Prediction breast cancer risk: Performance analysis data mining techniques

Solmaz Sohrabi^{1*}, Alireza Atashi²

¹Department of Medical Informatics, Shahid Beheshti University of Medical Sciences, Tehran, Iran

²Department of E-Health, Virtual School, Tehran University of Medical Sciences, Medical Informatics Department, Breast Cancer Research Center, Motamed Cancer Institute (ACECR), Tehran, Iran

Article Info

Article type:
Research

Article History:

Received: 2021-04-21
Accepted: 2021-07-11
Published: 2021-07-25

* Corresponding author:

Solmaz Sohrabi

Department of Medical Informatics,
Shahid Beheshti University of
Medical Sciences, Tehran, Iran

Email: elnazfatemi143@gmail.com

Keywords:

Breast Cancer
Data Mining
Classifiers

ABSTRACT

Introduction: Early detection breast cancer Causes it most curable cancer in among other types of cancer, early detection and accurate examination for breast cancer ensures an extended survival rate of the patients. Risk factors are an important parameter in breast cancer has an important effect on breast cancer. Data mining techniques have a growing reputation in the medical field because of high predictive capability and useful classification. These methods can help practitioners to develop tools that allow detecting the early stages of breast cancer.

Material and Methods: The database used in this paper is provided by Motamed Cancer Institute, ACECR Tehran, Iran. It contains of 7834 records of breast cancer patients clinical and risk factors data. There were 4008 patients (52.4%) with breast cancers (malignant) and the remaining 3617 patients (47.6%) without breast cancers (benign). Support vector machine, multi-layer perceptron, decision tree, K nearest neighbor, random forest, naïve Bayesian models were developed using 20 fields (risk factor) of the database because database feature was restrictions. Used 10-fold crossover for models evaluate. Ultimately, the comparison of the models was made based on sensitivity, specificity and accuracy indicators.

Results: Naïve Bayesian and artificial neural network are better models for the prediction of breast cancer risks. Naïve Bayesian had accuracy of 93%, specificity of 93.32%, sensitivity of 95.05%, ROC of 0.95 and artificial neural network had accuracy of 93.23%, specificity of 91.98%, sensitivity of 92.69%, and ROC of 0.8.

Conclusion: Strangely the different artificial intelligent calculations utilized in this examination yielded close precision subsequently these techniques could be utilized as option prescient instruments in the bosom malignancy risk considers. The significant prognostic components affecting risk pace of bosom disease distinguished in this investigation, which were approved by risk, are helpful and could be converted into choice help devices in the clinical area.

Cite this paper as:

Sohrabi S, Atashi A. Prediction breast cancer risk: Performance analysis data mining techniques. Front Health Inform. 2021; 10: 83. DOI: [10.30699/fhi.v10i1.296](https://doi.org/10.30699/fhi.v10i1.296)

INTRODUCTION

In worldwide more than one million new instances of female bosom malignant growth are analyzed every year. Bosom malignancy presents genuine danger to the existences of individuals and it is the subsequent driving reason for death in ladies today and It is the most well-known reason for female demise in industrialized nations, the second most normal reason on the planet and the third generally regular in non-industrial nations [1-3]. Roughly 10-15% of patients with bosom malignancy bite the dust of

disease metastasis or repeat, and early determination of it can improve guess. Death pace of ladies due to bosom malignant growth can be diminished if can be distinguished at a generally beginning phase. Early forecast of bosom disease assumes a basic part in fruitful treatment and saving existences of thousands of patients consistently. Information mining has become a basic approach for processing applications in the space of medication [4-6]. With the assistance of most recent, proficient and early showing techniques, most of such malignant growths are analyzed when the sickness is still at a restricted

