Shiraz E-Med J. 2022 July; 23(7):e120140. doi: 10.5812/semj-120140.

Published online 2022 April 20. **Research Article**

Comparing Data Mining Algorithms for Breast Cancer Diagnosis

Mostafa Shanbehzadeh [1], Raoof Nopour [2,*], Leila Erfannia [3], Morteza Amraei [4], Nahid Mehrabi [5] and Mehrnaz Mashoufi [6]

[1]Department of Health Information Technology, School of Paramedical, Ilam University of Medical Sciences, Ilam, Iran
[2]Student Research Committee, School of Health Management and Information Sciences Branch, Iran University of Medical Sciences, Tehran, Iran
[3]Department of Health Information Management, Health Human Resource Research Center, School of Health Management and Information Sciences, Shiraz University of Medical Sciences, Shiraz, Iran
[4]Department of Health Information Technology, School of Allied Medical Sciences, Lorestan University of Medical Sciences, Khorramabad, Iran
[5]Department of Health Information Technology, Aja University of Medical Sciences, Tehran, Iran
[6]Department of Health Information Technology, School of Medicine, Ardabil University of Medical Sciences, Ardabil, Iran

[*]*Corresponding author*: Student Research Committee, School of Health Management and Information Sciences Branch, Iran University of Medical Sciences, Tehran, Iran. Email: raoof.n1370@gmail.com

Received 2021 October 08; **Revised** 2022 January 12; **Accepted** 2022 February 02.

Abstract

Background: Early screening and diagnosis of breast cancer (BC) is critical for improving the quality of care and reducing the mortality rate.
Objectives: This study aimed to construct and compare the performance of several machine learning (ML) algorithms in predicting BC.
Methods: This descriptive and applied study included 1,052 samples (442 BC and 710 non-BC) with 30 features related to positive and negative BC diagnoses. The data mining (DM) process was implemented using the selected algorithm, including J-48 and random forest (RF) decision tree (DT), multilayer perceptron (MLP), Naïve Bayes (NB), Adaboost (AB), and logistics regression (LR) classifier. Then, we obtained the best algorithm by comparing their performances using the confusion matrix and area under the receiver operator characteristics (ROC) curve (AUC). Finally, we adopted the best model for BC prognosis.
Results: The results of evaluating various DM algorithms revealed that the J-48 DT algorithm had the best performance (AUC = 0.922), followed by the AB, MLP, LR, and RF algorithms (AUC: 0.899, 0819, 0.716, and 0.703, respectively). Also, the NB algorithm achieved the lowest performance in this regard (AUC = 0.669).
Conclusions: The ML presents a reasonable level of accuracy for an early diagnosis and screening of breast malignancies. Also, the empirical results showed that the J-48 DT algorithm yielded higher performance than other classifiers.

Keywords: Data Mining, Machine Learning, Artificial Intelligence, Breast Cancer, Decision Tree

1. Background

Today, breast cancer (BC) is considered the leading mortality culprit in the female population. In the Western communities, about 10% of women are susceptible to the disease, which makes up 12.5% worldwide (1, 2). In the United States, one in eight women is at greater risk of BC during their lifetime (3, 4). Additionally, it is estimated that about 200 million people will suffer from the disease annually in India by 2030, which is, per se, an epidemic for the country (5). Today, evidence implies that the BC is considered a global challenge due to its heterogeneous, multifactorial, violent nature, and destructive effects on health (6, 7). According to reports, it has been well established that malignant BC is often invasive and develops in the early stages in the mammary glands and ducts (8). It is followed by diffusion to the surrounding tissues, adjacent lymph nodes and metastasizes to the bones, liver, brain, or lungs in the advanced stages (9, 10). Most regretfully, many malignancies are diagnosed late in the advanced stages, with the tumor metastasizing to tissues around the breast, axillary lymph nodes, and even other organs (11, 12). Reportedly, Numerous clinical and nonclinical factors may affect the incidence of BC (13). Hence, the most effective way to reduce BC mortality is timely detection and treatment, which, in turn, necessitates faster diagnosis in the early stages.

Moreover, it is very demanding to differentiate between benign and malignant cancers in the initial diagnosis (14, 15). Therefore, to come up with an accurate and correct method for early detection is of great significance. A biopsy is the best way to diagnose benign or malignant cancers. However, it is an invasive and expensive proce-

dure (16). Also, physicians and cancer specialists usually analyze clinical and laboratory data manually and then opt for a relevant decision, making the method slow, expensive, time-consuming, and subjective (1). Given the different stages and severity of the disease and some ambiguities and unpredictable conditions about its consequences, it is imperative to adopt innovative technologies for screening (17). Also, so much research has focused on statistical methods and artificial intelligence (AI) in predicting cancer (16).

Recently, researchers have shown great interest in developing new and non-invasive digital technologies such as AI that can effectively prompt accurate and timely detection of malignancies (18). It is claimed that these technologies may minimize diagnostic errors and discrepancies among observers at any level of prediction, prognosis, and treatment. Therefore, diagnostic and prognostic models can help identify at-risk patients and adopt the most effective support and treatment programs (3, 19-21). Machine learning (ML), a branch of AI, can extract high-quality knowledge and patterns from a substantial raw dataset. Also, it can ease evidence-based risk analysis, screening, predictive, and care planning research and support reliable clinical decisions. Thus, it might improve patient care outcomes and quality and reduce uncertainty and ambiguity (3, 4, 22).

Data mining (DM) methods are used for BC in various areas, including early detection, differentiation of benign or malignant nature, prediction of patient survival after treatment, and the possibility of its recurrence (1, 2, 5, 23). It can also help physicians achieve significant results without dependency on invasive and complicated procedures (1). In this regard, many ML-based algorithms are applied for predicting and classifying BC outcomes.

2. Objectives

This study aimed to develop and evaluate the selected ML classifiers for early detection of BC and choose the best ones. We answered the following questions: (1) 'What are the most relevant predictors for predicting the BC?' and (2) 'How would the best ML algorithm be considered for the BC diagnosis, and how can it improve the BC diagnosis using the selected diagnostic factors?'

