ISSN: 2345-2897; Health Education and Health Promotion. 2022;10(1):89-97.

Discovering the Clinical Knowledge about Breast Cancer Diagnosis Using Rule-Based Machine Learning Algorithms

ARTICLE INFO

Article Type
Descriptive study

Authors
Nopour R.[1] *MSc,*
Kazemi-Arpanahi H.[2] *PhD,*
Shanbehzadeh M.*[3] *PhD*

How to cite this article
Nopour R, Kazemi-Arpanahi H, Shanbehzadeh M. Discovering the Clinical Knowledge about Breast Cancer Diagnosis Using Rule-Based Machine Learning Algorithms. Health Education and Health Promotion. 2022;10(1):89-97.

ABSTRACT

Aims Breast cancer represents one of the most prevalent cancers and is also the main cause of cancer-related deaths in women globally. Thus, this study was aimed to construct and compare the performance of several rule-based machine learning algorithms in predicting breast cancer.

Instrument & Methods The data were collected from the Breast Cancer Registry database in the Ayatollah Taleghani Hospital, Abadan, Iran, from December 2017 to January 2021 and had information from 949 non-breast cancer and 554 breast cancer cases. Then the mean values and K-nearest neighborhood algorithm were used for replacing the lost quantitative and qualitative data fields, respectively. In the next step, the Chi-square test and binary logistic regression were used for feature selection. Finally, the best rule-based machine learning algorithm was obtained based on comparing different evaluation criteria. The Rapid Miner Studio 7.1.1 and Weka 3.9 software were utilized.

Findings As a result of feature selection the nine variables were considered as the most important variables for data mining. Generally, the results of comparing rule-based machine learning demonstrated that the J-48 algorithm with an accuracy of 0.991, F-measure of 0.987, and also AUC of 0.9997 had a better performance than others.

Conclusion It's found that J-48 facilitates a reasonable level of accuracy for correct BC risk prediction. We believe it would be beneficial for designing intelligent decision support systems for the early detection of high-risk patients that will be used to inform proper interventions by the clinicians.

Keywords Machine Learning; Artificial Intelligence; Data Mining; Breast Neoplasms; Decision Tree

[1]Department of Health Information Management, Student Research Committee, School of Health Management and Information Sciences Branch, Iran University of Medical Sciences, Tehran, Iran
[2]Department of Health Information Technology, Abadan University of Medical Sciences, Abadan, Iran
[3]Department of Health Information Technology, School of Paramedical, Ilam University of Medical Sciences, Ilam, Iran

Correspondence
Address: Ilam, Bangangab, Research Blvd., Ilam University of Medical Sciences Campus. Postal code: 6939177143
Phone: +98 (930) 0833691
Fax: +98 (84) 32223039
mostafa.shanbezadeh@gmail.com

Article History
Received: September 14, 2021
Accepted: November 28, 2021
ePublished: April 10, 2022

CITATION LINKS

[1] Automated breast cancer diagnosis based on ... [2] Comparison of the performance of machine learning algorithms ... [3] Breast cancer population screening program results in ... [4] Aggressive behavior of Her-2 positive colloid ... [5] A paradigm shift toward a more aggressive ... [6] Metastatic ovarian cancer spreading into mammary ducts ... [7] Advanced stage at diagnosis and worse clinicopathologic ... [8] Late-stage diagnosis and associated factors among ... [9] Perspectives of patients, family members, and health ... [10] New insights into the screening, prompt ... [11] A comparative study of mammography, sonography ... [12] Validity and reliability of health belief model ... [13] Clinical breast examination and breast ... [14] Integration of clinical variables for the prediction of ... [15] Predicting breast cancer metastasis by ... [16] Computational radiology in breast cancer screening ... [17] Drug and hormone resistance in ... [18] Handbook of research on applications ... [19] Personalized pancreatic cancer management ... [20] Application of machine learning techniques ... [21] Machine learning with applications in breast ... [22] A machine learning approach to uncovering ... [23] Prediction of breast cancer using rule ... [24] Breast cancer diagnosis using feature ... [25] Breast cancer disease classification ... [26] Survival prediction of patients with breast ... [27] Integration of data mining classification ... [28] Imbalanced machine learning based techniques ... [29] A hybrid supervised machine learning ... [30] Applications of machine learning techniques ... [31] Analysis of breast cancer detection using ... [32] Comparison of decision tree methods for breast ... [33] Prediction of breast cancer recurrence ... [34] Comparative study on different classification ... [35] Classifying breast cancer by using ... [36] An analysis of classification of breast ...

Introduction

Breast Cancer (BC) is the most fatal and frequent malignancy among women with an estimated 11.7% of all cancer cases and about 20% of all cancer-related deaths. Globally it t is the second leading cause of cancer death among people (men and women) after lung malignancies in both developing and developed countries [1]. Based on the global cancer report, BC was the most commonly diagnosed cancer in 2020, with 2.3 million new cases [2]. Early detection and screening can significantly decrease patient costs, improve the overall likelihood of treatment and survivability [3].