stage [7, 8]. The utility of artificial intelligent (AI) strategies in medical care examination is developing dynamically [9-11]. Albeit gigantic clinical information identified with the patients is being gathered and put away by medical services associations, just a little subset of the prescient elements has been utilized in anticipating results [2, 12-14]. The vast majority of the current methodologies center around applying factual strategies on little arrangement of properties suggested by the space specialist's illness diagnostics [15-18]. These traditional methodologies typically make unreasonable suspicions, e.g., ordinariness, freedom or linearity connections, which may not be in every case valid in down to earth information. Then again, progressed factual methodologies may address a portion of the above deficiencies; in any case, they are computationally costly and may not appropriate to gigantic datasets [19]. For the most part, in the clinical world, there are two stages for settling on the choices. These two stages are (1): differential diagnosis (DD); in this stage, all data of patients including their clinical history, side effects of sickness, consequences of different testing, for example, blood testing and so forth are seen by specialists as the info information. This information is handled by specialists dependent on their clinical information for infection analysis. In some cases, a few infections have some comparative side effects; accordingly, clinical specialists should be relegated subjective loads to every last one of info sand make designs, match these examples with the examples of different illnesses lastly select the nearest match and analysis the specific sickness. (2) Final or provisional diagnosis (FD): in this stage, the fundamental proposals and medicines would be beginning as per the distinguished illness. In this progression, a doctor with clinical information and his/her rationale, proceeds with exams and records the consequences of consistently sees or tests, and chooses the last medicines and anticipation [20]. Information mining has different procedures (for example, classification, clustering, regression, association rules and so forth) and calculations (for example, decision trees, genetic algorithm, nearest neighbor strategy and so on) for breaking down the tremendous measure of crude or multi-dimensional information [21]. In different words, information digging has capacities for clever information examination to remove concealed information from huge data sets of clinical or clinical information that are gathered from clinical focuses or clinics. This information gives helpful data to improve choice help, counteraction, finding and therapy in clinical world. Further, information mining has capacity to recognize affiliation administers or build up connections between different highlights like patient's very own information, infection side effects and so on [22, 23]. This examination endeavors to address the consequences of a few exploration works which are distributed identified with information

mining applications in expectation, conclusion or treatment of breast cancers [24, 25]. The recent breakthrough of machine learning and data mining techniques has opened a new door for healthcare diagnostic and prediction. In this paper, we utilize these advancements for predicting breast cancer risk based on a labeled dataset [26, 27]. Particularly, we utilize the clinical data as well as the associated symptoms of patients to construct predictive models which can classify patients into different breast cancer categories, benign or malignant [28, 29]. The model is constructed by using the industry standard software package Rapidminer (ver. 20.9). The proposed approaches offer significant advantages over the conventional approaches with high classification accuracy. The organization of the paper is as follows [30]. We describe our dataset, methods and background on data mining techniques. We discuss our data exploration and predictive construction for predicting breast cancer. We present and discuss our results.

MATERIAL AND METHODS

The present study is of a fundamental analytical type. Databases used in this paper is provided by Motamed Cancer Institute, ACECR Tehran, Iran which contained the data of patients who referred to this center from 2010 to 2020. This study aimed to compare the performance of six data mining algorithm including support vector machine (SVM), multi-layer perceptron (MLP), decision tree (DT), K nearest neighbor (KNN), random forest (RF), naïve Bayesian (NB) for the prediction of breast cancer risk. This database contained 7834 information breast cancer records. Firstly, all columns that were not related to disease risk factors were removed from the database. Records that lost 50 percent of their data were omitted on condition that three important risk factors, such as personal history of breast cancer, family history of breast cancer, and hormone therapy were also lost, the database did not contain complete gene mutation modifications. The number of missing data was very large, so it was not included in the study because the placement of the lost data with the techniques would create fictitious data for use in the models. Eventually, the number of database records reached 7687. Records with lost data were then filled in with the median placement method [31] in the SPSS (ver. 26) software; with this action, the number of records in the database reached 7600 records. After this stage, the number of malignant records was 1900 (25%) and the number of benign records was 5700 (75%), which was used to balance the classes in the database using the SMOTE technique. This technique is capable of balancing unbalanced databases [32, 33].

The total number of records increased to 13460,

6999 records (52.4%) were data related to malignant breast cancer patients and 6460 records (47.6%) were data related to benign breast cancer patients. After balancing and selecting the risk factors, prioritization was performed with the CART decision tree algorithm [34]. The models were constructed using seventeen risk factors for breast cancer in Rapidminer (ver. 20.9) based on sensitivity, specificity and accuracy, taking into account 9 parts of the data for training and 1 part for testing. Cross Validation 10-Fold evaluation method (10-point mutual validation) were divided .

