Clinical Epidemiology and Global Health 7 (2019) 293–299

Contents lists available at ScienceDirect

Clinical Epidemiology and Global Health

journal homepage: www.elsevier.com/locate/cegh

Prediction of survival and metastasis in breast cancer patients using machine learning classifiers

Leili Tapak[a], Nasrin Shirmohammadi-Khorram[b,*], Payam Amini[c], Behnaz Alafchi[d], Omid Hamidi[e], Jalal Poorolajal[f]

[a] Department of Biostatistics, School of Public Health and Modeling of Noncommunicable Diseases Research Center, Hamadan University of Medical Sciences, Hamadan, 65175-4171, Iran
[b] Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, 65175-4171, Iran
[c] Department of Epidemiology and Reproductive Health, Reproductive Epidemiology Research Center, Royan Institute for Reproductive Biomedicine, ACECR, Tehran, Iran
[d] Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran
[e] Department of Science, Hamedan University of Technology, Hamedan, 65155, Iran
[f] Department of Epidemiology, School of Public Health and Modeling of Noncommunicable Diseases Research Center, Hamadan University of Medical Sciences, Hamadan, 65175-4171, Iran

ARTICLE INFO

Keywords:
Breast cancer survival prediction
Metastasis prediction
Classification
Machine learning models

ABSTRACT

Background: Breast cancer (BC) is one of the most common malignancies in women. Early diagnosis of BC and metastasis among the patients based on an accurate system can increase survival of the patients to > 86%. This study aimed to compare the performance of six machine learning techniques two traditional methods for the prediction of BC survival and metastasis.

Methods: We used a dataset that include the records of 550 breast cancer patients. Naive Bayes (NB), Random Forest (RF), AdaBoost, Support Vector Machine (SVM), Least-square SVM (LSSVM) and Adabag, Logistic Regression (LR) and Linear Discriminant Analysis were used for the prediction of breast cancer survival and metastasis. The performance of the used techniques was evaluated with sensitivity, specificity, likelihood ratio and total accuracy.

Results: Out of 550 patients, 83.4% were alive and 85% did not experience metastasis. In prediction of survival, the average specificity of all techniques was ≥94% and the SVM and LDA have greater sensitivity (73%) in comparison to other techniques. The greater total accuracy (93%) belonged to the SVM and LDA. For metastasis prediction, the RF had the highest specificity (98%), the NB had highest sensitivity (36%) and the LR and LDA had the highest total accuracy (86%).

Conclusions: Our finding showed that the SVM outperformed other machine learning methods in prediction of survival of the patients in terms of several criteria. Nevertheless, the LDA technique as a classical method showed similar performance.

1. Introduction

Breast cancer (BC) is the most common cancer among women that is the first leading cause of cancer-related deaths among women[1,2] and the second leading cause of cancer deaths, worldwide.[1] The main cause of BC is the uncontrolled growth of cells in breast tissues which can be either benign or malignant.[3] Malignant tumors are cancerous and their cells can spread to other parts of the body that causes metastasis, while benign type is non-intensive.[4] Metastasis, a sign of disease progression is related to survival of BC patients.[5] Less than 10% of breast cancers are thought to be hereditary, and others are because of genetic abnormalities.[6] It has been also shown that the age of about 1.9% of all BC patients are under 35 years old.[7]

Early detection of cancer is critical to improve breast cancer survival and to reduce the high mortality rate of BC.[1].Despite early detection and the advent of new treatments, about 50% of patients will develop distant metastases during their follow-up time.[8] Therefore, a precise and reliable system is required for the early diagnosis of tumors.[9]

Many modern medical diagnosing tools are based on the information achieved by clinical observation or some available tests which can

* Corresponding author. Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, 65175-4171, Iran.
E-mail addresses: l.tapak@umsha.ac.ir (L. Tapak), n.shirmohamadi.kh@gmail.com (N. Shirmohammadi-Khorram), payam.amini87@gmail.com (P. Amini), behnaz.alafchi@gmail.com (B. Alafchi), omid_hamidi@hut.ac.ir (O. Hamidi), poorolajal@umsha.ac.ir (J. Poorolajal).

https://doi.org/10.1016/j.cegh.2018.10.003
Received 2 July 2018; Received in revised form 16 September 2018; Accepted 5 October 2018
Available online 06 October 2018

help physicians to diagnose BC.[9] Most of these methods are based on classification and many researchers have adopted them to improve their precision. Methods which have better classification precision would improve diagnostic accuracy. Machine learning, which is a process to design a model which learns through experience to improves its performance, belongs to the artificial intelligence framework and are increasingly used in different fields of science.[10] Although the main objective of these models is to identify effective variables and their relationships, they can be used in prediction problems as well.[11,12]

Machine learning techniques are widely used in medicine for diagnosing BC.[10].There have been introduced several machine learning methods in various studies.[13–15] Montazeriet al have compared performances of Naive Bayes, Trees Random Forest, 1-Nearest Neighbor, AdaBoost, Support Vector Machine, RBF Network, and Multilayer Perceptron machine learning techniques.[16] Chao et al. have used support vector machine, logistic regression, and a C5.0 decision tree model to predict BC survival.[17] The main advantage of these techniques is that they overcome issues like collinearity, heteroskadacity, complex interactions between variables and higher order interactions between predictors that classical techniques are encountered with.[16]

While there are several studies showing the lower error rates and the higher accuracy in classification problems for data mining techniques compared with the traditional methods (LDA and LR), there can be found studies that shows this excellence is not the case for all data sets. There is inconsistency across the results of various studies regarding the classification accuracy of data mining techniques compared with the traditional methods that are less computer-demanding.[18–20]

The results of different studies have also introduced different methods as the most reliable one for prediction of survival of BC patients.[16,17] In addition to survival, metastasis as an important sign of disease progression is a consequential outcome in cancer studies and its effective variables is of interest.[21].The present study aimed to compare the performance of six machine learning techniques(including NB, LSSVM, Adaboost, Adabag, SVM, RF) and two traditional methods (LR and LDA) in prediction of survival and metastasis outcomes in patients with breast cancer.