Today, evidence suggests that BC is a global challenge due to its heterogeneous, multifactorial, violent nature, and destructive health effects [4, 5]. Reportedly, it is now well established that the malignant BC is often aggressive, forming in the early stages in the glands and mammary ducts [6] and then metastasizing to the surrounding tissues, adjacent lymph nodes, and, specifically, to the bones, liver, brain, or lungs in the advanced stages [7]. Most regrettably, many cases of malignancy are detected late in the advanced stages of the disease such that the tumor has metastasized to the tissues around the breast, axillary lymph nodes, and even other organs [8, 9]. Therefore, there is a growing body of literature that recognizes the benefits of systematic and up-to-date screening policies in this regard [10].

The most well-known methods for screening the disease include mammography, thermography, and tissue sampling techniques which are thoroughly implemented more seriously in many developed countries [11]. However, the mentioned screening methods are time-consuming, expensive, and highly complicated. On the other hand, recently, there has been renewed interest in some techniques, including breast self-examination (BSE) and clinical breast examination (CBE). Despite their cheapness and availability, studies have reported challenging and different results on their effectiveness [12, 13].

There are several clinical and non-clinical factors influencing the incidence of BC [14]. Due to the different stages and severity of BC and the existence of some ambiguities and unpredictable situations regarding its outcomes, which, in turn, necessitates adopting innovative technologies for screening [15].

Recently, researchers have shown an increased interest in the deployment of newly-developed digital technological and non-invasive methods such as artificial intelligence (AI) systems which can be effective in rapid, accurate, and timely diagnosis of malignancies [16]. Specifically, the rapid diagnosis of cancers in the early stages is considered the most significant factor for definitive treatment of the disease, prevention of unpleasant complications, and increasing patients' survival chances [17]. Machine learning (ML), a subset of AI, has many applications in many industries, including healthcare [18]. The ML plays a crucial role in managing malignancies such as prognosis, diagnosis, and treatment outcomes from the big data available in the medical field [19].

In the last few decades, several ML-based methods have been developed for the effective and timely prognosis and screening of BC [20, 21]. These methods will support decisions by extracting hidden patterns and applied knowledge from the raw dataset [22].

The clinical decision support systems (CDSSs) based on rule-based logic [23] and decision tree (DT) algorithms [24] are considered useful, practical, and flexible tools for modeling medical diagnoses and supporting complex decisions [23]. Rule-based machine learning (RBML) is increasingly adopted due to different stages and degrees of severity and some ambiguities and unpredictable situations in the behavior and outcome of the disease besides various clinical and non-clinical factors involved in BC emergence and progression [23, 25]. So far, several studies have been evaluating the application of ML algorithms in BC risk classification and prediction based on clinical variables.

Momenyan *et al.* developed an optimum ML-based intelligent model for classifying the BC risk [26]. Researchers compared three different ML algorithms for BC risk classification [27, 28]. In another study conducted by Solanki and their colleagues, they investigated the prediction of benign or malignant BC using selected ML techniques [29]. Finally, Salod *et al.* in their work compared the performance of eight ML algorithms in BC screening and detection [2].

In recent years, many RBML techniques are applied to predicting BC and classifying disease outcomes. Therefore, this study was aimed to develop an appropriate and scientific screening model based on the selected RBML for earlier detection of the disease, improve diagnostic efficiency and decrease the risk of mortalities caused by BC.

Instrument and Methods

This retrospective single-center study aimed to develop a BC risk prediction model using seven popular RBML algorithms and selecting the best performing. Models were trained and evaluated on the data of suspected BC from December 2017 to January 2021. BC cases were extracted from the BC Registry database in the Ayatollah Taleghani Hospital, Abadan, Iran. The Registry database contains 2854 patient records with 30 features. The independent features (input) are categorized into 6 main classes patient characteristics, nutritional factors, medical history, history of BC and related interventions, clinical manifestations, and epidemiological factors input variables. The dependent variable (output) is the diagnosis of BC by two values of 0 and 1 associated with non-BC and BC cases, respectively. Primary variables of the registry database associated with the BC prognosis

are listed below:
- Demographic: Age, job, education, nationality, the ratio of waist to breast, and Body Mass Index;
- History of diseases: Salt, vegetable, dairy, fruit (average in days from 5 years ago), fast food, and oil consumption;
- Nutritional factors: Diabetes, common cold, hyperlipidemia, hyperglyceridaemia, hypercholesterolemia, hypertension, and fatness;
- History of breast cancer and interventions: A personal history of breast cancer, history of breast sampling, history of chest radiotherapy, and family history of breast cancer;
- Clinical manifestations: Exist a mass in the upper quarter of the breast or the unspecified region of the breast;
- Epidemiological factors: Walking, heavy job, physical, optimal physical activities, and alcohol consumption;
- Outcome: BC and non-BC.

After applying exclusion criteria, ultimately the 1668 case records were chosen for the study (Diagram 1).