Table 1: Factors Risk Cancer Breast Considered in the Study

Risk factors	Type	Range
Age of Diagnosis	Quantitative_Discrete	38-89
Age of the first Menstruation	Quantitative_Discrete	16-11
Menopausal age	Quantitative_Discrete	62-48
The age of first Pregnancy	Quantitative_Discrete	45-18
History of Breastfeeding	Qualitative_Classified	0=No, 1=Yes
Use OCPs	Qualitative_Classified	0=No, 1=Yes
History of hormone therapy after Menopause	Qualitative_Classified	0=No, 1=Yes
History of breast cancer	Qualitative_Classified	0=No, 1=Yes
Family history of breast cancer	Qualitative_Classified	0=No, 1=Yes
Infertility history	Qualitative_Classified	0=No, 1=Yes
Smoking/ alcohol	Qualitative_Classified	0=No, 1=Yes
Marriage status	Qualitative_Classified	0=No, 1=Yes
Education	Qualitative_Classified	0=No, 1=Yes
Bad event of life	Qualitative_Classified	0=No, 1=Yes
BMI	Qualitative_Classified	0=No, 1=Yes
ER	Qualitative_Classified	0=No, 1=Yes
PR	Qualitative_Classified	0=No, 1=Yes
Type of disease(Malignant or Benign)	Qualitative_Classified	Malignant=1 Benign=0

Evaluation method

In order to evaluate the performance of a classification model, several approaches can be used. We employ the confusion matrix for the evaluation of classification models. According to the confusion matrix, the accuracy of a classification model is calculated as follows:

Accuracy

The accuracy of a classifier is the probability of

correctly predicting the class of an unlabeled instance and it can be estimated in several ways.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Sensitivity

Sensitivity describes how well classifier catches all of positive cases. Sensitivity is calculated by dividing the number of true-positive results by the total number of positives (which include false positives).

$$\text{Sensitivity} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Specificity

Specificity describes how well classifier catches negative cases as negatives. Specificity is calculated by dividing the number of true-negative results by the total number of negatives (which include false negatives).

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (3)$$

ROC

One way to evaluate and evaluate the performance of binary classification is the receiver operating characteristic or ROC in short. The performance of binary classifier algorithms is usually measured by indicators called sensitivity or recall. But in the ROC diagram, both of these indicators are combined and displayed as a curve. ROC curves are often used to evaluate the performance of classification algorithms or to generate string data. This issue has received more attention in the field of supervised machine learning.

RESULTS

A total of 7834 records were extracted from the reports and included in the dataset. The dataset contained 1900 records, which were related to the women who had breast cancer (BREAST CANCER=1), and 5700 records, which were for those who had not breast cancer (BREAST CANCER=0).

The NB model is the most sensitive and performance classifier with the values of 95.56% and 95% respectively. The ANN classifier has the best sensitivity and accuracy with the values of 92.69% and 93.23% respectively. Performance comparison of the breast cancer prediction models shown in Table 1 and ROC classifier curve comparison shown in Fig 1.

DISCUSSION

In this study, classifier models have been proposed for helping clinicians to decide about referring the woman to biopsy or not. Although the NB model has been shown to have the best ROC and sensitivity

(84.06%), it seems more sensitive models such as random forest and ANN with the sensitivity about 92% in this case might be recommended to avoid misdiagnosing breast cancer. Treatment choices are especially difficult in beginning phase of breast cancer patients with clashing prognostic risk highlights, particularly hub negative ones, in which whether or not to seek after an adjuvant treatment with chemotherapy or endocrine treatments is as yet hazy [35].

Table 1: Performance comparison of the breast cancer prediction models

Test Results	NB	K-NN	SVM	C5	ANN	RF
Accuracy	93	87.95	91.36	91.98	93.23	91.73
Specificity	93.32	90.01	95.37	92.36	91.98	93.84
Sensitivity	95.56	85.64	93	90.23	92.69	91.66
ROC	0.95	0.90	0.92	0.82	0.80	0.88

Fig 1: ROC of the classifiers

Utilization of information mining strategies in a choice emotionally supportive network for foreseeing bosom malignancy helps and help doctor in making ideal, precise and convenient choice, and lessen the general expense of treatment [36].

Researchers have used a wide variety of machine learning methods for predicting susceptibility diagnosis, recurrence and survivability. Several studies reported a higher accuracy for predicting breast cancer, this could be due to the difference in chosen datasets. This shows that more records can help create more efficient models. Kate et al. [37] used 1.4 million records for breast cancer prediction, reported logistic regression was able to provide higher performance. Diaz et al. [26] proposed a CAD for breast cancer diagnosis. The best accuracy reported for the proposed system was 89.3%. The researchers used naive Bayesian, random forest, and SVM classifiers. Another study by Kaushik et al. [38] utilized data mining techniques for the prediction of breast tissue biopsy results. The researchers proposed a model with an accuracy of 83.5% and a

ROC of 0.907. In [39], the breast cancer prediction was studied using various classification algorithms such as RepTree, J48, and random forest. Dataset used in this study was extracted from SEER repository with 762,691 samples and 134 features. Furthermore, data cleaning and dimensionality reduction techniques were used and seven features were selected for the final classification. Finally, different data mining techniques were compared and analyzed using the WEKA software. Results revealed that the RepTree algorithm in the set of decision tree algorithms acted better in the breast cancer prediction.