2. Materials and methods

2.1. Data collection

We used a data set originates from a retrospective cohort study that was conducted in 2014 in Tehran. We used the information of patients who developed breast cancer (International Classification of Diseases for Oncology 3rd edition sites C50.0-C50.9) and registered with the Comprehensive Cancer Control Center associated with Shahid Beheshti University of Medical Sciences from 1998 to 2013. All patients diagnosed pathologically and the patients with unknown pathology were excluded from analysis. Our focus was on the information about survival status (dead/alive) and metastasis (yes/no) as outputs and their risk factors among Iranian women. We selected 9 risk factors that are believed associated with survival of breast cancer patients, including age, Grade (well, moderate and poor), stage, Estrogen receptor (ER as negative or positive), Progesterone receptor (PR as negative or positive), Human epidermal growth factor receptor 2 (HER2 as negative or positive), Pathological type (Ductal/lobular carcinoma in situ, Invasive lobular carcinoma and Invasive ductal carcinoma), and Surgical approach (Modified Radical Mastectomy and Breast-conserving surgery) to compare the performance of the selected models.

2.1.1. Naïve Bayesian (NB) classification

The Naïve Bayes classification model works based on the famous Bayes' theorem following a clear, simple, and very fast classifier.[22,23] Using the Bayes rule, the prior probability of belonging to each class can be learnt and estimated using the training data with ignoring the marginal probabilities based on the conditional probability of each variable X_j (age, Grade, stage, ER, PR, HER2, Pathological type and Surgical approach) given the class label C (for survival c is one of dead or alive status and for metastasis c is yes or no). Then classification is conducted through the Bayes rule to calculate the probability of C given $\underline{X}_1, \ldots, \underline{X}_n$, by the formula:

$$P(C = c|\underline{X}_1 = \underline{x}_1, \ldots, \underline{X}_n = \underline{x}_n)$$

So, the posterior probability of each of the classes is calculated as follows:

$$P(C = c|\underline{X}_1 = \underline{x}_1, \ldots, \underline{X}_n = \underline{x}_n) = P(C = c) \times \prod_{x_j} P(X_j = x_j|C = c)$$

The new subject belongs to the class with the highest posterior probability.[16]

2.1.2. Logistic regression (LR)

This approach assumes that the binary outcome follows a binomial distribution. The model can be written as:

$$log\left(\frac{\pi}{1-\pi}\right) = \sum_{i=1}^{k} \beta_i X_i$$

In this model X's are the covariates and β_is are the regression coefficients indicating the measure of effect size.[24]

2.1.3. Linear discriminant analysis (LDA)

LDA is similar to LR and relates the dependent variable to a linear combination of predictors which best explain and classify the outcome. LDA solves the problem using conditional probability of the predictors given the class of the outcome. This approach minimizes the dispersion between the cases of the same class, and maximizes the dispersion between the cases of different classes.[25]

2.1.4. Random forest (RF)

This approach was first introduced by Leo Breiman.[26] RF assembles classification and regression trees. The dataset is sampled by replacement to form the trees in RF. Random sets of predictors are selected at the nodes which are created by the trees. It is possible to find the most important predictors using mean decrease Gini and mean decrease accuracy. The important variables classify the binary outcome so that the prediction is carried out with the highest accuracy.[26]

2.1.5. Support vector machine (SVM)

SVM is a machine learning technique that has been widely used in regression and classification problems. In this method, the classification equation for two groups (for survival are dead and alive status and for metastasis are yes and no) based on feature space (age, Grade, stage, ER, PR, HER2, Pathological type and Surgical approach) is given as follows:

$$y_i = \sum_{i=1}^{N} \alpha_i \gamma_i(x) + c = 0$$

where $\{\gamma_i(x)\}_{i=1}^{N}$ are features, c and $\{\alpha_i\}_{i=1}^{N}$ denote bias and value estimated weights from the data and $\{y(i)\}_{i=1}^{N}$ represent a set of response samples (survival/metastasis) where $y(i) \in \{-1, +1\}$. The optimal weights and the bias value are evaluate by solving the quadratic optimization problem as:

$$\text{Minimize } K(\alpha,\ c,\ \varepsilon) = \frac{1}{2}||\boldsymbol{\alpha}||^2 + B\left(\sum_{i=1}^{N} \varepsilon_i\right)^k$$

with the inequality constraint as $y_i(\boldsymbol{\alpha}^T\varphi(x_i) + c) \geq 1 - \varepsilon_i$, $(\varepsilon_i \geq 0)$. The function φ maps the training data to a higher dimensional space. B is a positive real constant that controls the trade-off between misclassifications and the complexity of the model. Therefore, the classifier function takes the form $f(x) = sign(\sum_{i=1}^{N} \beta_i y_i K(x, x_i) + c)$ where x is

any testing vector and the term $K(x, x_i)$ is denoted as the kernel function.[27] Choosing the kernel function and the parameters in the SVM make it a flexible method.[28] In the present study we used the radial basis kernel function because of its superior performance.