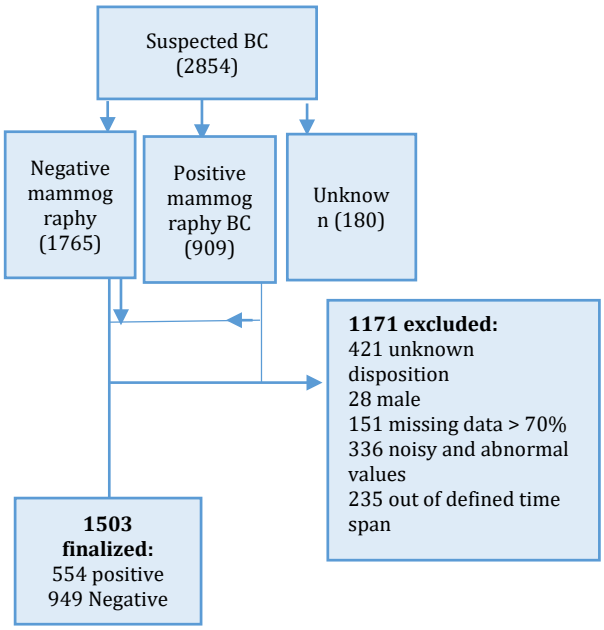

Diagram 1) Flow chart describing patient selection

The Abadan University of Medical Science ethics board approved the study design. Before implementing the ML algorithms, preprocessing was performed on the raw dataset. This stage is a common requirement for many ML predictions. For this purpose, we removed the samples with more than 70% missing data from the analysis. Then, for other missing fields, we used the average of the existing available values and the K-nearest neighborhood (KNN) Euclidean distance for the quantitative and qualitative variables, respectively. The model's implementation was done by Rapid Miner Studio 7.1.1 environment.

To select the best predictors and reduce the dataset dimension, the independent Chi-square test was used for determining the relationship between each independent variable (30 variables) with the dependent (BC diagnosis: Yes or No) as the output class. The $p<0.01$ is considered as a statistically significant level in this respect. After determining the most important affecting factors in BC, we trained a set of RBML algorithms such as J-48, random-forest (RF), random-tree (RT), and REP-Tree, decision table (DT), J-RIP, and Part were applied to classify the diagnosis value of the dataset and eliciting the knowledge about BC classification with IF-THEN structure. These techniques are used for discovering the knowledge and hidden patterns that existed for diagnosing BC in the dataset. The Weka software 3.9 was utilized in this respect.

In the last phase, the performance of all algorithms was assessed by criteria such as positive predicted value (PPV), negative predicted value (NPV), sensitivity, specificity, accuracy, F-score, and are calculating the area under receiver operator characteristics (AUC-ROC). The confusion matrix has been used for measuring the capabilities of each data mining algorithm in classification. They are calculated as follows: The True Positive (TP) and True Negative (TN) are the numbers of positive and negative cases that have and do non-having BC and are truly classified by algorithms as positive and negative, respectively. False Positive (FP) and False Negative (FN) are also the numbers of non-BC and BC cases that are incorrectly classified as positive and negative cases by algorithms, respectively. The 10-fold cross-validation has been utilized for determining and comparing all data mining performance for considering the errors that existed in algorithms performance calculation in this respect. After determining the best algorithm using different performance criteria, in the last step, the best knowledge for diagnosing BC has been obtained using the IF-THEN structure, and the rules with the more classified samples were considered the main knowledge for diagnosing BC.

Findings

The 554 and 949 cases associated with the positive and negative BC cases, respectively have remained and were included for statistical analysis. The mean age of the afflicted women was 48.146±13.074 years and in non-afflicted cases was 43.212±9.70 years. Table 1 shows the basic data of the two groups of individuals.

Based on the results, 18 variables had a significant relationship with diagnosing the BC using the Chi-square test at $p<0.01$. The variables of upper in quadrants breast cancer, history of chest radiotherapy, and fatness were considered as the most three important factors for diagnosing the BC at $p<0.001$ (Table 2).

Table 1) The frequency results of demographic variables

Variable		Number (%)
Age (year)	25-35	178 (11.84)
	36-45	312 (20.75)
	46-55	495 (32.94)
	>55	518 (34.46)
Height (centimeter)	92-195	168.53 (8.5)
Weight (kilogram)	6.5-163	75.20 (13.0)
Marital status	single	623 (41.45)
	married	752 (50.03)
	widow	128 (8.52)
Literacy	under diploma	487 (32.40)
	diploma	448 (29.81)
	associate degree	312 (20.75)
	bachelor's degree	152 (10.12)
	master's degree	78(5.19)
	PhD and above	26 (1.73)
Disease category	BC cases	554 (36.85)
	non-BC cases	949 (63.15)

Table 2) The most important BC prediction factors at p<0.01

Variable	Percent	Mean±SD	χ^2
Personal history of breast cancer			148.49
Yes	741	-	
No	762	-	
History of breast sampling			693.33
Yes	550	-	
No	953	-	
Family history of breast cancer			917.852
Yes	394	-	
No	1109	-	
History of colorectal cancer			1117.804
Yes	455	-	
No	1048	-	
Hypertension			1698.279
Yes	151	-	
No	1352	-	
Fruit consumption gr			1518.126
<100	193	-	
100-200	388	-	
>200	921	-	
Vegetable consumption gr			1519.569
<150	149	-	
150-300	422	-	
>300	939	-	
Alcohol consumption			2457.507
Yes	405	-	
No	1097	-	
Physical activities hours per day			1595.936
0-0.5	179	-	
0.5-1	831	-	
> 1	493	-	
Hypercholesterolemia			1568.890
Yes	64	-	
No	1439	-	
Fatness			1538.347
Yes	31	-	
No	1472	-	
Hyperlipidemia			1564.487
Yes	46	-	
No	1457	-	
Diabetes			1641.256
Yes	79	-	
No	1424	-	
Upper in Quadrants breast cancer			1519.623
Yes	12	-	
No	1491	-	
History of chest radio therapy			1539.431
Yes	91		
No	1412		
Age	-	38.490±11.114	950.06
Body Mass Index	-	24.022±12.912	742.465
The ratio of waist to pelvic	-	66.674±40.303	253.509