Using more features can help create better results. In [40], cancer patient data was collected from Wisconsin dataset of the UCI machine learning. This dataset contains 35 features, which are selected using the feature selection methods and are computed using classification algorithms. According to the results, naive Bayesian algorithm and decision tree provide better and higher accuracy.

This suggests that the use of modified and non-proprietary publicly accessible databases can be effective in creating models with higher results. In this study, the dataset was obtained specific registry dataset format. Kaushik et al. [38] used naïve Bayesian, RBF network, and decision tree techniques to predict breast cancer on Wisconsin dataset provided by the University of California Irvine machine learning repository [39]. They achieved the best accuracy of 97.36% related to the naïve Bayesian classifier. Pritom et al. [40] used decision tree, naïve Bayesian, KNN, and SVM techniques to predict breast cancer risk, they achieved an accuracy of 97.13% related to SVM classifier. Another study [41] classified breast tissue based on mammography reports on public datasets have reached to the accuracy of 99%. Li et al. [42] performed on two databases, which are separately from BCCD and WBCD reported random forest had higher AUC (0.99).

In Iran, studies have been conducted to predict breast cancer, which are mentioned below. In [43] researchers attempt to predict the period of staying alive among those diagnosed with breast cancer. Data were gathered from 5673 patients at Shiraz University of Medical Sciences. The proposed algorithm was logistic regression. The study has achieved the highest level of sensitivity and detection criteria, and it has used the smallest set of attributes like age, tumor size, the ratio of lymphatic glands involvement, invasion and so on. The results demonstrate that although the effect of age on the patient is a controversial matter, aging leads to 60 month decrease of survival. Lotfinezhad Afshar et al. [44] used 899 record from Omid treatment and research center Urmia, Iran, reported C5 in all the evaluation criteria had higher sensitivity (92.21%).

Sohrabi et al. [45] used data mining models. Results showed that MLP had higher sensitivity of 90.88%. Tanha et al. [46] used risk factors features from breast cancer patients from Mahdiah clinic, Kermanshah, Iran database which contains 2555 records. They used 3 data mining models. Results showed that J48 reported higher recall 0.94. Zand [47] used SEER public data base with 3 data mining models for analysis. Results showed that C4.5 accuracy was 86.7%. Mehri Dehnavi [48] used a hybrid method for breast cancer prediction which proposed for identifying more predictive gene signatures from microarray datasets with 1338 records, for analysis utilize neuro-fuzzy classifier, result reported the model has a good ability to analyze data. Lotfnezhad Afshar et al. [44] used 3 models for SEER data analysis for patient survival predicts, result shown SVM reported higher sensitivity of 97.7%.

CONCLUSION

Like some other examination, this investigation has various limits. One of the impediments of the examination is that, because of the accessibility and accessibility of information identified with hazard factors for bosom malignancy, a portion of the information of more persuasive elements identified with bosom disease, for example, sorts of quality changes in this assortment. Changes of two qualities known as BRCA1 and BRCA2 have for quite some time been known to bring about higher dangers of bosom and ovarian malignant growth in ladies. Researchers have additionally as of late found that men with specific transformations of these two qualities may have an expanded danger of beginning stage prostate disease was not accessible. Albeit the data set contained the consequences of some quality change tests, their utilization didn't prompt the development of genuine models because of their

missing qualities, so finishing the information identified with patients' quality transformation tests could be a significant point in information assortment. Utilizing various quantities of folds of cross-approval for a specific arrangement of information sources, the presentation of a choice tree at contrast sizes would be to some degree extraordinary. For similar arrangement of perceptions estimated on a set of information sources, the mistake rate for this set won't be the very same when assessed by three, five, or ten folds at specific sizes of choice tree. The outcome has been shown how information mining models can help specialists in Libyan emergency clinics better comprehend disease hazard factors to make a precise diagnosis. In the end new examination showed prescient bosom disease hazard can be improved by consolidating segment hazard factors, germ line hereditary variations, and mammography highlights. The worth of mammography anomaly includes gives proof that this moderate aggregate consolidates hereditary and ecological danger factors as well as fuses data to improve sickness forecast.