3. Methods

This retrospective single-center study aimed to develop a BC risk prediction model using the most popular ML algorithms and selecting the best performance.

3.1. Study Roadmap and Experiment Environment

All experiments on the ML models described in the present paper were run using Weka (version 3.9) in three phases, including dataset preprocessing, training, and evaluation. The ML models developed in Weka environment software are applied to various real-world issues. The Weka environment also provides an excellent framework for developers to run and evaluate their ML algorithms. The road map of the proposed system for the detection of BC is displayed in Figure 1.

3.2. Data Collection and Dataset Characteristics

In this study, models were trained and evaluated on the dataset of suspected BC cases from December 2017 to January 2021. The BC suspected case records were attained from the BC registry database in the Ayatollah Taleghani hospital in Abadan, Iran. Furthermore, the ethics board of Ilam University of Medical Science (ILUMS) approved the study design (code: IR.MEDILAM.REC.1399.294). Also, the registry database contained 2,854 records with 30 diagnostic variables. The variables are classified into six main categories as follows: (1) basic information such as nationality, education, age, job, body mass index (BMI), and the ratio of waist to the breast; (2) nutritional features, including salt intake, vegetable, and fruit consumption, dairy consumption, oil consumption, and fast food eating; (3) history of diseases such as fatness, hyperglyceridemia, hyperlipidemia, hypercholesterolemia, hypertension, common cold, and diabetes; (4) history of BC and interventions, e.g., the history of breast sampling, family history of BC, chest radiotherapy, and personal history of BC; (5) clinical manifestations such as a mass in the upper quarter of the breast or unspecified region of the breast; and (6) epidemiological factors like alcohol consumption, walking, physical activities, optimal physical activities, and heavy job activities as the independent variables in this study. The dependent variable was the BC diagnosis with two values of 0 and 1, which were associated with negative and positive BC diagnoses, respectively.

3.3. Preprocessing Dataset

Firstly, two health information managers (HIMs) (R-N and M-SH) and two cancer and gynecological specialists thoroughly reviewed the information of the dataset concerning quantitative and qualitative investigations. Also, implementing the ML algorithms was preceded by preprocessing the raw dataset. Of course, this stage is a common requirement for many ML predictions. For this purpose, we removed the samples with more than 70% missing values from the study, that were insignificant in statistical analysis. In the next step, the two methods of the K-Nearest Neighborhood with a specific amount of K and average

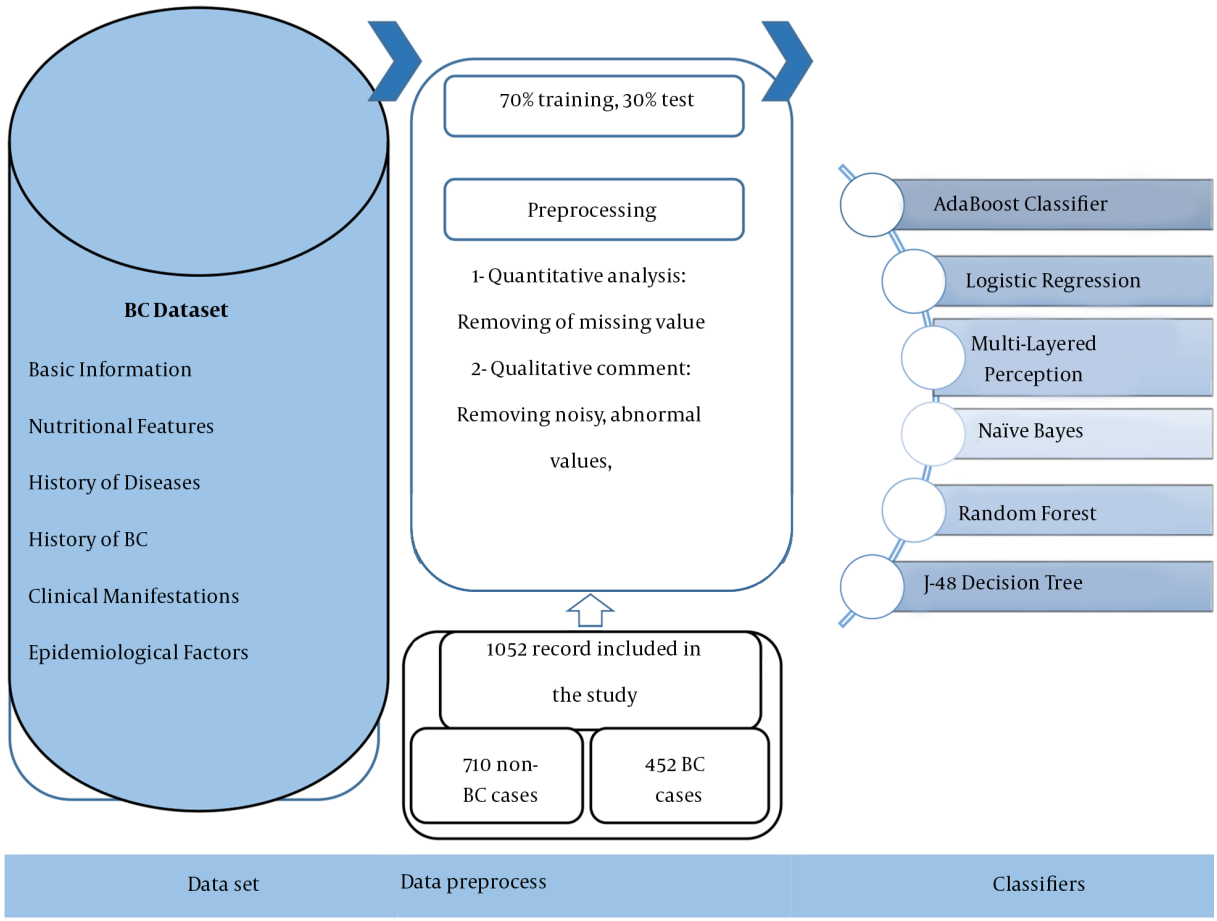

Figure 1. The study road map

available values were used for the cases having less than 70% missing values to embed the qualitative and quantitative variables, respectively.