The results of determining the combinational correlation between the BC diagnostic factors and the dependent variable using binary logistic regression (BLR) and forward logistic regression method have been brought in IF-Term Removed Table (Table 3). As depicted in Table 3, in the 9th step of the BLR, by entering the 9 variables of history of breast sampling, history of chest radiotherapy, family history of BC, alcohol consumption, vegetable consumption, diabetes, physical activity, age, and upper in quadrants breast cancer, the average of log-likelihood of the model has been obtained -61.91 at p<0.01. In conclusion, by selecting these nine variables in the BLR model and reducing the Log-likelihood, the performance of the BLR has been increased and therefore, these variables had a significant hybrid correlation coefficient with output class at p<0.01.

The results of comparing the performance of selected RBML algorithms in BC classification using the confusion matrix showed that the DT was the only algorithm that by FP=0 and TN=949, has classified all the non-BC samples correctly, and was a better algorithm than others in this regard. The J-48 decision tree algorithm with FP=1 and TN=948 had also the pleasant capability of classifying the non-BC cases. Also, this algorithm with FN=12 and TP=542 had a better performance in classifying the positive cases than other algorithms.

The results of measuring the evaluation criteria of PPV, NPV, sensitivity, specificity, accuracy, and F-score of these algorithms have been demonstrated in Diagram 2. Based on the results, although, the DT rule-based algorithm with NPV=1 demonstrated the best capability in just classifying the negative BC cases, generally, the J-48 decision tree algorithm with accuracy=0.991 and F-measure=0.987 has obtained the best performance in classifying all research samples than other algorithms. The ROC of all RBML algorithms has been shown in Diagram 3.

Generally, investigating all the algorithms classification performance using different evaluation criteria showed that the J-48 decision tree algorithm with PPV of 0.998, NPV of 0.987, the sensitivity of 0.978, specificity of 0.998, accuracy of 0.991, F-measure of 0.987, and also AUC of 0.9997 yielded the best performance than other algorithms for predicting the BC risk. In Diagram 4, the J-48 decision tree algorithm has been depicted and all technical characteristics used in this study have been mentioned. Finally, the best knowledge about diagnosing BC with the more classified sample extracted from this algorithm with IF-THEN structures has been brought and then interpreted. The most important technical features utilized for building J-48 with the best performance include the number of batch size=100, binary split=False, collapse tree=True, confidence factor=0.25, number of minimal objects=2, number of decimal places=2,

number of folds=3, reduced error pruning=True, and number of seeds=1.

Some knowledge extracted from the J-48 decision tree algorithm with highly classified samples:

1- IF Radio therapy=Yes THEN Diagnosis=breast cancer;

2- IF Radio therapy=No & Alcohol=Yes THEN Diagnosis=breast cancer;

3- IF Radio therapy=No & Alcohol=No & Age <=38 THEN Diagnosis=Non-breast cancer.

Based on the J-48 decision tree algorithm's diagram, the history of chest radiotherapy has been considered as the most important factor for diagnosing BC. Generally, three rules have been obtained as the most important patterns, as below:

1- The first rule was only based on the history of the chest radiotherapy as a condition, this means that in 455 of the positive cases, the history of chest radiotherapy has been seen and if one person has this risk factor, the probability of afflicting BC can be 82.1%;

2- In the second rule, if the person without any history of chest radiotherapy with alcohol consumption, the probability of afflicting BC can be 11.3% (63 positive samples have been classified truly);

3-The third rule is very important for diagnosing the non-BC cases, and if a person without any history of chest radiotherapy, non-alcoholic and less than 38 years, the probability of non-afflicting BC can be 89.5% (850 truly classified samples/ 949 total negative samples).

Table 3) IF-Term removed table for BC diagnostic factors (p<0.001)