2.1.6. Least square support vector machine (LS-SVM)

LS-SVM is a modification to the Support Vector Machine (SVM) model with a least squares loss function and equality constraints, where the dual solution can be found by solving a linear system instead of quadratic programming problem. As for SVM, LS-SVM maps the data into a high dimensional feature space. The primal formulation of the LSSVM classification model is Minimize $f(\alpha,\ c,\ \varepsilon) = \frac{1}{2}||\boldsymbol{\alpha}||^2 + \frac{1}{2}B(||\boldsymbol{\varepsilon}||^2)$ with the equality constraint as $y_i(\boldsymbol{\alpha}^T\varphi(x_i) + c) = 1 - \varepsilon_i$, $(\varepsilon_i \geq 0)$.[29]

2.1.7. AdaBoost (AD)

AdaBoost belongs to the machine learning techniques family and can be considered as a meta-algorithm that improves the performance together with other learning techniques. In a classification problem, AdaBoost focuses on the sequentially applying weak classifiers. In this way, the algorithm is repeatedly applied on the modified data. For example, let $Y \in \{-1, +1\}$ be the output variable with the -1 for death and $+1$ for alive statuses. Moreover, let X be a vector of potential risk factors (here age, grade, etc.). So, any classifier, say $G(X)$, predicts the status of the patients in $\{-1, +1\}$ set and the error rate on the training set and the expected error rate on the test set are calculated as follows:

$$e = \frac{1}{N}\sum_{i=1}^{N} I(y_i \neq G(x_i))$$

$$e_{test} = E_{XY}(I(Y \neq G(X))$$

AdaBoost generates M week classifiers $(G_m(X),\ m = 1,\ ...,M)$ and then produces the final prediction by combing them using a weighting process by the following rule:

$$G(x) = sign\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right)$$

Where α_m are computed using the boosting algorithm.[16]

2.1.8. Adabag

Bagging is a machine learning technique that works based on combining bootstrapping and aggregating. In this method, the number of B bootstrap samples is selected fromthe training set, say Tb $(b = 1, 2, \ldots, B)$. By bootstrapping, the noisy observations are reduced and even eliminated from some of Tbs. Therefore, these sets will provide the classifiers with a better behavior compared with the original set. This makes bagging technique a useful tool to build a better classifier at the presence of noisy observations in the training set. Finally, better results can be achieved by the ensemble of these B classifiers compared with the single classifiers. For the BC data set, the algorithm for bagging is as follows:

Step1. For each $b = 1, 2, \ldots, B$, a single classifier $C_b(age, grade, stage, ...) = \{alive, dead\}$ is constructed based on a bootstrap sample obtained from the original data set.

Step2. These basic classifiers are combined by using the most often predicted class (alive/dead or having metastasis/not having) to create the final decision rule.

2.1.9. Evaluation criteria and cross validation

In order to compare the discriminative power of the used classifiers, several criteria of sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (LR +), negative likelihood ratio (LR–) and total accuracy were provided that are calculated as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},\quad \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}},\quad \text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

$$\text{LR} + = \frac{\text{Sensitivity}}{1 - \text{Specificity}},\ \text{LR} - = \frac{1 - \text{Sensitivity}}{\text{Specificity}}$$

$$\text{Total Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP+TF} + \text{TN+FN}}$$

Where FP stands for alive people with breast cancer that were incorrectly identified as dead, TP stands for dead people with breast cancer that were correctly diagnosed as dead, TN stands for alive people with breast cancer that were correctly identified as dead and FN stands for dead people with breast cancer incorrectly identified as alive. We also used likelihood ratio criteria to compare the methods.

Positive likelihood ratio is the ratio of the sensitivity to 1 minus specificity and takes values greater than zero. A zero value (the worst case) is related to a test with zero sensitivity. Larger values of the positive likelihood ratio criterion indicate more valuable information is included in a test. Negative likelihood ratio is calculated as the ratio of 1 minus sensitivity to specificity and takes values greater than zero. The smaller the negative likelihood ratio, the greater the information can be provided by a test.[30]

We divided the data into two sets of training and testing set. So we considered two different scenarios: a) a 70% and 30% and b) a 50% and 50% for the train and test sets respectively. We also repeated this process 100 times for each scenario and reported the evaluation criteria as average over 100 repetitions.

2.1.10. Software

In the present study all analysis was implemented in R version 3.4.4 using packages e1071, naivebayes, kernlab, adabag.

3. Results

3.1. Data description

The data set included 550 patient records in which 463 (83.4%) patients were alive and 92 (16.6%) patients were dead. In addition, about 85% of the patients did not experience metastasis. We used these two variables as the target variable each included two categories (alive or dead categories for survival and experience or not metastasis).

Characteristics of the patients with BC were shown in Table 1. The patients with BC at diagnosis aged 47.86 ± 11.79 (mean ± SD) year in average with a minimum and maximum of 17 and 84 years respectively. Most of the patients were presented with grade II (~52%) and were at stage II (~42%), ER+ (~71%), PR+ (~68%), HER2- (~76%), diagnosed with pathological type of invasive ductal carcinoma (~90%) and received breast-conserving surgery (~65%) (Table 1). The distributions of the characteristics of the patients divided randomly into two sets (train and test) were also provided (Table 1). As seen over 100 repetition, there was no significant differences between train and test sets ($P > 0.05$).

3.1.1. Performance of the models in predicting survival

Table 2 shows the performance of the eight classifiers for prediction of survival of the patients with BC in terms of sensitivity, specificity, PPV, NPV, LR+, LR- and total accuracy obtained from 100 times of repetition of cross validation strategy under two scenarios. As can be seen in Table 2, all of the used algorithms had high specificity (≥94%) for the test sets. Nevertheless, the values of the sensitivity for the models were moderate and varied between 0.61 and 0.73 on averages for the test sets with the lowest and highest belonged to the AdaBoost and the SVM respectively. The sensitivity of the LDA was similar to that of the SVM.

Table 1
Characteristics of the patients with breast cancer (n = 550).