3.4. Model Development and Assessment

Having normalized the dataset through cleaning the records with the high rate of missing values, we performed the predictive models using various DM algorithms with specified technical features in Weka V 3.9 software environment. The selected DM algorithms, including AB, LR, MLP, NB, J-48, and RF algorithms were leveraged in this respect as described below:

LR: This model can predict variables with two values, negative and positive diagnoses, using independent variables as the probabilistic model. It can also be named a classification algorithm in the ML process to classify the samples based on the training data (24, 25).

AB: This algorithm is commonly used for binary classification and can be categorized as the optimization algorithms (Boosting type) because of augmenting various DM performances, which can be used with the AB algorithm. Also, this algorithm is sensitive to noisy data, but a normalized dataset without any distorting data has a higher performance than other algorithms (26, 27).

NB: Another classification method for the ML process is NB. This algorithm is a statistical prediction model that predicts output variables using the input in a way that none of the input variables affect each other. On the other hand, no combinational input variables have strength for determining the probability of occurring the output variable (28, 29).

MLP: This algorithm consists of computational units known as neurons that exist in the input, hidden, and out layers of the artificial neural network (ANN). This algorithm tries to simulate the process structure of humans' mind by using the activation and linkage between neurons during training methods. The input layers get the informa-

tion from the environment. In the hidden layers having the neuron, the data can be processed, and finally, in the output layer, the required information will be represented. So far, the ANNs have had application in various fields, such as medicine, with their high calculation performance (30-32).

RF: This algorithm is known as the decision tree with a large size. It includes different tree algorithms named subtrees and uses the voting way for gaining performance. The subtrees with performance are in the majority and can be selected as the RF general performance. Also, in this algorithm, the splitting process occurs randomly and has the flexibility in classifying the research samples. The vast volume dataset is suitable for this type of decision tree (33, 34).

J-48: This algorithm is a newer version of the ID3 with high flexibility and capability. The splitting process in this decision tree occurs using the variables with the highest entropy difference than other variables. Therefore, research samples can be classified with the highest performance and most discriminative capability. The capability of J-48 decision tree algorithms allow to embed the continuous variables for DM, use the most technical features to prevent overfitting, and adjust the decision size with confidence factors (35, 36).

The confusion matrix has been used to measure each DM algorithm's capabilities in classification. They are calculated as follows: The true positive (TP) and true negative (TN) are considered the numbers of positive and negative cases with or without BC, and correctly classified through algorithms as positive and negative, respectively. False positive (FP) and false negative (FN) are also regarded as the numbers of non-BC and BC cases incorrectly classified as positive and negative with algorithms, respectively. Also, 70% and 30% of research data were used for training and test processes, respectively.

4. Results

The predetermined inclusion criteria revealed 2,148 cases belonging to the afflicted group, and non-afflicted cases were removed from the study. Also, 1,052 records (442 cases of positive BC diagnosis and 710 cases of negative BC diagnosis) were obtained and used for analysis. The descriptive analysis of all study variables with frequency in each of the two groups, including Negative and Positive of BC cases with bivariate statistical analysis using independence test of Chi-square in train and test modes, is represented in Table 1.

Based on the information given in Table 1, the variables of history of the common cold ($P_{\text{-train}} = 0.04$) ($P_{\text{-test}} = 0.01$), BC in the unspecified region ($P_{\text{-train}} = 0.01$) ($P_{\text{-test}} = 0.02$), and history of breast sampling ($P_{\text{-train}} = 0.03$) ($P_{\text{-test}} = 0.04$) had a statistically significant relationship.

The results of the selected DM algorithms comparison based on the training and testing confusion matrix are demonstrated in Tables 2 and 3.

According to the information in Tables 2 and 3, the J-48 decision tree algorithm with $TP_{\text{-train}} = 298$, $TP_{\text{-test}} = 128$, $FN_{\text{-train}} = 12$, and $FN_{\text{-test}} = 4$ had a more practical upper hand in classifying the cases that belonged to positive BC diagnosis than other DM algorithms during training and testing processes. Also, the MLP with $TN_{\text{-train}} = 477$, $TN_{\text{-test}} = 207$, $FP_{\text{-train}} = 20$, and $FP_{\text{-test}} = 6$ demonstrated the best strength in categorizing the cases with negative diagnoses in this regard.

Figure 2 depicts the AUC of each algorithm regarding each DM algorithm's capability for classifying the research samples in train and test modes. In this figure, the horizontal and vertical vertices are presented as specificity and sensitivity, respectively.

Comparing the AUC of all DM algorithms indicated that the J-48 with $AUC_{\text{-train}} = 0.9$ and $AUC_{\text{-test}} = 0.832$ had the best performance compared to other DM algorithms for classifying the cases associated with positive and negative diagnoses of BC cases. Moreover, the two algorithms of MLP ($AUC_{\text{-train}} = 0.813$ and $AUC_{\text{-test}} = 0.809$) and AB ($AUC_{\text{-train}} = 0.856$ and $AUC_{\text{-test}} = 0.802$) had an acceptable performance for classifying the cases with $AUC > 0.8$. In contrast, the NB algorithm with $AUC_{\text{-train}} = 0.663$ and $AUC_{\text{-test}} = 0.552$ obtained a lower capability than other DM algorithms. Generally, evaluating various DM algorithms' performance in this research showed that the J-48 decision tree algorithm with $AUC_{\text{-train}} = 0.9$ and $AUC_{\text{-test}} = 0.832$ had the best performance in diagnosing the negative and positive cases associated with the BC screening. It could also be considered a suitable clinical diagnostic model for BC screening. For this purpose, we have drawn the J-48 decision tree algorithm for diagnosing BC and described it in more detail (Figure 3).

All essential technical characteristics for building the tree model with more details are described as: number of batch size = 100, binary split = True, collapse tree = True, confidence factor = 0.15, number of decimal places = 2, number of folds = 3, the minimum number of objects = 2, subtree raising = True, number of seeds = 1, and unpruned = False.

As depicted in Figure 3, the variable of family history of BC is placed in the root node as the most crucial factor for diagnosing BC in this decision tree with Size =41 and leaves= 24. However, we used the pruning process for shortening the tree and augmenting the performance so that some variables may be removed in this process. We interpreted the two most straightforward clinical rules extracted from the J-48 decision tree algorithm for diagnosing BC.