AUTHOR'S CONTRIBUTION

All authors contributed to the literature review, design, data collection and analysis, drafting the manuscript, read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this study.

FINANCIAL DISCLOSURE

No financial interests related to the material of this manuscript have been declared.

REFERENCES

1. Chen HL, Yang B, Liu J, Liu DY. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Systems with Applications*. 2011; 38(7): 9014-22.
2. Kumari M, Singh V. Breast cancer prediction system. *Procedia Computer Science*. 2018; 132: 371-6.
3. Salim EI, Jazieh AR, Moore MA. Lung cancer incidence in the Arab league countries: Risk factors and control. *Asian Pac J Cancer Prev*. 2011; 12(1): 17-34. PMID: 21517227 [PubMed]
4. Chaurasia V, Pal S, Tiwari BB. Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology*. 2018; 12(2): 119-26.
5. Chaurasia V, Pal S. A novel approach for breast cancer detection using data mining techniques. *International Journal of Innovative Research in Computer and Communication Engineering*. 2014; 2(1): 1-17.
6. Venkatesan E, Velmurugan T. Performance analysis of decision tree algorithms for breast cancer classification. *Indian Journal of Science and Technology*. 2015; 8(29): 1-8.
7. Verma D, Mishra N. Analysis and prediction of breast cancer and diabetes disease datasets using data mining classification techniques. *International Conference on Intelligent Sustainable Systems*. IEEE; 2017.
8. Poornima S, Pushpalatha M. A survey of predictive analytics using big data with data mining. *International Journal of Bioinformatics Research and Applications*. 2018; 14(3): 269-82.
9. No Authour. Application of data mining techniques to predict breast cancer. *Procedia Computer Science*.