Variable	Number (%)	Train set		Test set		*P*-value
		Mean	Sd	Mean	Sd	
Stage						0.833
I	110 (20.00)	77.80	4.88	32.20	4.88	
II	228 (41.46)	162.49	4.96	68.51	4.96	
III	188 (34.18)	131.72	5.03	57.28	5.03	
IV	24 (4.36)	16.99	2.23	7.01	2.23	
Grade						0.835
1	66 (12.00)	46.51	3.36	19.49	3.36	
2	288 (52.36)	204.92	5.35	87.08	5.35	
3	196 (35.64)	137.57	5.63	59.43	5.63	
Metastasis						0.903
No	467 (84.91)	331.1	3.74	140.9	3.74	
Yes	83 (15.09)	87.9	3.74	25.1	3.74	
Estrogen receptor						0.841
Negative	158 (28.73)	112.21	5.04	47.79	5.04	
Positive	392 (71.27)	276.79	5.04	118.21	5.04	
Progesterone receptor						0.916
Negative	174 (31.67)	124.09	4.99	51.91	4.99	
Positive	376 (68.36)	264.91	4.99	114.09	4.99	
Human epidermal growth factor receptor 2						0.871
Negative	420 (76.36)	296.36	4.53	127.36	4.53	
Positive	130 (23.64)	92.36	4.53	38.64	4.53	
Pathological type						0.931
Ductal/lobular carcinoma in situ	29 (5.27)	20.74	2.43	8.26	2.43	
Invasive lobular carcinoma	25 (4.54)	17.44	2.22	7.56	2.22	
Invasive ductal carcinoma	496 (90.19)	350.82	3.25	150.18	3.25	
Surgical approach						0.780
Modified Radical Mastectomy	192 (34.91)	252.74	5.85	108.26	5.85	
Breast-conserving surgery	358 (65.09)	136.26	5.85	57.74	5.85	
Age (mean (sd))	47.86 (11.79)	52.61	0.38	52.50	0.80	0.181

Sd: Standard deviation.

The mean PPV of the methods ranged between 0.69 and 0.82 for the test sets over 100 repetitions with the lowest and highest belonged to the AdaBoost and the RF respectively and the mean NPV performance of all methods was greater than 0.92. Furthermore, the total accuracy of the used classification schemes ranged between 0.89 (for the AdaBoost) and 0.93 (SVM and LDA).

The mean LR+ of the methods ranged between 10.17 (for the AdaBoost) and 24.33 (for the SVM and LDA) for the test sets over 100 repetitions. The values of the LR- for the models were varied between 0.27 and 0.41 on averages for the test sets with the lowest and highest belonged to the LDA and the LSSVM and AdaBoost respectively.

3.1.2. Performance of the models in predicting metastasis

Table 3 shows the performance of the eight classifiers for prediction of metastasis of the patients with BC in terms of sensitivity, specificity, PPV, NPV, LR+, LR- and total accuracy obtained from 100 time of repetition by cross validation strategy under two scenarios. As shown, all of the used algorithms had high specificity (≥90%) on average for the test sets with the minimum for the AdaBoost and the maximum for the RF (0.98). On the other hand, all the methods had low sensitivity (on average) ranged between 0.07 (for the RF) and 0.36 (for the Naïve Bayes) on average.

The mean PPV of the methods ranged between 0.32 (AdaBoost and SVM) and 0.61 (for the LDA) for the test sets over 100 repetition and the mean NPV performance of all methods was greater than 0.82 (the minimum belonged to the Adabag and the maximum belonged to the Naïve Bayes). Furthermore, the total accuracy of the used classification schemes ranged between 0.80 (for the AdaBoost) and 0.86 (for the LR and the LDA).

The mean LR+ of the methods ranged between 2.81 (for the AdaBoost) and 9.05 (for the LDA) for the test sets over 100 repetitions. The values of the LR- for the models were varied between 0.71 and 0.94 on averages for the test sets with the lowest and highest belonged to the NB and the RF respectively.

4. Discussion

In the present study, several machine learning methods were utilized and compared to predict survival and metastasis in patients with breast cancer. In this regard, the six models of machine learning NB, LSSVM, AdaBoost, Adabag, SVM, RF as well as two classical methods of LR and LDA were applied in the prediction of breast cancer survival and metastasis.

Based on total accuracy, it was shown that all the classification methods performed almost similar for classifying BC survival and metastasis over the test sets (ranged between 0.80 and 0.93). In other words, all the classification methods were quite efficient in predicting the classes for BC survival status and metastasis.

In terms of sensitivity, the SVM and LDA predicted survival of the patients. The minimum sensitivity was 0.61 (AdaBoost), and the maximum value was 0.73 (SVM and LDA). Moreover, despite good performance of the methods in terms of specificity and total accuracy, the sensitivity for predicting metastasis of the breast cancer patients was relatively poor (ranged between 0.07 and 0.36). So, none of the methods has the sensitivity greater than 0.5.

Considering that surviving and experiencing metastasis are the key predictions in this biomedical application, a classifier that has higher sensitivity is preferred. So, an efficient classification method must have the ability to predict a potentially future survived/metastatic patient using the predictor variables. As shown here, none of the classifiers performed well enough to predict the metastasis of the patients in our study. However, for survival outcome, all the methods had sensitivity greater than 0.5 and at the same time had good specificity and total accuracy. Another point that should be mentioned here is that in spite

Table 2
The values of the performance criteria for NB, LSSVM, AdaBoost, Adabag, SVM, RF, LR and LDA by cross validation strategy over 100 repetitions in predicting survival of the breast cancer patientsunder two scenarios.