(1) IF (FH of BC = 0) then BC = 0

Table 1. Description of All Research Variables with Bivariate Analysis Among Negative and Positive Samples

Variables	Value Codes	Negative Cases	Positive Cases	P-Value (Train)	P-Value (Test)
History of chest radiotherapy	Yes (1); No (2)	Yes (31%); No (69%)	Have (56%); Haven't (44%)	0.21	0.16
History of colorectal cancer	Yes (1); No (2)	Yes (18%); No (82%)	Have (36%); Haven't (64%)	0.42	0.27
Personal history of BC	Yes (1); No (2)	Yes (24%); No (76%)	Have (31%); Haven't (69%)	0.16	0.12
Hypertension	Yes (1); No (2)	Yes (28%); No (72%)	Have (64%); Haven't (36%)	0.18	0.1
Family history of BC	Yes (0); No (1)	Yes (26%); No (74%)	Have (39%); Haven't (61%)	0.63	0.27
Fruit consumption (average in days for five years)	< 100 (1); 100 - 200 (2); > 200 (3)	Low (< 100g) (15%); Medium (100 - 200g) (30%); High (> 200g) (55%)	Low (< 100g) (41%); Medium (100 - 200g) (48%); High (> 200g) (11%)	0.08	0.11
Alcohol consumption	Yes (1); No (2)	Yes (33%); No (67%)	Have (23%); Haven't (77%)	0.14	0.16
Hypercholesterolemia	Yes (1); No (2)	Yes (13%); No (87%)	Have (35%); Haven't (65%)	0.21	0.17
Physical activities (hours per day)	< 0.5 hours (1); 0.5 - 1 hours (2); >1 hours (3)	Low (0 - 0.5 hours) (25%); Medium (0.5 - 1 hours) (35%); High (> 1 hours) (40%)	Low (0 - 0.5 hours) (40%); Medium (0.5 - 1 hours) (40%); High (> 1 hours) (20%)	0.32	0.25
Fatness	Yes (1); No (2)	Yes (46%); No (54%)	Have (57%); Haven't (43%)	0.08	0.1
Vegetable consumption	<150 grams (1); 150 - 300g grams (2); >300 grams (3)	Low (< 150g) (11%); Medium (150 - 300g) (44%); High (> 300g) (45%)	Low (< 150g) (36%); Medium (150 - 300g) (51%); High (> 300g) (13%)	0.12	0.08
Age	-	44.28 (10.662)	39.26 (9.411)	0.14	0.12
BMI	-	19.996 (6.256)	24.441 (7.351)	0.09	0.16
Diabetes	Yes (1); No (2)	Yes (24%); No (76%)	Have (60%); Haven't (40%)	0.13	0.17
Upper in quadrants BC	Yes (1); No (2)	Yes (17%); No (83%)	Have (50%); Haven't P(50%)	0.1	0.11
The ratio of waist to pelvic	-	71.556 (11.225)	62.128 (7.253)	0.11	0.09
History of breast sampling	Yes (1); No (2)	Yes (33%); No (67%)	Have (54%); Haven't (46%)	0.03	0.04
Hyperlipidemia	Yes (1); No (2)	Yes (25%); No (75%)	Have (60%); Haven't (40%)	0.07	0.12
Heavy job activities	Yes (1); No (2)	Yes (18%); No (82%)	Have (31%); Haven't (69%)	0.07	0.16
BC in unspecified regions	Yes (1); No (2)	Yes (31%); No (69%)	Have (58%); Haven't (42%)	0.01	0.02
Hard job	Yes (1); No (2)	Yes (25%); No (75%)	Have (45%); Haven't (55%)	0.16	0.08
Walking	Yes (1); No (2)	Yes (23%); No (77%)	Have (50%); Haven't (50%)	0.17	0.15
Hyperglyceridaemia	Yes (1); No (2)	Yes (32%); No (68%)	Have (50%); Haven't (50%)	0.11	0.16
Genetic	Yes (1); No (2)	Yes (21%); No (79%)	Have (50%); Haven't (50%)	0.11	0.07
High consuming salt intake	Yes (1); No (2)	Yes (35%); No (65%)	Yes (59%); No (41%)	0.23	0.15
High dairy consumption	Yes (1); No (2)	Yes (14%); No (86%)	Yes (53%); No (47%)	0.13	0.16
High fast food consumption	Yes (1); No (2)	Yes (12%); No (88%)	Yes (58%); No (42%)	0.1	0.15
High oil consumption	Yes (1); No (2)	Yes (9%); No (91%)	Yes (57%); No (435)	0. 5	0.25
History of the common cold	Yes (1); No (2)	Yes (23%); No (77%)	Yes (45%); No (55%)	0.04	0.01
Optimal physical activities	Yes (1); No (2)	Yes (15%); No (85%)	Yes (48%); No (52%)	0.25	0.31

Table 2. All Selected Algorithms' Train Confusion Matrix Along with Technical Characteristics

NO	Algorithms	Important Technical Characteristics	TP	FN	FP	TN
1	AB	Number of decimal places = 2; Number of iterations = 10; Weight threshold = 100; Classifier type= decision stump	265	45	40	457
2	LR	Number of decimal places = 4; Ridge = 10e-8; Maximum number of iterations = 48	210	100	63	429
3	MLP	Number of hidden layers = 25; Learning rate = 0.3; Training time = 500; Validation threshold = 20	255	55	20	477
4	NB	Use kernel estimator = true; Use supervised discretization = false; Number of decimal places = 2	205	105	85	412
5	J-48	Confidence factor = 0.15; Number of folds = 3; Number of seeds = 1; Unpruned = false	298	12	21	476
6	RF	Number of iterations = 100; Number of execution slots = 1; Break Tie randomly = true	243	67	80	417

Table 3. All Selected Algorithms' Test Confusion Matrix Along with Technical Characteristics