Variable		Model Log-Likelihood	Change in -2 Log-Likelihood	df
Step 1	History of chest radiotherapy	-987.283	1319.042	1
Step 2	History of chest radiotherapy	-431.990	555.056	1
	Alcohol consumption	-327.762	346.599	2
Step 3	History of chest radiotherapy	-364.370	503.914	1
	Alcohol consumption	-261.954	299.081	2
	Upper in Quadrants breast cancer	-154.462	84.099	1
Step 4	History of chest radiotherapy	-207.348	255.777	1
	Alcohol consumption	-146.989	135.059	2
	Age	-112.413	65.907	1
	Upper in Quadrants breast cancer	-107.476	56.034	1
Step 5	History of chest radiotherapy	-173.157	245.248	1
	Alcohol consumption	-118.811	136.556	2
	Vegetable consumption	-79.459	57.853	2
	Age	-85.523	69.981	1
	Upper in Quadrants breast cancer	-70.753	40.441	1
Step 6	History of chest radiotherapy	-172.819	251.310	1
	Alcohol consumption	-112.812	131.296	2
	Vegetable consumption	-79.390	64.451	2
	Diabetes	-50.533	6.737	1
	Age	-82.903	71.476	1
	Upper in Quadrants breast cancer	-66.211	38.093	1
Step 7	History of breast sampling	-47.164	7.592	1
	History of chest radiotherapy	-103.301	119.865	1
	Alcohol consumption	-102.845	118.952	2
	Vegetable consumption	-73.136	59.536	2
	Diabetes	-47.101	7.464	1
	Age	-81.153	75.568	1
	Upper in Quadrants breast cancer	-64.240	41.742	1
Step 8	History of breast sampling	-42.427	8.412	1
	History of chest radiotherapy	-43.616	10.789	1
	Family history of breast cancer	-43.369	10.294	1
	Alcohol consumption	-96.486	116.528	2
	Vegetable consumption	-65.658	54.873	2
	Diabetes	-42.044	7.646	1
	Age	-72.760	69.077	1
	Upper in Quadrants breast cancer	-59.439	42.434	1
Step 9	History of breast sampling	-36.666	8.215	1
	History of chest radiotherapy	-37.083	9.049	1
	Family history of breast cancer	-38.143	11.170	1
	Alcohol consumption	-85.233	105.349	2
	Vegetable consumption	-54.881	44.646	2
	Diabetes	-37.015	8.914	1
	Physical activity	-38.221	11.326	2
	Age	-69.489	73.861	1
	Upper in Quadrants breast cancer	-54.975	44.834	1

Table 4) The selected RBML confusion matrix

Algorithms	Technical features	TN	FP	FN	TP
DT	Number of batch size=100 Number of decimal places=2 The search method=best first	949	0	63	491
J-RIP	Number of decimal places=2 Optimization=2 Number of seeds=1	836	113	84	470
Part	Number of decimal places=2 Fold numbers=3 Confidence factor=0.25 Number of seeds=1	825	124	104	450
RF	Number of iterations=1 Number of execution slots=1 Break Ties randomly=false Number of decimal places=2	901	48	109	445
J-48	Number of decimal places=2 Minimum number of objects=2 Confidence factor=0.25 Number of seeds=1 Number of folds=3	948	1	12	542
RT	Number of decimal places=2 Number of seeds=1 Break Ties randomly=false Minimum variance properties=0.001	785	164	73	481
REP-Tree	Number of folds=3 Number of decimal places=2 Maximum depth=-1 Number of seeds=1	801	148	52	502

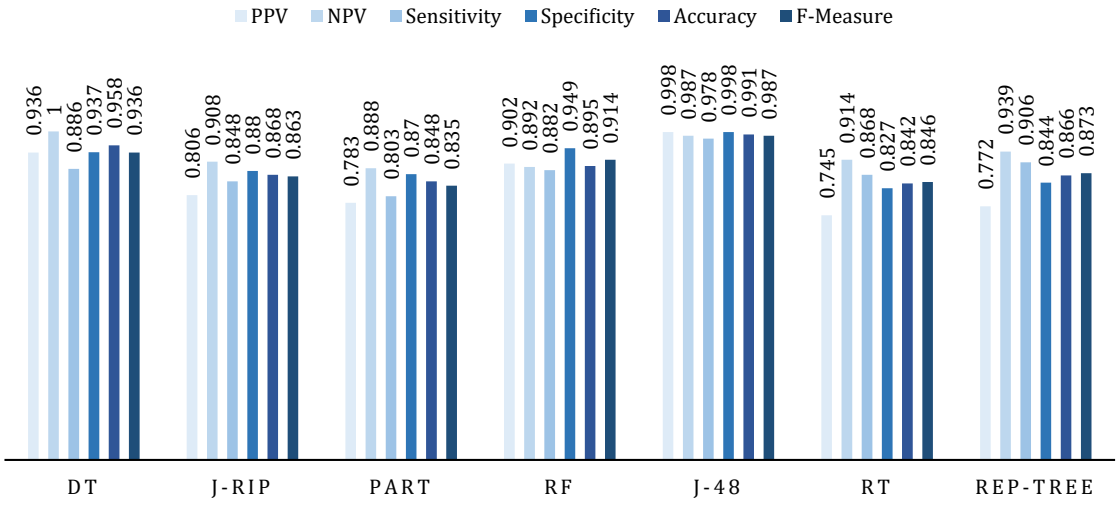

Diagram 2) Various performance evaluation criteria of different RBML algorithms (The vertical and horizontal vertices of the diagram show the True Positive Rate (TPR) and False Positive Rate (FPR), respectively)