Scenario	Method	Set	Sensitivity	Specificity	Positive predicted value	Negative predicted value	Positive Likelihood Ratio	Negative Likelihood Ratio	Total accuracy
(70% train, 30% test)	Naïvebayes	Train	0.74 ± 0.04	0.96 ± 0.01	0.80 ± 0.03	0.95 ± 0.01	18.50 ± 7.68	0.27 ± 0.08	0.93 ± 0.01
		Test	0.69 ± 0.07	0.96 ± 0.02	0.78 ± 0.07	0.94 ± 0.02	17.25 ± 6.27	0.32 ± 0.07	0.92 ± 0.02
	LS-SVM	Train	0.79 ± 0.05	0.98 ± 0.01	0.90 ± 0.03	0.96 ± 0.01	39.50 ± 6.23	0.21 ± 0.04	0.95 ± 0.01
		Test	0.62 ± 0.09	0.97 ± 0.01	0.78 ± 0.08	0.93 ± 0.02	20.67 ± 11.73	0.39 ± 0.07	0.91 ± 0.02
	Adabag	Train	0.73 ± 0.03	0.97 ± 0.01	0.82 ± 0.02	0.95 ± 0.01	24.33 ± 6.23	0.28 ± 0.04	0.93 ± 0.01
		Test	0.70 ± 0.08	0.97 ± 0.01	0.81 ± 0.07	0.94 ± 0.02	23.33 ± 11.73	0.31 ± 0.07	0.92 ± 0.02
	AdaBoost	Train	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.01	99.00 ± 4.27	0.01 ± 0.04	0.99 ± 0.01
		Test	0.61 ± 0.10	0.94 ± 0.02	0.69 ± 0.08	0.92 ± 0.02	10.17 ± 11.73	0.41 ± 0.07	0.89 ± 0.02
	RF	Train	0.72 ± 0.04	0.97 ± 0.01	0.85 ± 0.03	0.95 ± 0.01	24.00 ± 4.30	0.29 ± 0.04	0.93 ± 0.01
		Test	0.68 ± 0.08	0.97 ± 0.01	0.82 ± 0.07	0.94 ± 0.02	22.67 ± 11.73	0.33 ± 0.07	0.92 ± 0.02
	SVM	Train	0.73 ± 0.03	0.97 ± 0.01	0.81 ± 0.03	0.95 ± 0.01	24.33 ± 4.27	0.28 ± 0.04	0.93 ± 0.01
		Test	0.73 ± 0.07	0.97 ± 0.01	0.81 ± 0.07	0.95 ± 0.01	24.33 ± 11.73	0.28 ± 0.07	0.92 ± 0.02
	Logit	Train	0.72 ± 0.03	0.97 ± 0.01	0.84 ± 0.02	0.95 ± 0.01	24.00 ± 4.30	0.29 ± 0.04	0.93 ± 0.01
		Test	0.68 ± 0.07	0.97 ± 0.02	0.81 ± 0.08	0.94 ± 0.02	22.67 ± 11.73	0.33 ± 0.07	0.92 ± 0.02
	LDA	Train	0.73 ± 0.03	0.96 ± 0.01	0.80 ± 0.03	0.95 ± 0.01	18.25 ± 4.27	0.28 ± 0.04	0.93 ± 0.01
		Test	0.73 ± 0.07	0.97 ± 0.01	0.81 ± 0.07	0.95 ± 0.01	24.33 ± 11.72	0.27 ± 0.07	0.93 ± 0.02
(50% train,50% test)	Naïvebayes	Train	0.74 ± 0.05	0.96 ± 0.01	0.80 ± 0.05	0.95 ± 0.01	18.5 ± 8.74	0.27 ± 0.05	0.93 ± 0.01
		Test	0.69 ± 0.07	0.96 ± 0.01	0.78 ± 0.06	0.94 ± 0.02	17.25 ± 9.61	0.32 ± 0.07	0.91 ± 0.01
	LS-SVM	Train	0.83 ± 0.06	0.99 ± 0.006	0.95 ± 0.04	0.97 ± 0.01	83.00 ± 8.74	0.17 ± 0.05	0.97 ± 0.01
		Test	0.62 ± 0.07	0.96 ± 0.01	0.76 ± 0.06	0.93 ± 0.02	15.25 ± 9.61	0.41 ± 0.07	0.90 ± 0.02
	Adabag	Train	0.71 ± 0.06	0.97 ± 0.009	0.83 ± 0.04	0.94 ± 0.01	23.67 ± 8.74	0.30 ± 0.05	0.93 ± 0.01
		Test	0.70 ± 0.08	0.97 ± 0.01	0.81 ± 0.05	0.94 ± 0.02	23.33 ± 9.61	0.31 ± 0.07	0.92 ± 0.01
	Adaboost	Train	0.99 ± 0.02	0.99 ± 0.002	0.99 ± 0.01	0.99 ± 0.01	99.00 ± 8.75	0.01 ± 0.05	0.99 ± 0.002
		Test	0.61 ± 0.07	0.94 ± 0.02	0.69 ± 0.06	0.92 ± 0.02	10.33 ± 9.61	0.40 ± 0.07	0.89 ± 0.01
	RF	Train	0.71 ± 0.06	0.98 ± 0.01	0.87 ± 0.04	0.94 ± 0.01	35.5 ± 8.74	0.30 ± 0.05	0.93 ± 0.01
		Test	0.66 ± 0.07	0.97 ± 0.01	0.82 ± 0.04	0.93 ± 0.02	22.00 ± 9.61	0.35 ± 0.07	0.92 ± 0.01
	SVM	Train	0.72 ± 0.06	0.97 ± 0.01	0.82 ± 0.04	0.95 ± 0.01	24.00 ± 8.75	0.29 ± 0.05	0.93 ± 0.01
		Test	0.72 ± 0.06	0.97 ± 0.01	0.81 ± 0.04	0.95 ± 0.01	24.00 ± 8.75	0.29 ± 0.05	0.93 ± 0.01
	Logit	Train	0.72 ± 0.06	0.97 ± 0.01	0.85 ± 0.04	0.95 ± 0.01	24.00 ± 8.75	0.29 ± 0.05	0.93 ± 0.01
		Test	0.68 ± 0.07	0.97 ± 0.01	0.80 ± 0.06	0.94 ± 0.01	22.67 ± 9.61	0.33 ± 0.07	0.92 ± 0.01
	LDA	Train	0.73 ± 0.04	0.97 ± 0.01	0.81 ± 0.04	0.95 ± 0.01	24.33 ± 8.74	0.28 ± 0.05	0.93 ± 0.01
		Test	0.73 ± 0.05	0.97 ± 0.01	0.81 ± 0.04	0.95 ± 0.01	24.33 ± 9.61	0.28 ± 0.07	0.93 ± 0.02

of similarity between the characteristics of the patients in the train and test sets over 100 repetitions, there can be observed differences between the results over training and test sets except for the SVM. It is well known that the SVM minimizes the structural risk instead of empirical risk. This avoids the SVM to be trapped in a local minimum instead of the global minimum and the overfitting issue.[27].Considering two criteria of positive and negative likelihood ratios for predicting survival of the patient, the SVM outperformed other machine learning techniques.