NO	Algorithms	Important Technical Characteristics	TP	FN	FP	TN
1	AB	Number of decimal places = 2; Number of iterations = 10; Weight threshold = 100; Classifier type = decision stump	113	19	20	193
2	LR	Number of decimal places =4; Ridge = 10e-8; Maximum number of iterations = 48	102	30	40	173
3	MLP	Number of hidden layers = 25; Learning rate = 0.3; Training time = 500; Validation threshold = 20	116	16	6	207
4	NB	Use kernel estimator = true; Use supervised discretization = false; Number of decimal places = 2	85	47	39	174
5	J-48	Confidence factor = 0.15; Number of folds = 3; Number of seeds = 1; Unpruned = false	128	4	9	204
6	RF	Number of iterations = 100; Number of execution slots = 1; Break Tie randomly = true	105	27	26	187

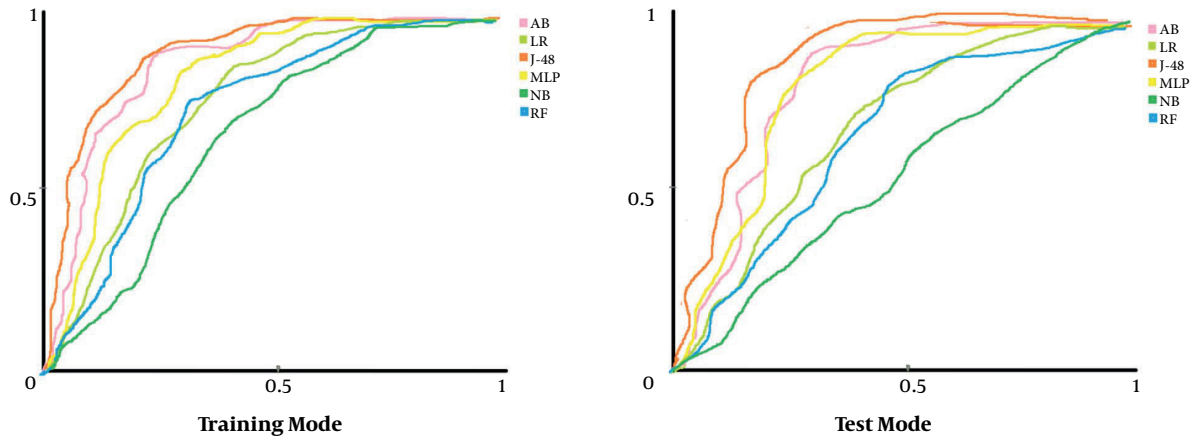

Figure 2. The ROC diagrams of selected DM algorithms in two training and test modes

(2) IF (FH of BC = 1) & (hypertension =1) then BC = 1

Rule 1 implies that the model will classify the individual with no family history of BC as the negative diagnosis of BC (148 samples were identified based on this pattern). Rule 2 states that one person has a family history of BC, and a history of hypertension is assigned to the positive group via the J-48 decision tree algorithm with 14 confirmed cases.

5. Discussion

Owing to the heterogeneous, complex, and invasive nature of BC, which requires understanding the non-linear interrelation between the modifiable and non-modifiable risk factors, the ML algorithms are applicable for cancer prognosis and screening (3). This study aimed to construct an intelligent predictive model via leveraging the selected ML algorithms to predict the BC and effectively differentiate between positive and negative BC cases. We trained six well-known classification algorithms, including AB, LR, MLP NB, J-48, and RF, according to the top related parame-

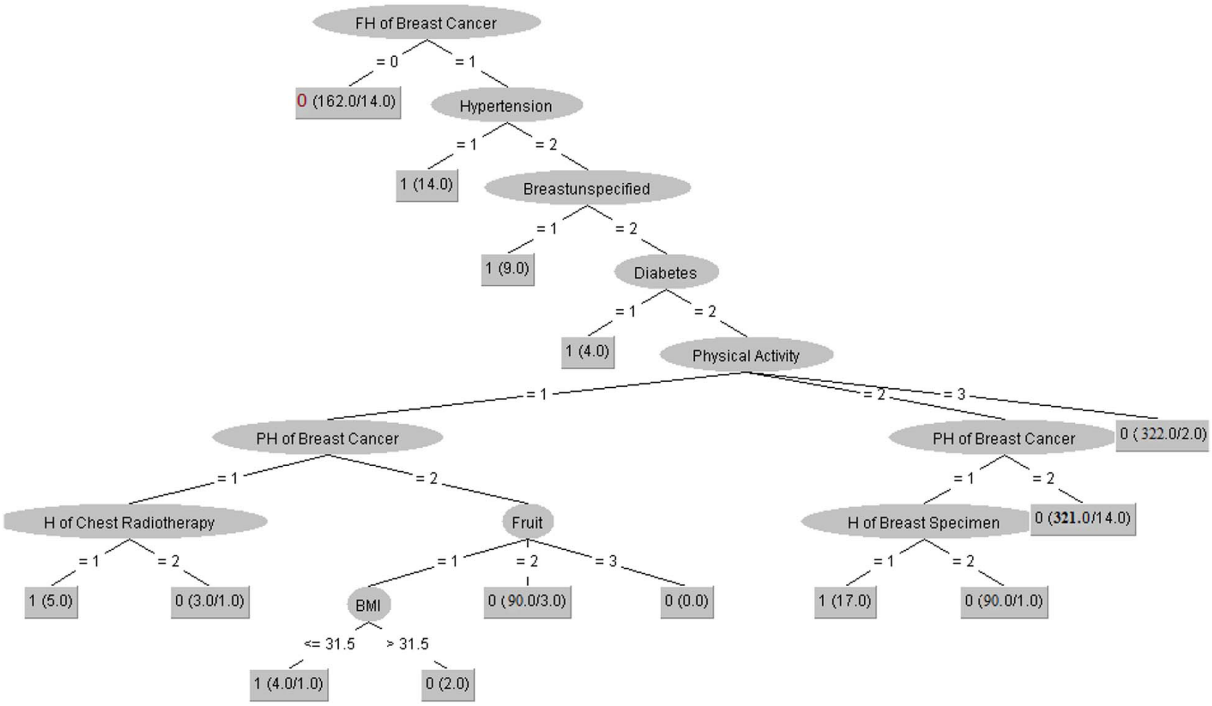

Figure 3. The pruned J-48 decision tree algorithm

ters affecting the risk of BC derived. Several prediction factors are investigated in the studies for BC prediction, such as breast medical images (34, 35), lesion biopsy (21), blood tests (36), etc. However, we considered more cost-benefit and available data with the minor intervention features for our prediction models.