- 2019; 163: 11-8.
10. Wang H, Yoon SW. Breast cancer prediction using data mining method. *Industrial and Systems Engineering Research Conference. IISE*; 2015.
 11. Burt JR, Torosdagli N, Khosravan N, Ravi Prakash H, Mortazi A, Tissavirasingham F, et al. Deep learning beyond cats and dogs: Recent advances in diagnosing breast cancer with deep neural networks. *Br J Radiol.* 2018, 91(1089): 20170545. PMID: 29565644 DOI: 10.1259/bjr.20170545 [[PubMed](#)]
 12. Hueman MT, Wang H, Yang CQ, Sheng L, Henson DE, Schwartz AM, et al. Creating prognostic systems for cancer patients: A demonstration using breast cancer. *Cancer Med.* 2018, 7(8), 3611–21. PMID: 29968970 DOI: 10.1002/cam4.1629 [[PubMed](#)]
 13. Thakur SS, Li H, Chan AMY, Tudor R, Bigras G, Morris D, et al. The use of automated Ki67 analysis to predict Oncotype DX risk-of-recurrence categories in early-stage breast cancer. *PLoS One.* 2018; 13(1): e0188983. PMID: 29304138 DOI: 10.1371/journal.pone.0188983 [[PubMed](#)]
 14. Lavanya D, Usha Rani K. Performance evaluation of decision tree classifiers on medical datasets. *International Journal of Computer Applications.* 2011; 26(4): 1-4.
 15. Shomona GJ, Ramani R. Discovery of knowledge patterns in clinical data through data mining algorithms: Multi-class categorization of breast tissue data. *International Journal of Computer Applications.* 2011; 32(7): 46-52.
 16. Shajahaan SS, Shanthi S, ManoChitra V. Application of data mining techniques to model breast cancer data. *International Journal of Emerging Technology and Advanced Engineering.* 2013; 3(11): 1-5.
 17. Kharya S, Dubey D, Soni S. Predictive machine learning techniques for breast cancer detection. *International Journal of Computer Science and Information Technologies.* 2013; 4(6): 1023-8.
 18. Nalini C, Meera D. Breast cancer prediction system using data mining methods. *International Journal of Pure and Applied Mathematics.* 2018; 119(12): 10901-11.
 19. Thongkam J, Xu G, Zhang Y, Huang F. Support vector machine for outlier detection in breast cancer survivability prediction. *Advanced Web and Network Technologies, and Applications.* Springer; 2008.
 20. Rana M, Chandorkar P, Dsouza A, Kazi N. Breast Cancer Diagnosis and Recurrence Prediction Using Machine Learning Techniques. *International Journal of Research in Engineering and Technology.* 2015; 4: 372-6.
 21. Sriramakrishnan GV, Muthu Selvam M. Early detection of breast cancer using data mining algorithm based on historical medical records. *International Journal of Advanced Science and Technology.* 2020; 29(7): 8949-55.
 22. Abdull MAS. Data mining techniques and breast cancer prediction: A case study of Libya [PhD Thesis]. Sheffield Hallam University, United Kingdom; 2011.
 23. Rath M, Gupta C. An approach to predict breast cancer and drug suggestion using machine learning techniques. *International Journal on Information Technology.* 2014; 4(1): 1-9.
 24. Eltalhi S, Kutrani H. Breast cancer diagnosis and prediction using machine learning and data mining techniques: A review. *IOSR Journal of Dental and Medical Sciences.* 2019; 18(4): 85-94.
 25. Swetha K, Ranjana R. Breast cancer predication using machine learning and data mining. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology.* 2020; 6(3): 610-5.
 26. Diz J, Marreiros G, Freitas A. Applying data mining techniques to improve breast cancer diagnosis. *J Med Syst.* 2016; 40(9): 203. PMID: 27498205 DOI: 10.1007/s10916-016-0561-y [[PubMed](#)]
 27. Mandadi S, Tejashwini B, Vijayakumar S. Breast cancer classification using data mining. *International Journal of Research in Engineering and Science.* 2020; 8(11): 43-7.
 28. Sultana J, Jilani AK. Predicting breast cancer using logistic regression and multi-class classifiers. *International Journal of Engineering & Technology.* 2018; 7(4); 22-6.
 29. Leenavinmalar F, Kumarkombaiya A. Application of data mining techniques in early detection of breast cancer. *International Journal of Engineering Trends and Technology.* 2018; 56(1): 1-5.
 30. Nishara Banu MA, Gomathy B. Disease predicting system using data mining techniques. *International Journal of Technical Research and Applications.* 2013; 1(5): 41-5.
 31. Han J, Kamber M. *Data mining: Concepts and techniques.* 3rd ed. Morgan Kaufman; 2012.
 32. Liou DM, Chang WP. Applying data mining for the analysis of breast cancer data. *Methods Mol Biol.* 2015; 1246: 175-89. PMID: 25417087 DOI: 10.1007/978-1-4939-1985-7_12 [[PubMed](#)]
 33. Karapinar Senturk Z, Kara R. Breast cancer diagnosis via data mining: Performance analysis of seven different algorithms. *Computer Science & Engineering: An International Journal.* 2014; 4(1): 35-46.
 34. D'Alisa S, Miscio G, Baudo S, Simone A, Tesio L, Mauro A. Depression is the main determinant of quality of life in multiple sclerosis: A classification-regression (CART) study. *Disabil Rehabil.* 2006; 28(5): 307-14. PMID: 16492625 DOI: 10.1080/09638280500191753 [[PubMed](#)]
 35. Garcia Jacob S, Geetha Ramani R. Efficient classifier for classification of prognostic breast cancer data through data mining techniques. *World Congress on Engineering and Computer Science. WCECS*; 2012.
 36. Elsayad AM, Elsalamony HA. Diagnosis of breast cancer using decision tree models and SVM. *International Journal of Computer Applications.* 2013;

- 83(5): 19-29.
37. Kate RJ, Nadig R. Stage-specific predictive models for breast cancer survivability. *Int J Med Inform.* 2017; 97: 304-11. PMID: 27919388 DOI: 10.1016/j.ijmedinf.2016.11.001 [[PubMed](#)]
 38. Kaushik D, Kaur K. Application of data mining for high accuracy prediction of breast tissue biopsy results. *International Conference on Digital Information Processing, Data Mining, and Wireless Communications.* IEEE; 2016.
 39. Hamsagayathri P, Sampath P. Decision tree classifiers for classification of breast cancer. *International Journal of Current Pharmaceutical Research.* 2017; 9(2): 31-6.
 40. Pritom AI, Munshi MAR, Sabab SA, Shihab S. Predicting breast cancer recurrence using effective classification and feature selection technique. *International Conference on Computer and Information Technology.* IEEE; 2017.
 41. Percha B, Nassif H, Lipson J, Burnside E, Rubin D. Automatic classification of mammography reports by BI-RADS breast tissue composition class. *J Am Med Inform Accoc.* 2012; 19(5): 913-6. PMID: 22291166 DOI: 10.1136/amiajnl-2011-000607 [[PubMed](#)]
 42. Li Y, Chen Z. