Technology and Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran, [2]Department of Health Information Technology, School of Paramedical, Ilam University of Medical Sciences, Ilam, Iran, [3]Health Management Research Center, Baqiyatallah University of Medical Sciences, Tehran, Iran, [4]Department of Health Information Management and Technology, Abadan University of Medical Sciences, Abadan, Iran, [5]Student Research Committee, Abadan University of Medical Sciences, Abadan, Iran

Address for correspondence:
Dr. Hadi Kazemi-Arpanahi,
Department of Health Information Management and Technology, Abadan University of Medical Sciences, Abadan, Iran.
E-mail: h.kazemi@abadanums.ac.ir

Received: 10-01-2022
Revised: 12-02-2022
Accepted: 21-02-2022
Published: 25-08-2022

Abstract:

BACKGROUND: Breast cancer (BC) is the most common cause of cancer-related deaths in women globally. Currently, many machine learning (ML)-based predictive models have been established to assist clinicians in decision making for the prediction of BC. However, preventing risk factor formation even with having healthy lifestyle behaviors or preventing disease at early stages can significantly lead to optimal population-wide BC health. Thus, we aimed to develop a prediction model by using a genetic algorithm (GA) incorporating several ML algorithms for the prediction and early warning of BC.

MATERIAL AND METHODS: The data of 3168 healthy individuals and 1742 patient case records in the BC Registry Database in Ayatollah Taleghani hospital, Abadan, Iran were analyzed. First, a modified hybrid GA was used to perform feature selection and optimization of selected features. Then, with the use of selected features, several ML algorithms were trained to predict BC. Afterward, the performance of each model was measured in terms of accuracy, precision, sensitivity, specificity, and receiver operating characteristic (ROC) curve metrics. Finally, a clinical decision support system based on the best model was developed.

RESULTS: After performing feature selection, age, consumption of dairy products, BC family history, breast biopsy, chest X-ray, hormone therapy, alcohol consumption, being overweight, having children, and education statuses were selected as the most important features for prediction of BC. The experimental results showed that the decision tree yielded a superior performance than other ML models, with values of 99.3%, 99.5%, 98.26% for accuracy, specificity, and sensitivity, respectively.

CONCLUSION: The developed predictive system can accurately identify persons who are at elevated risk for BC and can be used as an essential clinical screening tool for the early prevention of BC and serve as an important tool for developing preventive health strategies.

Keywords:

Breast cancer, lifestyle, machine learning, prevention

Introduction

Breast cancer (BC) is one of the most frequent cancers in women and the second leading cause of death after lung cancer, according to the World Health Organization (WHO) report.[1-3] BC is the most frequent malignancy among females, with an estimated 11.7% of all cancer cases and 2.3 million new cases in 2020.[4] Several biomarkers have been identified for detecting and predicting this disease.[5,6] Family history is one of the known risk factors for BC. Age is another factor that may be associated with the risk of BC. For example, studies state that women with a family history of BC and those who are more than 40 years old are highly susceptible to have the risk of BC.[7] Evidence indicates the relationship between fat consumption and BC. Several studies indicate a significant relationship between diet habits, especially fat consumption in postmenopausal women,

How to cite this article: Afrash MR, Bayani A, Shanbehzadeh M, Bahadori M, Kazemi-Arpanahi H. Developing the breast cancer risk prediction system using hybrid machine learning algorithms. J Edu Health Promot 2022;11:272.

and BC.[8,9] Other evidence demonstrated that hormonal factors, having a history of benign breast tumors, family history of BC, and genetic factors increase the risk of BC incidence.[5,10] Obesity, which is prevalent in about 20% of the population in developed countries, is another factor that increases the risk of BC during the postmenopausal period.[11] Diabetes is another risk factor; 5%–16% of patients with BC who are more than 65 years old have diabetes. The incidence of BC and type 2 diabetes mellitus is prevalent among older people who are a common risk factor in obese individuals.[10,12] Regarding the lifestyle of the people in the modern era, weight gain and obesity are prevalent and may thus increase the risk of BC.[7,13] However, the reports vary; some researchers suggest that people with a BMI of more than 30 are highly susceptible to the incidence of BC in premenopausal periods.[6,13-15] Although various factors influence the risk of BC, these factors are not linearly associated, and the relationship between them is complicated.[6] Therefore, using machine learning (ML) may be influential as these techniques do not consider the association of the variables nonlinear and are compatible with complicated relations.[16] In recent years, many ML techniques have been employed for predicting and classifying BC outcomes. ML algorithms consist of supervised and unsupervised methods. We considered supervised methods. In the supervised approach, we used a part of our data as a training dataset to train our model, and then we tested the model with the part of data that is new to the algorithm.[2,17-19]

To date, many studies have tried to predict BC by using outstanding ML techniques.[2,17,20-22] For example, Dhahri *et al*.[2] introduced an automated BC detection based on ML algorithms. Their study was based on GA and ML techniques; they aimed to develop a system to accurately differentiate between benign and malignant breast tumors. They applied several techniques to choose the best features and perfect parameter values as ML classifiers inputs. They found that the GA can automatically detect the best model by combining feature preprocessing methods and classifier algorithms.

Salod *et al*.,[4] in their study, compared the performance of ML algorithms in BC screening and detection. They used anthropometric blood analysis data from female BC patients and volunteer healthy controls of the UCI ML repository datasets. They applied eight ML algorithms, including logistic regression, support vector machine (SVM), k-nearest neighbors (KNN), decision tree (DT), random forest (RF), adaptive boosting (AdaBoost), gradient boosting machine (GBM), and eXtreme gradient boosting (XGBoost) and selected the best model considering the performance metrics of accuracy, precision, recall or sensitivity, specificity, F1 score, receiver operating characteristic (ROC) curve, and area under the ROC curve (AUC).

In another study conducted by Soltani Sarvestani *et al*.,[23] efficient networks for BC data mining from clinically collected datasets were investigated. By using various data mining techniques, they aimed to find out the percentage of disease development. The performance of the statistical neural network models, self-organizing map (SOM), radial basis function network (RBF), general regression neural network (GRNN), and probabilistic neural network (PNN) were evaluated on the Wisconsin Breast Cancer data (WBCD) and the Shiraz Namazi Hospital Breast Cancer Data (NHBCD). The results were considered, and the effectiveness and performance of the proposed networks were compared. The PNN yielded the best classification accuracy.