Some studies evaluated the combination of the lifestyle, history of diseases, and demographic data to predict BC. The selected features are used as input for developing ML-based predictive models (37, 38). By recognizing patterns within the large amounts of data, it may be applied to gain more insight into the diseases and produce knowledge that can potentially inspire further research in many areas of medicine (39-41).

In the present study, we identified an efficient and effective classifier for BC prediction, which may lead to a more accurate model. Recently, the application of ML algorithms in healthcare has been attracting the attention of researchers (42, 43). Many studies have been performed using different ML algorithms to diagnose and predict various malignancies (44). So far, several studies have evaluated the DT algorithm's application in BC risk classification and prediction based on clinical variables (44, 45).

A study conducted by Williams et al. (46) showed that DM approaches have significant predictive power for BC. They indicated that the DT had the best accuracy compared to other techniques. In a systematic review by Li et al., the results showed the most frequently used ML methods for BC prediction from 2013 to 2020 were DT classifiers (19 studies, 61.3%) (42). Besides, the study by Park et al. showed that the best meaningful results were observed from the DT model with an accuracy of 90% (47). In our research, the DT approach had a high accuracy of about 96%. This may prove the potent power of DTs in predicting BC. Accordingly, Higa et al. (48) introduced DT and neural network as the best models for diagnosing benign and malignant tumors of BC with 95% accuracy. Rajinikanth et al.'s research showed that DT had the best predictive performance with an accuracy of > 92% (43). Solanki et al. also investigated the prediction of benign or malignant BC using selected ML techniques. The results showed that the J-48 yielded the best classification performance with an accuracy of 98.83% (49).

The result of other studies confirm the better performance of DT than other similar algorithms in predicting the BC risk (44, 45, 50, 51). Similarly, in this research, different ML algorithms were used to classify the data, and the DT algorithm had a higher efficiency than other algorithms. Accordingly, in the current study, the results showed that the J-48 DT algorithm with AUC = 0.922 had the best capability for early prediction of BC.

However, this study had some significant limitations

to be addressed. First, this is a retrospective study that suffers from meaningless values. Second, the proposed model's generalizability was confined by a single-center dataset with a limited sample size. Third, we used six ML algorithms for prediction analyses based on some clinical features. Hence, the performance accuracy of our model and its generalizability will be enhanced provided that we test more ML techniques on a larger, multicenter, and prospective dataset enriched with more qualitative and validated data. We applied factors from different clinical and nonclinical aspects in this research. Hence, it provides a better plan for clinicians to improve patient outcomes and quality of care as a clinical guideline in terms of BC diagnosis to a large extent. It may also minimize the ambiguity and sophistication in BC diagnosis through a practical and systematic knowledge representation method, including various diagnostic factors to facilitate BC screening and optimize the episode of care planning.

5.1. Conclusions

The present study indicated that ML-based prediction systems are powerful tools to predict the BC based on the predictor variables. We identified some general clinical factors which can contribute to an accurate prognosis for patients with BC. The proposed prediction model (J-48) can predict the BC risk for each case with an AUC_{train} = 0.9 and AUC_{test} = 0.832. As a result, it can be used as an essential clinical screening tool for the early prevention of BC. The underlying model may have the potential to augment informed decisions for early prognosis and effective screening of BC by offering an objective, systematic, and evidence-based approach.

Acknowledgments

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

Footnotes

Authors' Contribution: R.N. conceived and designed the evaluation and drafted the manuscript. M.SH. participated in designing the evaluation, performed parts of the statistical analysis and helped to draft the manuscript. L.E. and N.M re-evaluated the clinical data, revised the manuscript and performed the statistical analysis, and revised the manuscript. M.A. collected the clinical data, interpreted them, and revised the manuscript. M.M. re-analyzed the clinical and statistical data and revised the manuscript. All authors read and approved the final manuscript.

Conflict of Interests: The authors declare that they have no conflicts of interest.

Data Reproducibility: It was not declared by the authors.

Ethical Approval: The ethics board of Ilam University of Medical Science (ILUMS) approved the study design (code: IR.MEDILAM.REC.1399.294).

Funding/Support: This article has been extracted from a research project supported by the Ilam University of Medical Sciences.