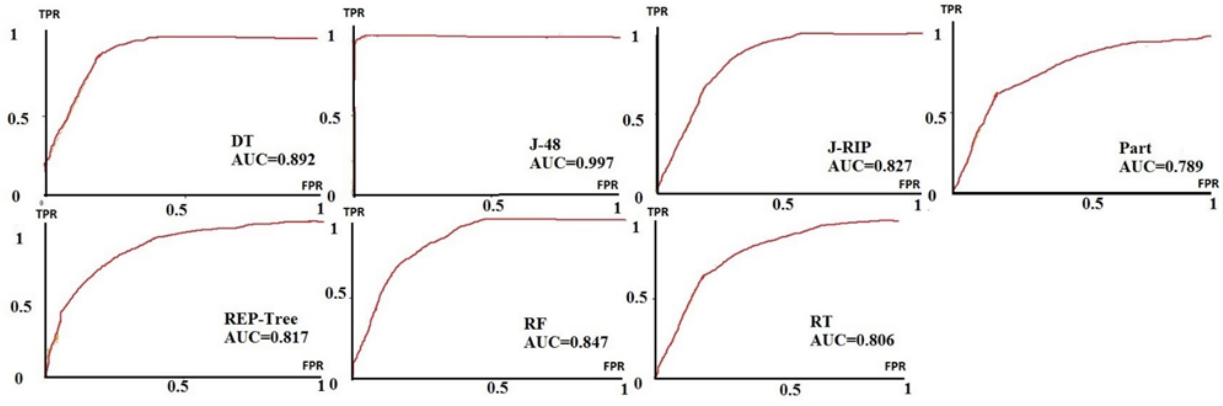

Diagram 3) The ROC of different RBML algorithms

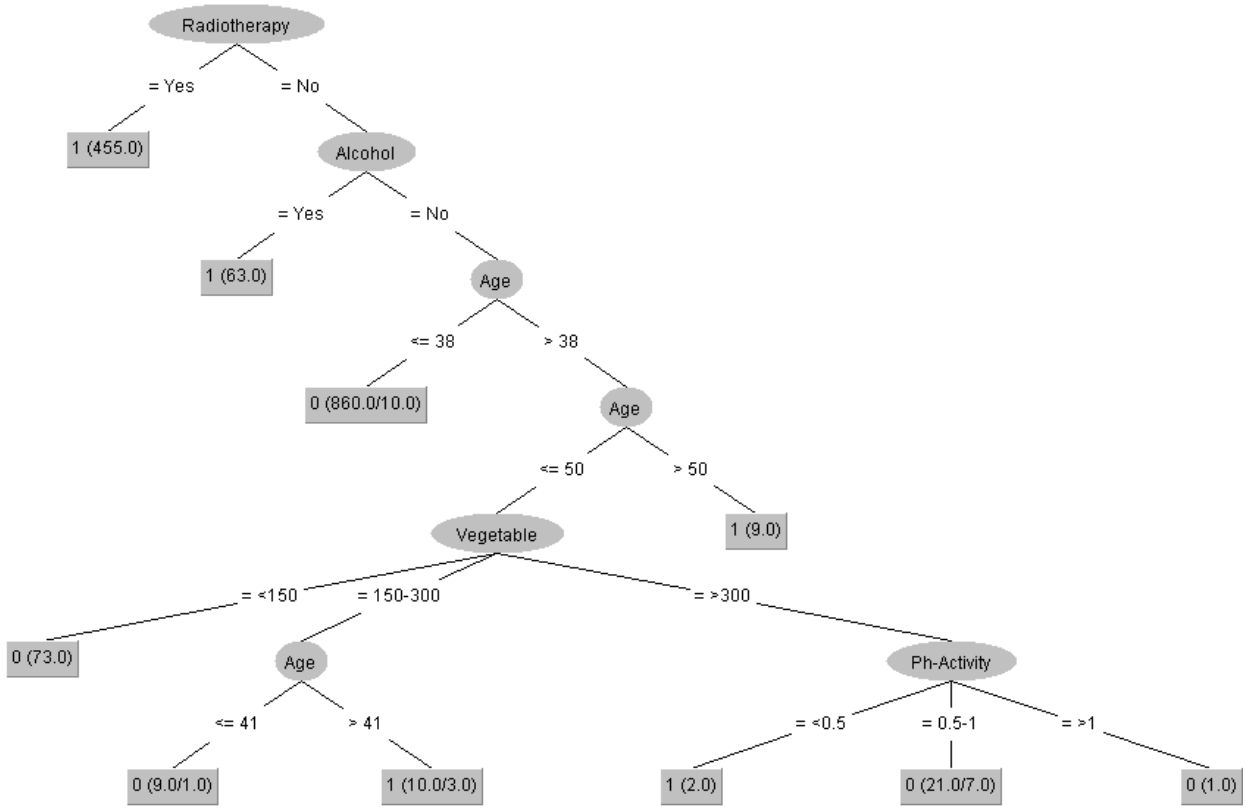

Diagram 4. The J-48 pruned decision tree

Discussion

The purpose of the current study was to effectively determine BC cases through intelligent RBML techniques. In the present study, multiple RBML-based predictive models were developed for early risk prediction of BC based on 1668 suspected BC clinical data. Thus, we trained seven RBML algorithms including J-48, RF, RT, and REP-Tree, DT, J-RIP, and Part according to the top related parameters affecting the risk of BC that derived from a correlation coefficient analysis. The selected algorithms were applied to the pre-processed dataset. This study first selected the most reliable and clinically relevant predictors related to BC by using the independence Chi-square test. Hence we identified nine highly correlated variables that had the meaningful hybrid correlation coefficient with output class at $P<0.05$. It is proven that ML can be an effective tool in dealing with BC problems [30]. Then to validate the system, the k-fold cross-validation method was used. To compare the performance of selected RBML classifiers, several evaluation metrics derived from confusion matrices such as PPV, NPV, sensitivity, specificity, accuracy; F-score, and AUC-ROC were used.