The goodness of classifying cases into a specific category of an outcome variable depends on several factors such as the imbalance sample size of outcome categories. In other words, some classification methods are good for high prevalent outcomes and some for low prevalent.[27] Several studies have shown unbalanced efficiency of different classification tools regarding the small/high frequency of the outcome.[31] In addition to the distribution of the outcome categories, the mechanism in which the predictors affect the outcome is responsible for determining the best classification method. Hence, controversy might be observed in the performance of classification methods in different areas of medicine and clinics. Training and testing sets of the data are chosen randomly. However, to validate the results, it is suggested to repeat the cross validation process. Therefore, to rely on the results, the cross validation strategy was performed over 100 repetitions to predict survival and metastasis of BC patients.

Many of the studies have compared the performance of various classification methods to predict an outcome of interest. Sountharrajan et al. conducted a study on automatic classification on bio medical prognosis of invasive breast cancer. They compared three different methods of SVM, C4.5 Decision tree, Naïve Bayes to provide for earlier prognosis of breast cancer which is helpful to improve the survivability of patients.[32] Salazar et al., found the same results in evaluating the difference between SVM and logistic regression.[33].To assess the factors associated with macrosomia among singleton live-birth, logistic regression, random forest and artificial neural network methods were compared. It was shown that random forest has the highest accuracy (89%) in comparison to logistic regression (64%) and artificial neural network (62%).[34] Amini et al. also found SVM as the best classifier of fatal suicide attempts in comparison to decision tree, logistic regression and artificial neural network. They suggested SVM against other classification tools due to its practical application according to the robustness.[35]

This study focused on the performance of different classifiers in detecting survival and metastasis of BC patients. Our finding showed that the SVM outperformed other machine learning methods in prediction of survival of the patients in terms of several criteria. Nevertheless, the LDA technique as a classical method showed similar performance. On the other hand, their performance in terms of sensitivity was poor for the prediction of metastasis. The used machine learning techniques are nonparametric and provide efficient solutions for classification problems without considering any special assumption regarding the distribution of data. They also deal with nonlinearity and high order interactions. However, the performance of a method is data dependent and in general, there is no method that always performs as the best technique in classification problems.

5. Limitations

There were some limitations in the present study. The data used here was based on a registry-based retrospective study that makes the analysis prone to potential biases for the estimations for criteria like sensitivity and so on. Censoring was another limitation in our study that may result in overestimation or underestimation of the results. Further

Table 3
The values of the performance criteria for NB, LSSVM AdaBoost, Adabag, SVM, RF, LR and LDA by cross validation strategy over 100 repetition in predicting metastasis of the breast cancer patientsunder two scenarios.