References

1. Oskouei RJ, Kor NM, Maleki SA. Data mining and medical world: breast cancers' diagnosis, treatment, prognosis and challenges. *Am J Cancer Res*. 2017;**7**(3):610–27. [PubMed: 28401016]. [PubMed Central: PMC5385648].
2. Gupta S, Kumar D, Sharma AK. Data mining classification techniques applied for breast cancer diagnosis and prognosis. *Indian J Comput Sci Eng*. 2011;**2**(2):188–95.
3. Chaurasia V, Pal S. Applications of Machine Learning Techniques to Predict Diagnostic Breast Cancer. *SN Comput Sci*. 2020;**1**(5). doi: 10.1007/s42979-020-00296-8.
4. Prasuna K, Rao KR, Saibaba CHMH. Application of Machine Learning Techniques in Predicting Breast Cancer-A Survey. *IJITEE*. 2019;**8**(8):826–32.
5. Chaurasia V, Pal S. Data mining techniques: to predict and resolve breast cancer survivability. *IJCSMC*. 2014;**3**(1):10–22.
6. Anwar SL, Dwianingsih EK, Avanti WS, Choridah L, Aryandono T, Suwardjo. Aggressive behavior of Her-2 positive colloid breast carcinoma: A case report in a metastatic breast cancer. *Ann Med Surg (Lond)*. 2020;**52**:48–52. doi: 10.1016/j.amsu.2020.02.010. [PubMed: 32211189]. [PubMed Central: PMC7082430].
7. Babiera GV. Metastatic breast cancer: a paradigm shift toward a more aggressive approach. *Cancer J*. 2009;**15**(1):78. doi: 10.1097/PPO.0b013e318197686b. [PubMed: 19197179].
8. Maeshima Y, Osako T, Morizono H, Yunokawa M, Miyagi Y, Kikuchi M, et al. Metastatic ovarian cancer spreading into mammary ducts mimicking an in situ component of primary breast cancer: a case report. *J Med Case Rep*. 2021;**15**(1):78. doi: 10.1186/s13256-020-02653-w. [PubMed: 33593410]. [PubMed Central: PMC7887787].
9. Beachler DC, de Luise C, Yin R, Gangemi K, Cochetti PT, Lanes S. Predictive model algorithms identifying early and advanced stage ER+/HER2- breast cancer in claims data. *Pharmacoepidemiol Drug Saf*. 2019;**28**(2):171–8. doi: 10.1002/pds.4681. [PubMed: 30411431].
10. Franzoi MA, Rosa DD, Zaffaroni F, Werutsky G, Simon S, Bines J, et al. Advanced Stage at Diagnosis and Worse Clinicopathologic Features in Young Women with Breast Cancer in Brazil: A Subanalysis of the AMAZONA III Study (GBECAM 0115). *J Glob Oncol*. 2019;**5**:1–10. doi: 10.1200/JGO.19.00263. [PubMed: 31730380]. [PubMed Central: PMC6882517].
11. Tesfaw A, Getachew S, Addissie A, Jemal A, Wienke A, Taylor L, et al. Late-Stage Diagnosis and Associated Factors Among Breast Cancer Patients in South and Southwest Ethiopia: A Multicenter Study. *Clin Breast Cancer*. 2021;**21**(1):e112–9. doi: 10.1016/j.clbc.2020.08.011. [PubMed: 33536135].
12. Gebremariam A, Addissie A, Worku A, Assefa M, Kantelhardt EJ, Jemal A. Perspectives of patients, family members, and health care providers on late diagnosis of breast cancer in Ethiopia: A qualitative study. *PLoS One*. 2019;**14**(8). e0220769. doi: 10.1371/journal.pone.0220769. [PubMed: 31369640]. [PubMed Central: PMC6675093].
13. Dowsett M, Sestak I, Regan MM, Dodson A, Viale G, Thurlimann B, et al. Integration of Clinical Variables for the Prediction of Late Distant Recurrence in Patients With Estrogen Receptor-Positive Breast Cancer Treated With 5 Years of Endocrine Therapy: CTS5. *J*

Clin Oncol. 2018;**36**(19):1941–8. doi: 10.1200/JCO.2017.76.4258. [PubMed: 29676944]. [PubMed Central: PMC6049399].
14. Che Mohamed N, Moey SF, Lim BC. Validity and Reliability of Health Belief Model Questionnaire for Promoting Breast Self-examination and Screening Mammogram for Early Cancer Detection. *Asian Pac J Cancer Prev*. 2019;**20**(9):2865–73. doi: 10.31557/APJCP.2019.20.9.2865. [PubMed: 31554389]. [PubMed Central: PMC6976832].
15. Anderson BO, Bevers TB, Carlson RW. Clinical Breast Examination and Breast Cancer Screening Guideline. *JAMA*. 2016;**315**(13):1403–4. doi: 10.1001/jama.2016.0686. [PubMed: 27046372].
16. Faisal KA. Evaluation of Breast Cancer Tumor Classification with Unconstrained Functional Networks Classifier. *the 4th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA-06)*. Dubai , UAE. King Fahd University of Petroleum and Minerals; 2006.
17. Tseng YJ, Huang CE, Wen CN, Lai PY, Wu MH, Sun YC, et al. Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. *Int J Med Inform*. 2019;**128**:79–86. doi: 10.1016/j.ijmedinf.2019.05.003. [PubMed: 31103449].
18. Halim E, Halim PP, Hebrard M. Artificial Intelligent Models for Breast Cancer Early Detection. *2018 International Conference on Information Management and Technology (ICIMTech)*. IEEE; 2018. p. 517–21.
19. Dawngliani MS, Chandrasekaran N, Lalmuanawma S, Thangkhanhau H. Prediction of Breast Cancer Recurrence Using Ensemble Machine Learning Classifiers. *International Conference on Security with Intelligent Computing and Big-data Services*. Springer; 2020.
20. Gupta P, Garg S. Breast Cancer Prediction using varying Parameters of Machine Learning Models. *Procedia Comput Sci*. 2020;**171**:593–601. doi: 10.1016/j.procs.2020.04.064.
21. Yue W, Wang Z, Chen H, Payne A, Liu X. Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis. *Designs*. 2018;**2**(2). doi: 10.3390/designs2020013.
22. Sathya D, Sudha V, Jagadeesan D. Application of Machine Learning Techniques in Healthcare. *Handbook of Research on Applications and Implementations of Machine Learning Techniques*. IGI Global; 2020. p. 289–304.
23. Silva J, Lezama OBP, Varela N, Borrero LA. Integration of Data Mining Classification Techniques and Ensemble Learning for Predicting the Type of Breast Cancer Recurrence. *International Conference on Green, Pervasive, and Cloud Computing*. Springer; 2019. p. 18–30.
24. Shaker Abdalrada A, Hashim Yahya O, Hadi M. Alaidi A, Ali Hussein N, Th. Alrikabi H, Al-Quraishi TA. A Predictive model for liver disease progression based on logistic regression algorithm. *Period Eng Nat Sci*. 2019;**7**(3). doi: 10.21533/pen.v7i3.667.
25. De Caigny A, Coussement K, De Bock KW. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *Eur J Oper Res*. 2018;**269**(2):760–72. doi: 10.1016/j.ejor.2018.02.009.
26. Yong Z, Jianyang L, Hui L, Xuehui G. Fatigue driving detection with modified ada-boost and fuzzy algorithm. *2018 Chinese Control And Decision Conference (CCDC)*. IEEE; 2018. p. 5971–4.
27. Zhang T, Chen W, Li M. Recognition of epilepsy electroencephalography based on AdaBoost algorithm. *Acta Physica Sinica*. 2015;**64**(12). doi: 10.7498/aps.64.128701.
28. Khotimah B, Miswanto M, Suprajitno H. Optimization of Feature Selection Using Genetic Algorithm in Naïve Bayes Classification for Incomplete Data. *Int J Intell Eng Syst*. 2020;**13**(1):334–43. doi: 10.22266/ijies2020.0229.31.
29. Marikani T, Shyamala K. Modified Multinomial Naïve Bayes Algorithm for Heart Disease Prediction. *Intelligent Communication Technologies and Virtual Mobile Networks*. Springer; 2020. p. 294–300.
30. Darvishan A, Bakhshi H, Madadkhani M, Mir M, Bemani A. Application of MLP-ANN as a novel predictive method for prediction of the higher heating value of biomass in terms of ultimate analysis. *Energy Sources A: Recovery Util Environ Eff*. 2018;**40**(24):2960–6. doi: 10.1080/15567036.2018.1514437.
31. Borghi PH, Zakordonets O, Teixeira JP. A COVID-19 time series forecasting model based on MLP ANN. *Procedia Comput Sci*. 2021;**181**:940–7. doi: 10.1016/j.procs.2021.01.250. [PubMed: 33936325]. [PubMed Central: PMC8076817].
32. Benyekhlef A, Mohammedi B, Hassani D, Hanini S. Application of artificial neural network (ANN-MLP) for the prediction of fouling resistance in heat exchanger to MgO-water and CuO-water nanofluids. *Water Sci Technol*. 2021;**84**(3):538–51. doi: 10.2166/wst.2021.253. [PubMed: 34388118].
33. Roshanaei G, Omidi T, Faradmal J, Safari M, Poorolajal J. [Determining affected factors on survival of kidney transplant in living donor patients using a random survival forest]. *Koomesh*. 2018;**20**(3):517–23. Persian.
34. Javeed A, Zhou S, Yongjian L, Qasim I, Noor A, Nour R. An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection. *IEEE Access*. 2019;**7**:180235–43. doi: 10.1109/access.2019.2952107.
35. Ripon SH. Rule induction and prediction of chronic kidney disease using boosting classifiers, Ant-Miner and J48 Decision Tree. *2019 international conference on electrical, computer and communication engineering (ECCE)*. IEEE; 2019.
36. Maliha SK, Islam T, Ghosh SK, Ahmed H, Mollick MRJ, Ema RR. Prediction of Cancer Using Logistic Regression, K-Star and J48 algorithm. *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*. IEEE; 2019. p. 1–6.
37. Al-Salihy NK, Ibrikci T. Classifying breast cancer by using decision tree algorithms. *Proceedings of the 6th International Conference on Software and Computer Applications - ICSCA '17*. New York, NY, United States. Association for Computing Machinery; 2017. p. 144–8.
38. Ozkan GY, Gunduz SY. Comparision of Classification Algorithims for Survival of Breast Cancer Patients. *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE; 2020. p. 1–4.
39. Ibrahim S, Nazir S, Velastin SA. Feature Selection Using Correlation Analysis and Principal Component Analysis for Accurate Breast Cancer Diagnosis. *J Imaging*. 2021;**7**(11). doi: 10.3390/jimaging7110225. [PubMed: 34821856]. [PubMed Central: PMC8625715].
40. Khandezamin Z, Naderan M, Rashti MJ. Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier. *J Biomed Inform*. 2020;**111**:103591. doi: 10.1016/j.jbi.2020.103591. [PubMed: 33039588].
41. Lopez NC, Garcia-Ordas MT, Vitelli-Storelli F, Fernandez-Navarro P, Palazuelos C, Alaiz-Rodriguez R. Evaluation of Feature Selection Techniques for Breast Cancer Risk Prediction. *Int J Environ Res Public Health*. 2021;**18**(20). doi: 10.3390/ijerph182010670. [PubMed: 34682416]. [PubMed Central: PMC8535206].
42. Li J, Zhou Z, Dong J, Fu Y, Li Y, Luan Z, et al. Predicting breast cancer 5-year survival using machine learning: A systematic review. *PLoS One*. 2021;**16**(4). e0250370. doi: 10.1371/journal.pone.0250370. [PubMed: 33861809]. [PubMed Central: PMC8051758].
43. Rajinikanth V, Kadry S, Taniar D, Damasevicius R, Rauf HT. Breast-Cancer Detection using Thermal Images with Marine-Predators-Algorithm Selected Features. *2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII)*. IEEE; 2021. p. 1–6.
44. Ortega JHJ, Resureccion MR, Natividad LRQ, Bantug ET, Lagman AC, Lopez SR. An Analysis of Classification of Breast Cancer Dataset Using J48 Algorithm. *Int J Adv Trends Comput Sci Eng*. 2020;**9**(1.3):475–80. doi: 10.30534/ijatcse/2020/7591.32020.
45. Ghiasi MM, Zendehboudi S. Application of decision tree-based ensemble learning in the classification of breast cancer. *Comput Biol Med*. 2021;**128**:104089. doi: 10.1016/j.compbiomed.2020.104089. [PubMed: 33338982].
46. Williams K, Adebayo Idowu P, Ademola Balogun J, Ishola Oluwaranti A. Breast Cancer Risk Prediction Using Data Mining Classification Techniques. *Trans Netw Commun*. 2015;**3**(2). doi: 10.14738/tnc.32.662.
47. Park EY, Yi M, Kim HS, Kim H. A Decision Tree Model for Breast Reconstruction of Women with Breast Cancer: A Mixed Method Approach. *Int J Environ Res Public Health*. 2021;**18**(7). doi: 10.3390/ijerph18073579.

[PubMed: 33808263]. [PubMed Central: PMC8036358].
48. Higa A. Diagnosis of breast cancer using decision tree and artificial neural network algorithms. *Cell*. 2018;**1**:10.
49. 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**(6). doi: 10.3390/electronics10060699.
50. Shankar JR, Nithish S, Babu MN, Karthik R, Afridi AS. Breast Cancer Prediction using Decision Tree. *Journal of Physics: Conference Series*. IOP Publishing; 2021.
51. Kurian B, Jyothi VL. Breast cancer prediction using an optimal machine learning technique for next generation sequences. *Concurr Eng*. 2021;**29**(1):49–57. doi: 10.1177/1063293x21991808.