So far, several studies have been evaluating the application of ML algorithms in BC risk classification and prediction based on clinical variables [31]. The Momenyan results showed J-48 gained optimum predictive performance with an accuracy of 93.3% [26]. Regarding the obtained results by Alickovic *et al.* portrayed J-48 DT method was able to predict the probability of BC more accurately compared with other classifiers [32]. The best meaningful results in Dawngliani's study were obtained from the J-48 model with an accuracy of 84.21% while a random tree demonstrates the lowest accuracy (76.49%) [33]. Saabith also stated that the J-48 was the best ML technique to predict BC with an accuracy of 79.97% [34]. Solanki *et al.* in their study revealed that J-48 with an accuracy of 98.83% and AUC of 0.983 gained the best performance for BC diagnosis and differentiating the benign and malignant patients [29]. Similarly, Al-Salihy showed J-48 DT algorithm is outperformed by other algorithms with an accuracy of 97.7% [35]. Ortega's study presented that the application of the J-48 algorithm in BC risk assessment had optimum accuracy (95%) in risk classification and disease screening [36]. In Silva's and Mohammed's work authors concluded that J-48 yielded better performance than others (accuracy of 91 and 98.2%, respectively) [27, 28]. The results of Solanki's research showed that the model developed by J-48 yielded the best performance in terms of classification accuracy [29]. Ultimately Salod's results showed that the model developed using J-48 with 0.81 of AUC was introduced as the best performing model [2].

Hence the purpose of these researches is to propose the most effective ML-based predictive models for early BC prognosis by classification of BC risks. In our study, we applied two feature selection methods

including the independence Chi-square test and BLR as a hybrid correlation method for determining the most important factors affecting BC. The independence Chi-square test showed that the 18 diagnostic variables acquired the Chi-square at $p<0.01$ and therefore, were considered as the most important factors determining BC. Also, the results of using the BLR showed that the nine variables including the history of breast sampling, history of chest radiotherapy, family history of breast cancer, alcohol consumption, vegetable consumption, diabetes, physical activity, age, upper in quadrants breast cancer at nine steps of the BLR had a common hybrid correlation with BC diagnosis at $p<0.05$, and therefore, were used for making the decision trees and aftermath knowledge representation. The experimental results of the present work similar to the reviewed studies showed that the J-48 decision tree with PPV=0.998, NPV=0.987, sensitivity=0.978, specificity=0.998, accuracy=0.991, F-measure=0.987, and also AUC=0.9997 has the best capability for earlier detection of the disease, improve diagnostic efficiency and decrease the risk of mortalities caused by BC.

The results of the present study may help physicians throughout correct, accurate, and timely diagnosis of the disease and reduce the severe complications of the disease and the resulting mortality. Despite the small amount of data fed into the models and the lack of clinical variables, the selected RBML models, especially the J-48 algorithm, performed well. On the other hand, this model application in real clinical environments will assist physicians owing to its simplicity, user-friendliness, and easy-to-use characteristics.

Given the power of the current study in the timely and accurate prediction of BC risk, this study had some limitations that need to be addressed. First, this is a retrospective study that suffers from low data quantity (missing or duplicate cells) and non-optimal quality (imbalanced, noisy, and meaningless values). Second, we deal with a single-center dataset with a limited sample size which undoubtedly confines the generalizability of the proposed model. Moreover, we used only seven RBML algorithms for prediction analyses based on some clinical features. Finally, the selected registry dataset lacks some important variables such as Para-clinical indicators. In the future, the performance accuracy of our model and its generalizability will be enhanced if we test more ML techniques, at the larger, multicenter, and prospective dataset which is equipped with more qualitative and validated data. The obtained results confirm the positive effect of nine selected features in predicting the risk of BC as a powerful optimizer which selected the best sub-set features to be included in the RBML algorithms. It has been inferred that by different ML algorithms, the prediction models have shown more promising performance compared to other traditional approaches. Hence, ML algorithms can construct complex models and make reliable decisions when fed by appropriate features.

Conclusion

The evaluation of the selected ML technique's performance demonstrates the suitability of the J-48 for predicting BC risk. Our proposed predictive model for BC discriminates persons at high and elevated risk for BC and non-BC cases based on the most important variables and can be used as an essential and non-invasive clinical screening tool for the early identification of BC.

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

Ethical Permissions: The Abadan University of Medical Science ethics board approved the study design (Ethics code: IR.ABADANUMS.REC.1400.040).

Conflicts of Interests: This article is extracted from a research project supported by the Abadan University of Medical Sciences.

Authors' Contributions: Nopour R. (First author), Methodologist/Statistical Analyst (40%); Kazemi Arpanahi H (Second author), Introduction Writer/Discussion Writer (20%); Shanbehzadeh M. (Third author), Introduction Writer/Methodologist/Assistant researcher (40%).