Scenario	Method	Set	Sensitivity	Specificity	Positive predicted value	Negative predicted value	Positive Likelihood Ratio	Negative Likelihood Ratio	Total accuracy
(%70 train, %30test)	Naïvebayes	Train	0.36 ± 0.04	0.94 ± 0.01	0.51 ± 0.05	0.89 ± 0.01	5.93 ± 0.98	0.68 ± 0.05	0.85 ± 0.01
		Test	0.33 ± 0.08	0.94 ± 0.02	0.49 ± 0.14	0.89 ± 0.02	5.50 ± 3.54	0.72 ± 0.09	0.85 ± 0.02
	LS-SVM	Train	0.46 ± 0.10	0.98 ± 0.01	0.81 ± 0.01	0.91 ± 0.01	30.20 ± 17.72	0.55 ± 0.10	0.90 ± 0.02
		Test	0.21 ± 0.07	0.95 ± 0.02	0.44 ± 0.14	0.87 ± 0.02	5.07 ± 3.07	0.84 ± 0.08	0.84 ± 0.02
	Adabag	Train	0.33 ± 0.06	0.98 ± 0.01	0.74 ± 0.05	0.89 ± 0.01	18.44 ± 10.24	0.69 ± 0.06	0.88 ± 0.01
		Test	0.20 ± 0.07	0.97 ± 0.02	0.53 ± 0.16	0.82 ± 0.02	6.67 ± 3.84	0.83 ± 0.07	0.85 ± 0.02
	Adaboost	Train	0.92 ± 0.03	0.99 ± 0.01	0.97 ± 0.02	0.99 ± 0.01	189.84 ± 56.78	0.08 ± 0.03	0.98 ± 0.01
		Test	0.25 ± 0.08	0.91 ± 0.02	0.32 ± 0.09	0.87 ± 0.02	2.81 ± 1.01	0.83 ± 0.08	0.81 ± 0.02
	RF	Train	0.20 ± 0.07	0.99 ± 0.01	0.86 ± 0.06	0.88 ± 0.01	35.24 ± 15.14	0.81 ± 0.07	0.87 ± 0.01
		Test	0.08 ± 0.05	0.98 ± 0.01	0.49 ± 0.20	0.86 ± 0.02	5.47 ± 3.82	0.94 ± 0.04	0.85 ± 0.02
	SVM	Train	0.58 ± 0.14	0.99 ± 0.01	0.94 ± 0.07	0.93 ± 0.04	59.75 ± 42.27	0.43 ± 0.26	0.93 ± 0.04
		Test	0.14 ± 0.11	0.95 ± 0.03	0.32 ± 0.17	0.87 ± 0.02	3.33 ± 2.47	0.89 ± 0.10	0.83 ± 0.02
	Logit	Train	0.26 ± 0.06	0.97 ± 0.01	0.63 ± 0.07	0.88 ± 0.02	10.27 ± 3.65	0.76 ± 0.06	0.86 ± 0.02
		Test	0.21 ± 0.07	0.97 ± 0.02	0.59 ± 0.15	0.87 ± 0.02	7.61 ± 3.73	0.81 ± 0.07	0.86 ± 0.02
	LDA	Train	0.28 ± 0.05	0.97 ± 0.01	0.66 ± 0.06	0.89 ± .01	11.88 ± 6.03	0.74 ± 0.05	0.87 ± 0.01
		Test	0.26 ± 0.09	0.97 ± 0.01	0.61 ± 0.13	0.88 ± 0.02	9.05 ± 3.90	0.77 ± 0.09	0.86 ± 0.02
(%50 train,%50 test)	Naïvebayes	Train	0.39 ± 0.07	0.94 ± 0.01	0.53 ± 0.07	0.90 ± 0.01	6.86 ± 2.07	0.65 ± 0.07	0.86 ± 0.02
		Test	0.34 ± 0.07	0.93 ± 0.02	0.46 ± 0.08	0.89 ± 0.02	5.20 ± 1.69	0.71 ± 0.07	0.84 ± 0.02
	LS-SVM	Train	0.55 ± 0.11	0.98 ± 0.01	0.84 ± 0.06	0.93 ± 0.02	30.20 ± 17.72	0.46 ± 0.11	0.92 ± 0.02
		Test	0.24 ± 0.07	0.93 ± 0.02	0.39 ± 0.09	0.88 ± 0.02	3.95 ± 1.61	0.81 ± 0.07	0.83 ± 0.02
	Adabag	Train	0.33 ± 0.09	0.98 ± 0.01	0.76 ± 0.06	0.89 ± 0.01	19.97 ± 8.19	0.68 ± 0.09	0.88 ± 0.01
		Test	0.21 ± 0.07	0.96 ± 0.02	0.51 ± 0.11	0.87 ± 0.02	6.59 ± 3.84	0.82 ± 0.07	0.85 ± 0.02
	Adaboost	Train	0.94 ± 0.04	0.99 ± 0.01	0.98 ± 0.02	0.99 ± 0.01	189.84 ± 56.78	0.06 ± 0.04	0.99 ± 0.01
		Test	0.27 ± 0.08	0.90 ± 0.02	0.33 ± 0.06	0.88 ± 0.02	2.83 ± 0.77	0.82 ± 0.07	0.81 ± 0.02
	RF	Train	0.18 ± 0.10	0.99 ± 0.01	0.86 ± 0.06	0.87 ± 0.01	35.24 ± 15.14	0.82 ± 0.10	0.87 ± 0.01
		Test	0.07 ± 0.05	0.98 ± 0.01	0.49 ± 0.20	0.86 ± 0.02	5.47 ± 3.82	0.94 ± 0.05	0.85 ± 0.01
	SVM	Train	0.49 ± 0.22	0.99 ± 0.01	0.91 ± 0.06	0.92 ± 0.03	65.67 ± 57.67	0.51 ± 0.21	0.92 ± 0.03
		Test	0.14 ± 0.09	0.96 ± 0.03	0.37 ± 0.16	0.87 ± 0.02	4.85 ± 6.58	0.89 ± 0.08	0.84 ± 0.02
	Logit	Train	0.28 ± 0.08	0.97 ± 0.01	0.64 ± 0.09	0.89 ± 0.01	11.47 ± 4.68	0.74 ± 0.08	0.86 ± 0.02
		Test	0.24 ± 0.09	0.96 ± 0.02	0.55 ± 0.11	0.87 ± 0.02	7.61 ± 3.73	0.79 ± 0.08	0.86 ± 0.02
	LDA	Train	0.30 ± 0.07	0.97 ± 0.01	0.67 ± 0.09	0.89 ± .01	13.65 ± 9.21	0.72 ± 0.07	0.87 ± 0.01
		Test	0.27 ± 0.09	0.97 ± 0.02	0.59 ± 0.09	0.88 ± 0.02	9.05 ± 3.90	0.75 ± 0.09	0.86 ± 0.01

studies with large sample sizes are required to investigate the performance of these techniques deeply and more reliably. There are many other factors that need to be considered and best addressed by a parsimonious, robust epidemiological model.

6. Conclusion

The focus of this study was on evaluating the performance of six machine learning and two classical techniques in predicting survival and metastasis occurrence in patients with breast cancer. Our finding showed that the SVM and LDA were the best models to predict survival in terms of several criteria and the LDA was the best technique to predict metastasis among BC patients in this study. Further investigation is needed using large data sets to recommend a useful tool for BC survival and metastasis prediction.

Acknowledgement

We would like to appreciate the Vice-chancellor of Research of Hamadan University of Medical Sciences for their approval and support of this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cegh.2018.10.003.

Conflicts of interest

The authors declare no conflict of interest.