Chaurasia *et al*.[24] investigated the prediction of benign and malignant BC by using data mining techniques. They employed three popular data mining algorithms—naive Bayes, RBF network, and J48—to develop prediction models by using breast-cancer Wisconsin having 699 instances, two classes (malignant and benign), including the features of tumor structures and historical data of patients. They reported, according to average accuracy, that naive Bayes exhibits the best capability with 97.36% accuracy; RBF network was second with 96.77% accuracy, and J48 came out third with 93.41% accuracy. Regardless of the various studies that investigated ML for BC prediction, few studies considered the lifestyle and historical data of patients.[22,25,26] Few studies considered the combination of the lifestyle, history of diseases, and demographics data to predict BC regarding that these data are more available and affordable. Therefore, in our study, we aimed to predict BC based on the patients' demographics, history, and lifestyle features by using GA incorporated with several ML approaches such as KNN, RBF, DT, fuzzy neural network (FNN), PNN, and pattern recognition networks. Finally, their performance was compared to introduce the best model for predicting BC.

Materials and Methods

Study design and setting

This retrospective cross-sectional study was conducted in 2021 at Ayatollah Taleghani Hospital, which is the cancer screening hub and treatment center in Abadan city, Southwest region of Khuzestan province, Iran. The study was approved by the Ethics Committee of Abadan University of Medical Sciences (code: IR. ABADANUMS. REC.1400.040). To protect the privacy and confidentiality of patients, we concealed the unique identification information of all patients in the process of data collection. All experiments on the classification algorithms described in this study were conducted using the Python programming language (version 3.7.7). The Python experiment environment offers a well-defined framework for researchers and developers to run and assess their ML models. The road map of the proposed system for predicting the risk of BC based on lifestyle factors is depicted in Figure 1.

Study participants and sampling

Data preprocessing

It is very important to prepare and clean the dataset before using it to construct ML models. In the preprocessing stage, to ensure effective use of data in classification algorithms, the raw data were inputted using several preprocessing methods such as deletion of missing values, minimum and maximum scalar, and standard scalar. The standard scalar guarantees that every feature has the mean as 0 and variance as 1, bringing all features to the same coefficient. Similarly, in minimum and maximum scalar transfers, the values were such that all attributes are between 0 and 1, and rows with missing values (greater than 50%) were removed. In addition, the remaining missing values were inputted with the mean or mode of each variable. Noisy and abnormal values, errors, duplicates, and meaningless data were checked by researchers in collaboration with two infectious disease specialists and oncologists.

Data sample

A total of 6870 supposed BC cases were referred to this center from February 2017 to 2020 for cancer screening or diagnosis. Of those, 5520 cases underwent mammography. By applying the predefined exclusion criteria, 3930 cases remained. Of those, 1270 patients (32.31%) were introduced as confirmed BC by mammography in combination with biopsy, and 2660 were diagnosed non-BC (67.69%) [Figure 2]. After removing the missing values, two classes (healthy and patient) and 32 integer-valued attributes were identified. The attributes of the dataset are presented in Table 1.

Feature selection

Feature selection is an effective technique that is used to determine relevant features, reduce the dimensions of the dataset, and improve the efficiency of the classifier. This method, along with the assessment of the technique to score diverse evolved feature subsets, is required to obtain the best or most favorable output.[27]

The genetic method, which is a type of feature selection algorithm based on a random optimization technique, was applied to select the most relevant features to predict BC. GA is based on the Darwinian theory, which tries to inspire the strategies of the natural evolution of living beings. The GA operates on a population of solutions at various sequential generations for choosing superior offspring based on the “survival of the fittest” principle. The process begins by making an accidental population of solutions; it does not assess solutions in sequence but assesses a set of solutions synchronously. There are three important operators in GA: selection, crossover, and mutation within chromosomes. In our study, GA

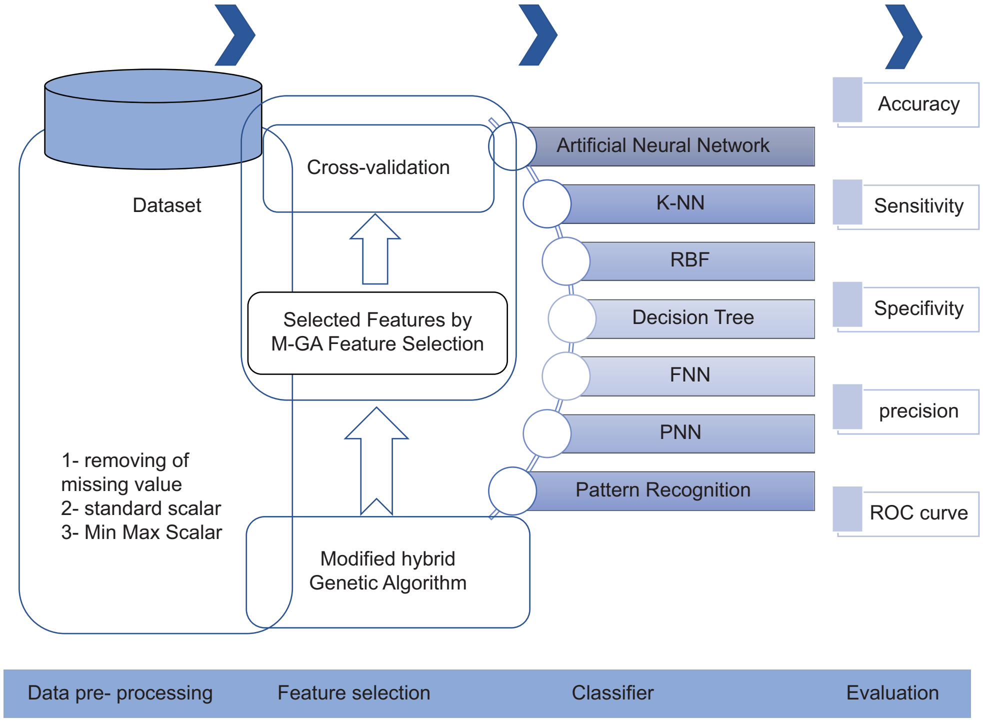


Figure 1: Block diagram of the proposed system for predicting the risk of breast cancer based on lifestyle factors

was combined with the best predictive ML algorithms to perform early diagnosis or prediction of BC based on selected features.

Model development and evaluation

Classifiers

In the present study, to predict the risk of BC based on lifestyle factors, several ML classification algorithms, including KNN, RBF, DT, artificial neural network (ANN), FNN, PNN, and pattern recognition, were used. These algorithms were selected because these are simple yet powerful models and can yield feasible results. In addition, the use of different approaches for developing the prediction models may increase the chances of obtaining a better prediction model with high classification accuracy.

ML algorithms usually have a set of parameters that must be set before running the models. The selection of parameters can notably impact the performance of models, but distinguishing the good value of parameters can be complex. In this study, we applied a set of parameters as shown in Table 2.

KNN algorithm

KNN is an algorithm for classifying variables by considering the nearest training data in the feature space. KNN applies an instance-based learning method, which is one of the simplest algorithms among data mining techniques. This method considers the nearest neighbors to each object and decides to dedicate the object to classes.[28,29]

Artificial neural network

ANN is an ML algorithm that imitates the biological neural network. In our study, we applied two types of networks: multilayer perceptron neural networks (MLPNNs) and RBF networks.[30] The MLPNN maps a set of input data to a set of appropriate output classes. RBF network is another type of neural network. As the input of the neurons in an MLP network takes the weighted sum of its inputs, every input value is multiplied by a coefficient and then the outputs are obtained by summation of the values.[24,31,32] A single MLP neuron is a simple linear classifier, but complex nonlinear classifiers can be built by combining these neurons into a network.[17]

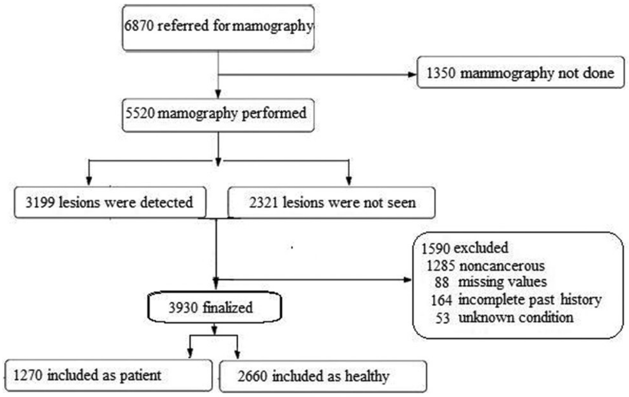

6870 referred for mamography
1350 mammography not done
5520 mamography performed
3199 lesions were detected
2321 lesions were not seen
1590 excluded
1285 noncancerous
88 missing values
164 incomplete past history
53 unknown condition
3930 finalized
1270 included as patient
2660 included as healthy

Figure 2: Flowchart describing patient selection

Decision trees

DT consists of two parts: nodes and rules. The main idea of this algorithm is to create a tree that contains a root node on top; each non-leaf node represents an attribute, and the final results are represented in the leaf

Table 1: Data set attributes and corresponding values for BC early detection

Variable name	Values	Variable name	Values
Age	Num	Meat Consumption Status	Rarely Some times highly
Gender	Male Female	Fish Consumption Status	Rarely Some times highly
Marital Status	Single Married	Vitamin Supplements	Yes No
Smoking	Yes No	Family Breast Cancer History	Yes No
Pregnancy Status	Yes No	Breast Biopsy History	Yes No
Having Children	Yes No	Chest Radiology History	Yes No
Regular Physical Activity	Yes No	Brest Examination History	Yes No
Diabetes	Yes No	Dairy Products Status	Rarely Some times highly
Fruit Consumption Status	Rarely Some times highly	Colorectal Cancer	Yes No
Having Job	Yes No	Hyperglyceridemia	Yes No
Higher Salt Intake	Yes No	Drinking Alcohol	Yes No
Breast Implants	Yes No	Body Mass Index	Num
Hyperlipidemia	Yes No	Fiber Consumption Status	Rarely Some times highly
Being overweight or obese	Yes No	Hormone Therapy History	Yes No
Vegetable Consumption Status	Rarely Some times highly	Education Statues	illiterate High school Bachelor High.
Hypertension	Yes No	Waist	Num
Hypercholesterolemia	Yes No	Breast Feeding History	Yes No
Contraceptives	Yes No		

Table 2: Parameters for machine learning algorithms

Number	Model	Parameters	
1	KNN	K=1, 3, 5	
2	RBF	32-3182-2, Spread=150	
3	DT		
4	ANN - MLP	32-10-5-2. Training Ratio: 80% Validation Ratio: 20%	
5	FNN	Using Matlab Toolbox	
6	PNN	32-3182-2, Spread=150	
7	Pattern recognition network	32-10-5-2	
	Descriptions	Value	Parameters
Genetic parameters	-	50	Population Size
	Uniform/Random Mutation[36]	0.3	mutation probability Rate(P_m)
	Uniform Crossover[36,37]	0.8	crossover probability rate(P_c)
	-	100	Maximum Number of Iterations
	-	10	Number of independent executions
	Roulette Wheel	-	Selection

nodes. DT algorithms have been widely used in data mining applications as they are one of the most powerful classification tools.[33]

Fuzzy neural networks and probabilistic neural networks

FNN integrates the advantages of both fuzzy rule-based systems and neural networks. This hybrid learning algorithm uses a fuzzy inference system and is implemented in the framework of adaptive neural networks. For the classification, for each feature, FNN applies several neurons and membership functions; the number of fuzzy rules is dependent on the number of inputs.[34] PNN is another classification algorithm. This model calculates the distances from the input vector to the training input vectors and generates a vector whose elements indicate how close the input is to a training input for each input in its first layer. The second layer computes the sums of these contributions for each class of inputs and generates as its net output a vector of probabilities. Finally, a complete transfer function on the output of the second layer selects the maximum of these probabilities and outputs 1 for that class and 0 for the other classes.[35]

Evaluation

In this study, we applied 10-fold cross-validation and five performance assessment metrics. Accuracy, specificity, sensitivity, KAPA, error rate, and ROC curve were measured for comparing the performance of the classifiers (Equations 1–4).

To better compare the performance of the algorithms, we assessed the effectiveness of five ML algorithms in terms of time required to build the model, correctly classified instances, incorrectly classified instances, Kappa statistic, mean absolute error, root mean squared error (RMSE), relative absolute error, and root relative squared error (RRSE).

$$1)\ \text{classification accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} * 100$$

$$2)\ \text{classification sensitivity} = \frac{\text{Tp}}{\text{TP} + \text{FN}} * 100$$

$$3)\ \text{classification specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} * 100$$

$$4)\ \text{classification error} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} * 100$$

The output results of the classification accuracy in each run using each of the methods are presented in Table 2. Case class data were considered as negative data, and control class data were considered as positive data. Our dataset contained 32 features.

Ethical consideration

The research deputy of Abadan University of Medical Sciences (ethical code: IR.ABADANUMS.REC.1400.040) approved the current study. To protect the privacy and confidentiality of patients, we concealed the unique identifying information of all patients in the process of data collection and presentation.

Results

Patient selection

We obtained data from 5520 patients in the NCBR. Eighty-eight incomplete case records that had a lot of missing data (more than 70%) were excluded from the analysis. In addition, the missing values were inputted with the mean or mode of each variable. After applying the exclusion criteria, the final analysis was performed on the data of 3930 BC and non-BC cases who were referred to Ayatollah Taleghani Hospital for BC screening and diagnosis. Of the 3390 study participants, 1270 cases (32.32%) were BC patients and 2660 (67.68%) were healthy women, and the median age of participants was 57.25 (interquartile range: 16–86). A flowchart to represent the patient selection methodology is shown in Figure 2.

Result of selected features by genetics algorithm

The modified GA selects the most important and highly related features based on the weights of the features. The experimental results of the modified GA showed that of the 35 included variables, 10 variables were the most significant features in predicting the risk of BC. Table 3 shows the important selected risk factor for BC and their scores.

According to the results, family history of BC, hormone therapy history, breast biopsy history, and dairy product consumption yielded high scores, meaning that these four variables have a high impact on the early prediction of BC.

In this experiment, the selected features of our dataset were checked on a GA‑DT modified ML classifier with 10‑fold cross‑validation methods. The average assessment criteria of 10‑fold cross‑validation were measured.

As can be seen from Tables 4 and 5, the GA‑DT shows good performance, with a mean classification accuracy of 92.9, a mean specificity of 92.7%, and a mean sensitivity of 91.8%. Table 6 displays the 10‑fold cross‑validation results of 10 independent runs of the modified GA‑DT algorithm with selected features.

Results of K‑fold cross‑validation on the dataset with the selected features

In this study, datasets were used for seven classification algorithms with k‑fold (k = 10) cross‑validation methods; 90% of the dataset was used for training the prediction models, and only 10% of the dataset was for testing. Finally, the mean metrics of 10‑fold cross‑validation methods were measured. Furthermore, the values of the determining parameters were obtained by running ML classifiers. Table 6 shows the 10‑fold cross‑validation results and five performance assessment metrics (mean, standard deviation, minimum and maximum classification accuracy, specificity, and sensitivity). The ROC and confusion matrix for the best running of algorithms are presented in Figures 3, 4, and 5.

Table 3: Features set selected by Relief‑feature selection algorithm and their scores

Feature name	Score
Age	0.497
Dairy products status	0.515
Family breast cancer history	0.610
Breast biopsy history	0.531
Chest radiology history	0.482
Hormone therapy history	0.579
Drinking alcohol	0.370
Being overweight or obese	0.401
Having children	0.468
Education statues	0.374

In this study, for the KNN algorithm, we performed experiments with different values of k (1, 3, and 5). The results revealed that at k = 1, the performance of KNN was superior than that for other values of k, as shown in Table 5. KNN with k = 1 yielded an accuracy of 95.1%, a specificity of 95.9%, and a sensitivity of 90.4. The specificity percent of the KNN algorithm demonstrates the probability that a prediction for BC was negative and the patient does have BC. Furthermore, 90.4% sensitivity demonstrates the probability that the patient was accurately predicted for BC. The ROC and confusion matrix for KNN (k = 1) are depicted in Figure 3.

Table 7 represents the 10‑fold cross‑validation results and metrics of performance for the GA‑DT for the selected feature.

Table 8 shows the classification performance of five other classifiers with 10‑ fold CV for selected The confusion matrix and ROC curve of the models are shown in Figures 4 and 5, respectively. features. The value for the error rate of the classifier and the computation time of classification models on the selected features are shown in Figure 6.

As shown in Table 7 and Figure 5, the performances of all the models are good. The ANN was trained on several inputs and hidden layers and neurons. The ANN model yielded a mean accuracy of 98.6%, a mean specificity of 97.05%, and a mean sensitivity of 99.03% in 10 independent iterations. The standard deviation for the accuracy was 1.1, and the best results in the ten runs were 98.8% accuracy, 97.7% specificity, and 99.1% sensitivity. For the FNN, the mean accuracy in 10 iterations was 97%, with the mean specificity and sensitivity of 98.8%

Table 4: The performance measures of GA‑DT with different parameters in 10 iterations

Run	Accuracy	Specificity	Sensitivity	Features Selected	The most important feature based on ten independent running of modified GA algorithm
1	91.5755	87.5312	92.7407	1,4,5,7,8,9,10,11,12,13,14,17,19,20,21,22,26,28	Age, dairy products status,
2	90.855	89.7467	91.8519	1,4,5,7,10,12,13,14,15,17,18,20,21,25,26,28	family breast cancer history,
3	90.1062	90.1389	95.1481	1,4,5,7,10,12,17,18,19,20,21,22,25,26,28	breast biopsy history,
4	90.5049	91.6962	91.5556	1,4,5,7,8,10,11,12,13,14,17,18,19,20,22,24,28,29	chest radiology history,
5	92.5788	92.3115	93.6667	1,4,5,7,8,12,14,15,18,20,24,25,27,28,31	hormone therapy history,
6	90.5788	90.4098	90.6296	1,4,5,7,8,13,14,15,17,18,20,24,25,28,29,31	drinking alcohol,
7	91.83	90.791	91.7407	1,4,5,7,11,13,14,17,18,20,21,25,27,28	being overweight or obese,
8	93.2373	93.4799	93.2593	1,4,5,7,9,10,12,14,17,20,21,24,26,28	having children,
9	95.7432	95.908	94.2407	1,4,5,7,8,10,11,14,17,18,20,22,25,28,29	education statues
10	95.5365	94.5622	94.1667	1,4,5,7,11,13,14,17,19,20,22,24,25,27,28,29,31	

and 96.9%, respectively. The standard deviation for the accuracy obtained was 0.07, and the best results in the 10 runs were 97.2% accuracy, 99% specificity, and 97% sensitivity. For the pattern net, we obtained the mean accuracy, mean specificity, and mean sensitivity as 98.7%, 97.4%, and 99%, respectively. The standard deviation for the accuracy in 10 -iterations was 0.12, and the best results in the 10 runs for the accuracy, specificity, and sensitivity were 98.8%, 97.7%, and 99.1%, respectively. After applying the probabilistic neural network, the mean accuracy, mean specificity, and mean sensitivity were 96%, 91.4%, and 96.9%, respectively. The standard deviation for accuracy was 0.11, and the best measure obtained was 96.2% for the accuracy, and 92.1% and 97% for the specificity and sensitivity, respectively. Table 5 shows that the mean accuracy in 10 iterations of the RBF method was 83%, with the mean specificity and sensitivity of 78% and 84%, respectively. The standard deviation for the accuracy was 1.1, and the best results in the 10 runs yielded an accuracy of 85%. The pattern net model has the best computational time (38 s), and the RBF model has the worst processing time (117 s). The AUC value of the GA-DT model was the highest (99.01) in comparison to other models.

Table 5: Results for running modified GA algorithm in 10 independent executions

	Accuracy	Specificity	Sensitivity
Mean	92.9259	92.7407	91.88996
Std	1.806388	2.356811	1.671002
MIN	90.6296	87.5312	90.1062
MAX	95.7407	94.5622	94.2407

Finally, upon comparing seven ML algorithms, we found that the performance of the GA-DT algorithm was excellent; the second important classification algorithm was pattern net, and the worst performance was noted for RBF.

System implementation

The system was implemented during February and April 2021. System programming consisted of three types of implementation codes: codes for the user interface implementation, codes for the logic layer implementation, and codes for the database implementation. The user interface in our study comprised four pages: Welcome page (sign up and log-in page), and CDSS module (2 pages). The user interface of the BC risk prediction system was developed using the C# programming language, see Figures 7 and 8.

Discussion

The prevalence of healthy lifestyle with healthy behaviors is low in all worldwide countries.[38] Thus, the use of an intelligence system for predicting BC risk based on lifestyle factors may assist clinicians in the evaluation of a limited number of important lifestyle features in an endeavor to recognize individuals at high risk for BC. Furthermore, our developed lifestyle-based

Table 6: The average performance of 10 independent runs of classifiers based on 10-fold cross-validation

	K=1			K=3			K=5		
	Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity
Mean	9.4901e+01	9.5852e+01	8.9350e+01	9.4234e+01	9.4613e+01	9.1157e+01	9.3275e+01	9.3644e+01	8.9668e+01
Std	1.7312e-01	1.0816e-01	5.8369e-01	9.7362e-02	1.0276e-01	3.0080e-01	8.3126e-02	6.4132e-02	3.1143e-01
Worst	9.4656e+01	9.5649e+01	8.8693e+01	9.4090e+01	9.4481e+01	9.0671e+01	9.3099e+01	9.3549e+01	8.9160e+01
Best	9.5193e+01	9.5982e+01	9.0472e+01	9.4401e+01	9.4787e+01	9.1677e+01	9.3411e+01	9.3746e+01	9.0304e+01

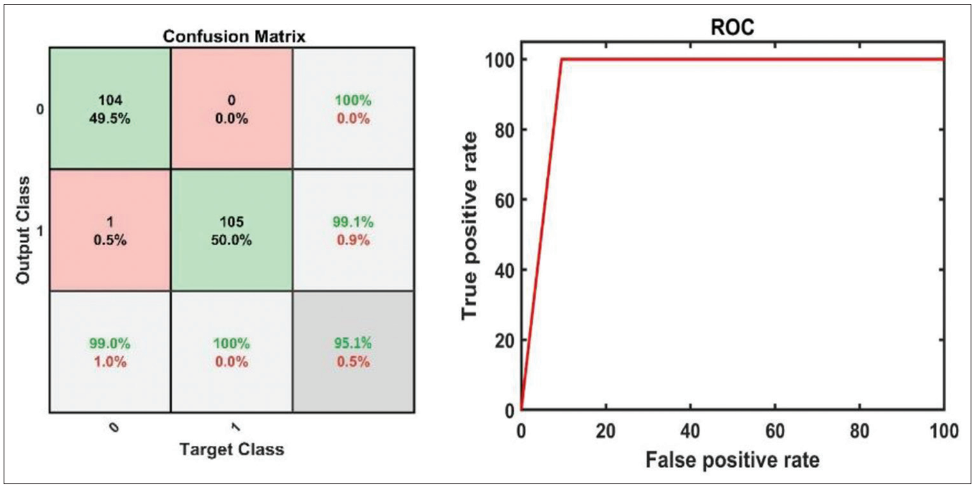


Figure 3: The best performance in confusion matrix and ROC curve for GA-KNN algorithm

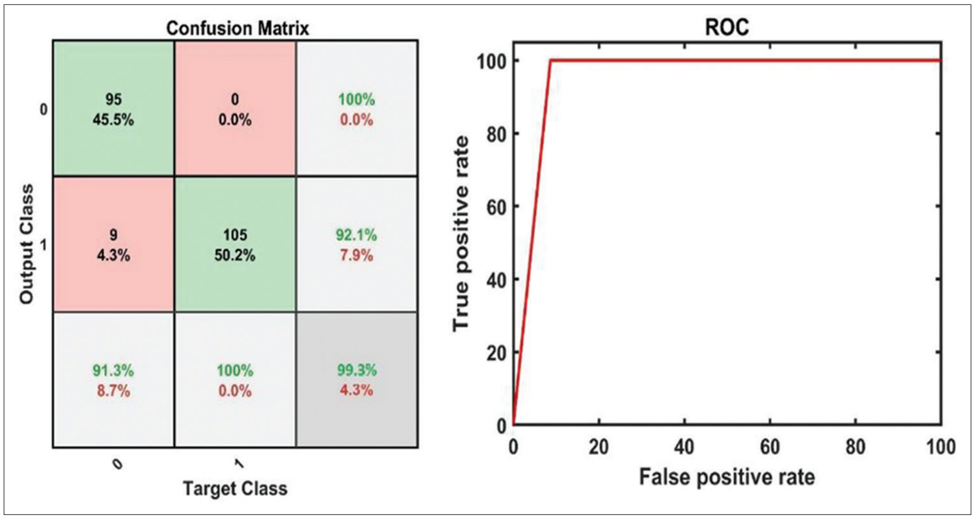


Figure 4: Confusion matrix and ROC curve for GA-DT algorithm

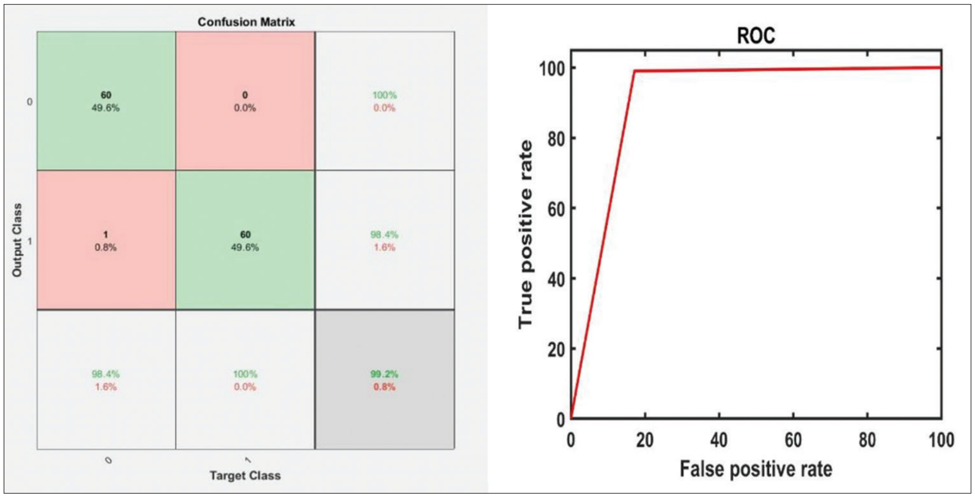


Figure 5: The best performance in confusion matrix and ROC curve for neural network algorithm

Table 7: Results for running GA-DT algorithm in 10 independent executions

	Accuracy	Specificity	Sensitivity
Mean	9.9222e+01	9.9505e+01	9.7949e+01
Std	6.4318e-02	4.4483e-02	1.4923e–01
Worst	9.9123e+01	9.9426e+01	9.7792e+01
Best	9.9321e+01	9.9554e+01	9.8239e+01

prediction system may enhance consciousness to detect true primordial inhibition through interventions on underlying unwell behaviors to ban the growth of any important risk factors of BC initially rather than treating patients only when the disease advanced.

In the present study, an ML-based predictive system was developed for early risk prediction of BC based on lifestyle factors. The ML algorithms were applied to a preprocessed dataset. Eight well-known classifier algorithms, including KNN (k = 1, 2, and 3), ANN, SVM, FNN, RFB, DT, pattern net, and PNN, were used with a feature selection algorithm (GA) to select the most important predictors. To validate the system, the k-fold cross-validation method was used. To compare the performance of the classifiers, several evaluation metrics derived from the confusion matrix were used. The feature selection method (GA) selects the most important variables that enhance the performance of the classification algorithms in terms of accuracy, specificity, sensitivity, ROC, different metrics of error rate, and the processing time of models.

The most important features of BC were age, consumption of dairy products, BC family history, breast biopsy, chest X-ray, hormone therapy, alcohol consumption, being overweight, having children, and education level. The DT algorithm with 10-fold cross-validation presented the best accuracy of 99.2%, a specificity of 99.5%, and a sensitivity of 97.9% when selected by the GA algorithm. Due to the high performance of the DT algorithm with GA, it was

Table 8: Classification performance of five other classifiers with 10- fold CV on selected features

Evaluation criteria	Classifier				
	RBF	Probabilistic neural network	Pattern net	FNN	Neural network MLP
Best Time to build a model (s)	117	83	38	63	91
Mean Accuracy (%)	83.4	96.0	98.7	97.09	98.6
Mean Specificity (%)	77.7	91.4	97.5	98.8	97.05
Mean Sensitivity (%)	84.1	96.9	99.07	96.9	99.03
STD	1.1109e+00	1.1832e−01	1.2861e−01	7.7757e−02	1.0787e−02
Worst Accuracy	81.3	95.9	98.4	97	98.5
Best Accuracy	85.0	96.2	98.8	97.2	99.8

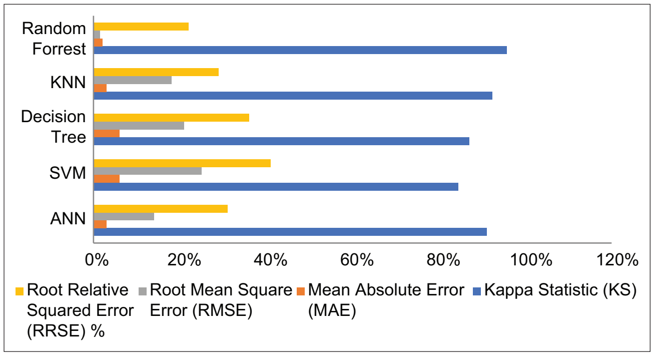
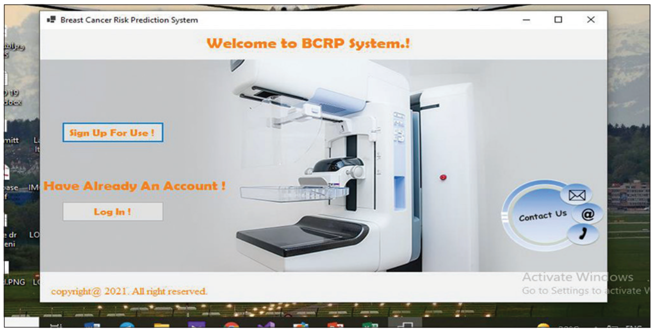


Figure 6: The error rate of classifiers on the given dataset

selected as the core of the clinical decision support system to better predict the risk of BC in terms of accuracy, specificity, and sensitivity.

In terms of specificity, the FNN algorithm has the best specificity among other models with 97% specificity and a standard deviation of 0.07. The best result for sensitivity among the eight models was obtained for pattern net, with a sensitivity of 99.1% and a standard deviation of 0.1. In addition, in terms of computation time, the pattern net with a feature selection (GA) algorithm was the best as compared to the computation time of the other seven algorithms, as shown in Figure 5. Williams *et al*.[39] showed that data mining approaches have significant predictive power for BC. They indicated that DT had the best accuracy in comparison with other techniques.

In our study, the DT model had a high accuracy of about 99% as well. This may prove the strong power of DT in predicting BC. Another study that compared different data mining algorithms is the research by Higa[40] that selected DT and neural network as the best models for diagnosing benign and malignant tumors of BC with 95% accuracy. Our study also showed that these two approaches (DT and ANN) had the best prediction powers, and we obtained higher maximum accuracy in our study. The proposed ML-GA model in the Jebarani study (2021) showed its effectiveness for classifying benign and malignant BC tumors.[41] Solanki *et al*.[42] (2021) also conducted a retrospective analysis for BC prognosis by using hybrid supervised ML classifiers.


Breast Cancer Risk Prediction System

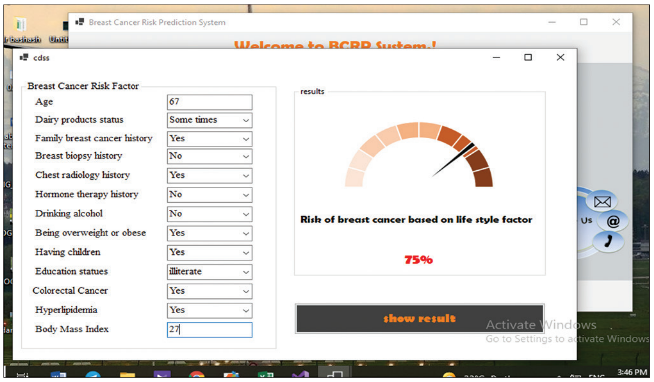
Welcome to BCRP System.!

copyright@ 2021. All right reserved.

Figure 7: The welcome page of the clinical decision support system

Finally, the best meaningful results were observed using the J-48 DT-GA classifier for feature selection with an accuracy of 98.83%, MCC = 0.974, sensitivity = 98.95%, specificity = 98.58%, and Kappa statistics = 0.9735.

Several prediction factors have been investigated in previous studies to predict BC, such as breast medical images,[43,44] the biopsy of the lesion,[24] and blood tests.[45] However, we considered a more cost-effective approach and available data with the least intervention features for our prediction models. In addition, in a previous study, several unimportant features reduced the accuracy and sensitivity of the prediction system and enhanced the processing time. Therefore, one of the innovations used in this study was applying the feature selection algorithm to select the most important factors that enhance the accuracy, specificity, and sensitivity of the classifiers as well as decrease the running time of the predictive system. In this study, the GA algorithm was utilized for solving these challenges. The results confirm the positive effect of 10 features in predicting the risk of BC, and such satisfactory results can be attributed to the use of GA as a powerful optimizer that selected the best subset features to be included in ML algorithms. It has been inferred that upon hybridizing different ML algorithms, the prediction models show more promising performance compared to a single model. Thus, ML algorithms can be used to construct complex models and make reliable decisions when fed with appropriate features. When there is a valuable set of features, the performance of ML algorithms is anticipated to be adequately acceptable. However, in specific applications, the dataset is often



Breast Cancer Risk Factor

Age	67
Dairy products status	Some times
Family breast cancer history	Yes
Breast biopsy history	No
Chest radiology history	Yes
Hormone therapy history	No
Drinking alcohol	No
Being overweight or obese	Yes
Having children	Yes
Education statues	illiterate
Colorectal Cancer	Yes
Hyperlipidemia	Yes
Body Mass Index	27

Risk of breast cancer based on life style factor

75%

Figure 8: The clinical decision support system for predicting the risk of breast cancer

insufficient or imbalanced. Therefore, it is important to train these algorithms and obtain good results with the most relevant set of features.

The predictive BC system identifies persons at high risk based on lifestyle behaviors. For individuals with a high risk of BC, protocols are in place for the optimal period of treatment. However, for persons with a low risk of BC, our developed system can provide information about long-term BC risk and BC overall burden. Ultimately, future studies are needed to evaluate the feasibility and impacts of our lifestyle BC risk prediction system, BC risk feature improvement, and overall BC risk evaluation when integrated into the clinical care environment, particularly in integration with other clinical-based risk systems.

Limitations and reconsecrations

This study had some limitations that are necessary to be recognized. Dealing with a retrospective-single center dataset, the present study suffered from the low quality (imbalanced, noisy, duplicates, and meaningless values), low quantity (missing cells), and non-optimal generalizability of data in the selected database. First, we removed noises, duplicates, and meaningless records manually as much as possible from the dataset. To solve the imbalanced dataset problem, by using the SMOTE technique, the bias was minimized via class balancing. Second, missing values were imputed with the mean or mode of each variable. Finally, it is recommended to use a dataset with a larger sample size in a multi-center setting in future studies.

Despite the limitations of our study, this is an important study on clinical prediction systems assessing the incidence risk of BC by routine lifestyle features in Iran. We expect this prediction clinical decision support system to play a significant role in enhancing the quality of decision-making and detecting individuals at high risk for BC and applying BC prevention strategies in the field of health care policy in our countries, where the screening program at the early stages is not included in the routine national health program. Additionally, the main novelty of our study is that we derived the lifestyle-based BC risk prediction empirically by using state-of-the-art and novel hybrid ML methods (i.e., hybrid GA-DT) by considering various features simultaneously. Moreover, we considered a Windows-based application to enable BC risk prediction in the first line of the healthcare system.

Conclusion

The predictive BC system identifies persons at high and elevated risk for BC based on lifestyle behaviors and can be used as an essential clinical screening tool for the early prevention of BC and serve as important tools for developing preventive health strategies. However, there is also an essential need to perform studies to evaluate the feasibility and impacts of such a lifestyle-BC risk prediction system, especially when integrated into the clinical care environment, particularly in integration with other clinical-based risk systems.

Ethics approval and consent to participate

This article is extracted from a research project supported by the Abadan University of Medical Sciences (grant no: IR. ABADANUMS. REC.1400.040). We appreciate the research deputy of the Abadan University of Medical Sciences who sponsored this project financially. We would also like to thank Ayatollah Taleghani Hospital officials who assisted the research team in conducting the research (research.ac.ir).

Consent for publication

Not applicable.

Availability of data and materials

All data generated and analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request and the Research Committee of Abadan University of Medical Science's approval.

Acknowledgments

We thank the Research Deputy of the Abadan University of Medical Sciences for financially supporting this project.

Abbreviations

ML: Machine Learning, GA: Genetic Algorithm, BC: Breast Cancer, ROC: Receiver Operating Characteristic, WHO: World Health Organization, LR: Logistic Regression, SVM: Support Vector Machine, KNN: K-Nearest Neighbors, DT: Decision Tree, RF: Random Forest, AdaBoost: Adaptive Boosting, GBM: Gradient Boosting Machine, XGBoost: eXtreme Gradient Boosting, SOM: Self-Organizing Map, RBF: Radial Basis Function Network, GRNN: General Regression Neural Network, and PNN: Probabilistic Neural Network, WBCD: Wisconsin Breast Cancer, FNN: Fuzzy Neural Networks, ANN: Artificial Neural Network, RMSE: Root Mean Squared Error, RRSE: Relative Absolute Error, and Root Relative Squared Error.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. World Health Organization. Cancer. 2018 Available from: https://www.who.int/news-room/fact-sheets/detail/cancer.
2. Dhahri H, Al Maghayreh E, Mahmood A, Elkilani W, Faisal Nagi M. Automated breast cancer diagnosis based on machine learning algorithms. J Healthc Eng 2019;2019. doi: 10.1155/2019/4253641.
3. Namini S, Elahi SA, Seirafi MR, Sabet M, Azadeh P. Predicting post-traumatic growth inventory (PTGI) based on the perceived social support; the mediating role of resilience in women with breast cancer: A structural equation modeling approach. Iran J Health Educ Health Promot 2021;9:172-86.
4. Salod Z, Singh Y. Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol. J Public Health Res 2019;8. doi: 10.4081/jphr.2019.1677.
5. Key TJ, Verkasalo PK, Banks E. Epidemiology of breast cancer. Lancet Oncol 2001;2:133-40.
6. Cheraghi Z, Poorolajal J, Hashem T, Esmailnasab N, Irani AD. Effect of body mass index on breast cancer during premenopausal and postmenopausal periods: A meta-analysis. PLoS One 2012;7:e51446. doi: 10.1371/journal.pone. 0051446.
7. Colditz GA, Willett WC, Hunter DJ, Stampfer MJ, Manson JE, Hennekens CH, *et al*. Family history, age, and risk of breast cancer: Prospective data from the Nurses' Health Study. JAMA 1993;270:338-43.
8. Farvid MS, Eliassen AH, Cho E, Liao X, Chen WY, Willett WC. Dietary fiber intake in young adults and breast cancer risk. Pediatrics 2016;137:e20151226.
9. Kotepui M. Diet and risk of breast cancer. Contemp Oncol 2016;20:13-9.
10. Wolf I, Sadetzki S, Catane R, Karasik A, Kaufman B. Diabetes mellitus and breast cancer. Lancet Oncol 2005;6:103-11.
11. Yancik R, Wesley MN, Ries LA, Havlik RJ, Edwards BK, Yates JW. Effect of age and comorbidity in postmenopausal breast cancer patients aged 55 years and older. JAMA 2001;285:885-92.
12. Park Y-MM, O'Brien KM, Zhao S, Weinberg CR, Baird DD, Sandler DP. Gestational diabetes mellitus may be associated with increased risk of breast cancer. Br J Cancer 2017;116:960-3.
13. Tehard B, Clavel-Chapelon F. Several anthropometric measurements and breast cancer risk: Results of the E3N cohort study. Int J Obes 2006;30:156-63.
14. Tian Y-F, Chu C-H, Wu M-H, Chang C-L, Yang T, Chou Y-C, *et al*. Anthropometric measures, plasma adiponectin, and breast cancer risk. Endocr Related Cancer 2007;14:669-77.
15. Barlow WE, White E, Ballard-Barbash R, Vacek PM, Titus-Ernstoff L, Carney PA, *et al*. Prospective breast cancer risk prediction model for women undergoing screening mammography. J Natl Cancer Inst 2006;98:1204-14.
16. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. Ann Intern Med 1993;118:201-10.
17. Chaurasia V, Pal S. Data mining techniques: To predict and resolve breast cancer survivability. International Journal of Computer Science and Mobile Computing IJCSMC 2014;3:10-22.
18. Lokeshkumar R, Mishra OA, Kalra S. Social media data analysis to predict mental state of users using machine learning techniques. J Educ Health Promot 2021;10:301.
19. Amirhajlou L, Sohrabi Z, Alebouyeh MR, Tavakoli N, Haghighi RZ, Hashemi A, *et al*. Application of data mining techniques for predicting residents' performance on pre-board examinations: A case study. J Educ Health Promot 2019;8.
20. Boeri C, Chiappa C, Galli F, De Berardinis V, Bardelli L, Carcano G, *et al*. Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation. Cancer Med 2020;9:3234-43.
21. Mariani MC, Tweneboah OK, Bhuiyan MAM. Supervised machine learning models applied to disease diagnosis and prognosis. AIMS Public Health 2019;6:405.
22. Valvano G, Santini G, Martini N, Ripoli A, Iacconi C, Chiappino D, *et al*. Convolutional neural networks for the segmentation of microcalcification in mammography imaging. J Healthc Eng 2019;2019:9360941. doi: 10.1155/2019/9360941.
23. Sarvestani AS, Safavi A, Parandeh N, Salehi M. Predicting breast cancer survivability using data mining techniques. 2010 2nd International Conference on Software Technology and Engineering. IEEE, 2010. p. V2-227-V2-231.
24. Chaurasia V, Pal S, Tiwari B. Prediction of benign and malignant breast cancer using data mining techniques. J Algorithm Comput Technol 2018;12:119-26.
25. Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst Appl 2009;36:3240-7.
26. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform 2006;2:117693510600200030. doi: 10.1177/117693510600200030.
27. Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. IEEE Trans knowl Data Eng 2005;17:491-502.
28. Medjahed SA, Saadi TA, Benyettou A. Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules. Int J Comput Appl 2013;62.
29. Odajima K, Pawlovsky AP. A detailed description of the use of the kNN method for breast cancer diagnosis. 2014 7th International Conference on Biomedical Engineering and Informatics. IEEE; 2014. p. 688-692.
30. Ting F, Sim K. Self-regulated multilayer perceptron neural network for breast cancer classification. 2017 International Conference on Robotics, Automation and Sciences (ICORAS). IEEE; 2017. p. 1-5.
31. Jouni H, Issa M, Harb A, Jacquemod G, Leduc Y. Neural Network architecture for breast cancer detection and classification. 2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET). IEEE; 2016. p. 37-41.
32. Afrash MR, Khalili M, Salekde MS. A comparison of data mining methods for diagnosis and prognosis of heart disease. Int J Adv Intell Paradig 2020;16:88-97.
33. Sumbaly R, Vishnusri N, Jeyalatha S. Diagnosis of breast cancer using decision tree data mining technique. Int J Comput Appl 2014;98.
34. Naghibi S, Teshnehlab M, Shoorehdeli MA. Breast cancer classification based on advanced multi dimensional fuzzy neural network. J Med Syst 2012;36:2713-20.
35. Azar AT, El-Said SA. Probabilistic neural network for breast cancer classification. Neural Comput Appl 2013;23:1737-51.
36. Engelbrecht AP. Computational Intelligence: An Introduction. Hoboken, New Jersey: John Wiley & Sons; 2007.

37. Umbarkar DA, Sheth P. Crossover operators in genetic algorithms: A review. ICTACT J Soft Comput 20156;6. doi: 10.21917/ijsc.2015.0150.
38. Lloyd-Jones DM, Hong Y, Labarthe D, Mozaffarian D, Appel LJ, Van Horn L, *et al*. Defining and setting national goals for cardiovascular health promotion and disease reduction: The American Heart Association's strategic impact goal through 2020 and beyond. Circulation 2010;121:586-613.
39. Williams K, Idowu PA, Balogun JA, Oluwaranti AI. Breast cancer risk prediction using data mining classification techniques. Tran Networks Commun 2015;3:1.
40. Higa A. Diagnosis of breast cancer using decision tree and artificial neural network algorithms. Cell 2018;1 (7):23-27.
41. Jebarani PE, Umadevi N, Dang H, Pomplun M. A novel hybrid K-means and GMM machine learning model for breast cancer detection. IEEE Access 2021;9:146153-62.
42. Solanki YS, Chakrabarti P, Jasinski M, Leonowicz Z, Bolshev V, Vinogradov A, *et al*. A hybrid supervised machine learning classifier system for breast cancer prognosis using feature selection and data imbalance handling approaches. Electronics 2021;10:699.
43. Antonie ML, Zaiane OR, Coman A. Application of data mining techniques for medical image classification. In Proceedings of the Second International Conference on Multimedia Data Mining 2001. p. 94-101.
44. Sinthia P, Malathi M. An effective two way classification of breast cancer images: A detailed review. Asian Pac J Cancer Prev 2018;19:3335-9.
45. Muthuselvan S, Sundaram KS. Prediction of breast cancer using classification rule mining techniques in blood test datasets. 2016 International Conference on Information Communication and Embedded Systems (ICICES). IEEE; 2016.