Funding/Sources: The Abadan University of Medical Sciences was support this project.

References

1- Dhahri H, Al Maghayreh E, Mahmood A, Elkilani W, Faisal Nagi M. Automated breast cancer diagnosis based on machine learning algorithms. J Healthcare Eng. 2019;2019:4253641.

2- 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(3):1677.

3- Homan SG, Yun S, Bouras A, Schmaltz C, Gwanfogbe P, Lucht J. Breast cancer population screening program results in early detection and reduced treatment and health care costs for medicaid. J Public Health Manag Pract. 2021;27(1):70-9.

4- Anwar SL, Dwianingsih EK, Avanti WS, Choridah L, Suwardjo, Aryandono T. Aggressive behavior of Her-2 positive colloid breast carcinoma:A case report in a metastatic breast cancer. Annal Med Surg. 2020;52:48-52.

5- Babiera GV. Metastatic breast cancer. A paradigm shift toward a more aggressive approach. Cancer J. 2009;15(1):78.

6- 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):1-7.

7- 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 Global Oncol. 2019, 5:1-10.

8- 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-e9.
9- 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.
10- Domeyer PRJ, Sergentanis TN. New insights into the screening, prompt diagnosis, management, and prognosis of breast cancer. J Oncol. 2020;2020:8597892.
11- Badiger S, Moger J. A comparative study of mammography, sonography and infrared thermography in detection of cancer in breast. Int Surg J.2020;7(6):1886.
12- Mohamed NC, 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 Prevent. 2019;20(9):2865-73.
13- Anderson BO, Bevers TB, Carlson RW. Clinical breast examination and breast cancer screening guideline. JAMA. 2016;315(13):1403-4.
14- Dowsett M, Sestak I, Regan MM, Dodson A, Viale G, Thürlimann 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 Clini Oncol. 2018;36(19):1941-8.
15- 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.
16- Tran WT, Sadeghi-Naini A, Lu FI, Gandhi S, Meti N, Brackstone M, et al. Computational radiology in breast cancer screening and diagnosis using artificial intelligence. Can Assoc Radiol J. 2021;72(1):98-108.
17- Elwood JM. Drug and hormone resistance in neoplasia. Boca Raton: CRC Press; 2019. pp. 39-56.
18- Sathya D, Sudha V, Jagadeesan D. Handbook of research on applications and implementations of machine learning techniques. Pennsylvania: IGI Global; 2020. p. 289-304.
19- Bradley A, Van Der Meer R, McKay C. Personalized pancreatic cancer management: a systematic review of how machine learning is supporting decision-making. Pancreas. 2019;48(5):598-604.
20- Prasuna K, Rama RK, Saibaba C. Application of machine learning techniques in predicting breast cancer – A survey. Int J Innov Technol Exploring Eng. 2019;8(8):826-32.
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):1-17.
22- Peng J, Zeng X, Townsend J, Liu G, Huang Y, Lin S. A machine learning approach to uncovering hidden utilization patterns of early childhood dental care among medicaid-insured children. Front Public Health. 2021;8:1025.
23- Singh NK. Prediction of breast cancer using rule based classification. Appl Med Inform. 2015;37(4):11-22.
24- Tian JX, Zhang J. Breast cancer diagnosis using feature extraction and boosted C5.0 decision tree algorithm with penalty factor. Math Biosci Eng. 2022;19(3):2193-205.
25- Idris NF, Ismail MA. Breast cancer disease classification using fuzzy-ID3 algorithm with FUZZYDBD method: automatic fuzzy database definition. Peer J Comput Sci. 2021;7:e427.
26- Momenyan S, Baghestani AR, Momenyan N, Naseri P, Akbari ME. Survival prediction of patients with breast cancer:comparisons of decision tree and logistic regression analysis. Int J Cancer Manag. 2018;11(7).
27- 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.
28- Mohammed A, Arunachalam N. Imbalanced machine learning based techniques for breast cancer detection. 2021 International Conference on System, Computation, Automation and Networking (ICSCAN), 30-31 July 2021, Puducherry, India. Piscataway: IEEE; 2021.
29- Solanki Y, 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):699.
30- Chaurasia V, Pal S. Applications of machine learning techniques to predict diagnostic breast cancer. SN Comput Sci. 2020;1(5):1-11.
31- Mohammed SA, Darrab S, Noaman SA, Saake G. Analysis of breast cancer detection using different machine learning techniques. International Conference on Data Mining and Big Data. 2020;1234:108-17.
32- Alickovic E, Subasi A. Comparison of decision tree methods for breast cancer diagnosis. The 6th International Conference on Information Technology (ICIT 2013), Amman, Jordan. Unknown Publisher; 2013.
33- Dawngliani M, 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; 2019:232-44.
34- Saabith ALS, Sundararajan E, Bakar AA. Comparative study on different classification techniques for breast cancer dataset. Int J Comput Sc Mob Comput. 2014;3(10):185-91.
35- Al-Salihy NK, Ibrikci T. Classifying breast cancer by using decision tree algorithms. Proceedings of the 6th International Conference on Software and Computer Applications, Unknown Date & location. Unknown Publisher; 2017.
36- Ortega JHJC, 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 Scie Eng. 2020;9(3):178-85.