References

1. World Health Organization. Breast cancer: breast cancer and early diagnosis. . Available from: http://wwwwhoint/cancer/prevention/diagnosis-screening/breast-cancer/en/; 2018.
2. Beiki O, Hall P, Ekbom A, Moradi T. Breast cancer incidence and case fatality among 4.7 million women in relation to social and ethnic background: a population-based cohort study. *Breast Canc Res.* 2012;14(1):R5.
3. Liu X, Wang L, Zhang J, Yin J, Liu H. Global and local structure preservation for feature selection. *IEEE Trans. Neural Networks. Learn. Syst.* 2014;25(6):1083–1095.
4. Applying a multi-agent classifier system with a novel trust measurement method to classifying medical data. In: Mohammed MF, Lim CP, bt Ngah UKeds. *The 8th International Conference on Robotic, Vision, Signal Processing & Power Applications.* Springer; 2014.
5. Zhang L, Riethdorf S, Wu G, et al. Meta-analysis of the prognostic value of circulating tumor cells in breast cancer. *Clin Canc Res.* 2012;18(20):5701–5710.
6. Reif M, Shafait F. Efficient feature size reduction via predictive forward selection. *Pattern Recogn.* 2014;47(4):1664–1673.
7. Hartmann S, Reimer T, Gerber B. Management of early invasive breast cancer in very young women (< 35 years). *Clin Breast Canc.* 2011;11(4):196–203.
8. Nahar J, Imam T, Tickle KS, Ali AS, Chen Y-PP. Computational intelligence for microarray data and biomedical image analysis for the early diagnosis of breast cancer. *Expert Syst Appl.* 2012;39(16):12371–12377.
9. Gayathri B, Sumathi C, Santhanam T. Breast cancer diagnosis using machine learning algorithms-a survey. *Int. J. Distr. Parallel Syst.* 2013;4(3):105.
10. Rajesh K, Anand S. Analysis of SEER dataset for breast cancer diagnosis using C4. 5 classification algorithm. *Int. J. Adv. Res. Comp. Commun. Eng,.* 2012;1(2):2278 1021.
11. Hashemian AH, Beiranvand B, Rezaei M, Bardideh A, Zand-Karimi E. Comparison of artificial neural networks and cox regression models in prediction of kidney transplant survival. *Int. J. Adv. Biol. Biomed. Res.* 2013;1(10):1204–1212.
12. Montazeri M, Montazeri M, Naji HR, Faraahi A, eds. *A Novel Memetic Feature Selection Algorithm. Information and Knowledge Technology (IKT), 2013 5th Conference on.* IEEE; 2013.
13. Das R, Sengur A. Evaluation of ensemble methods for diagnosing of valvular heart disease. *Expert Syst Appl.* 2010;37(7):5110–5115.
14. AlkıM E, Gürbüz E, Kılıç E. A fast and adaptive automated disease diagnosis method with an innovative neural network model. *Neural Network.* 2012;33:88–96.
15. Zheng B, Yoon SW, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Syst Appl.* 2014;41(4):1476–1482.
16. Montazeri M, Montazeri M, Montazeri M, Beigzadeh A. Machine learning models in breast cancer survival prediction. *Technol Health Care.* 2016;24(1):31–42.
17. Chao C-M, Yu Y-W, Cheng B-W, Kuo Y-L. Construction the model on the breast cancer

survival analysis use support vector machine, logistic regression and decision tree. *J Med Syst.* 2014;38(10):106.
18. Gelnarová E, Safarik L. Comparison of three statistical classifiers on a prostate cancer data. *Neural Netw World.* 2005;15(4):311.
19. Austin PC. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Stat Med.* 2007;26(15):2937–2957.
20. Finch H, Schneider MK. Classification accuracy of neural networks vs. Discriminant analysis, logistic regression, and classification and regression trees. *Methodology.* 2007;3(2):47–57.
21. Ferrer L, Rondeau V, Dignam J, Pickles T, Jacqmin-Gadda H, Proust-Lima C. Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer. *Stat Med.* 2016;35(22):3933–3948.
22. Witten IH, Frank E, Hall MA, Pal CJ. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann; 2016.
23. Bellaachia A, Guven E. Predicting breast cancer survivability using data mining techniques. *Age.* 2006;58(13):10–110.
24. Agresti A, Kateri M. *Categorical Data Analysis.* Springer; 2011.
25. Izenman AJ. *Linear Discriminant Analysis. Modern Multivariate Statistical Techniques.* Springer; 2013:237–280.
26. Breiman L. Random forests. *Mach Learn.* 2001;45(1):5–32.
27. Tapak L, Mahjub H, Hamidi O, Poorolajal J. Real-data comparison of data mining methods in prediction of diabetes in Iran. *Healthcare informatics research.* 2013;19(3):177–185.
28. Auria L, Moro RA. *Support Vector Machines (SVM) as a Technique for Solvency Analysis.* 2008; 2008.
29. Tripathy RK, Zamora-Mendez A, de la O Serna JA, Paternina MRA, Arrieta JG, Naik GR. Detection of life threatening ventricular arrhythmia using digital taylor fourier transform. *Front Physiol.* 2018;9:722.
30. Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J. Intern. Med.* 2013;4(2):627–635.
31. Meyer D, Leisch F, Hornik K. The support vector machine under test. *Neurocomputing.* 2003;55(1-2):169–186.
32. Sountharrajan S, Karthiga M, Suganya E, Rajan C. Automatic classification on bio medical prognosisof invasive breast cancer. *Asian Pac J Cancer Prev APJCP: APJCP.* 2017;18(9):2541.
33. Salazar DA, Vélez JI, Salazar JC. Comparison between SVM and logistic regression: which one is better to discriminate? *Rev Colomb Estadística.* 2012;35(2):223–237.
34. Amini P, Maroufizadeh S, Hamidi O, Samani RO, Sepidarkish M. Factors associated with macrosomia among singleton live-birth: a comparison between logistic regression, random forest and artificial neural network methods. *Epidemiol. Biostat. Public Health.* 2016;13(4).
35. Amini P, Ahmadinia H, Poorolajal J, Amiri MM. Evaluating the high risk groups for suicide: a comparison of logistic regression, support vector machine, decision tree and artificial neural network. *Iran J Public Health (English ed).* 2016;45(9):1179.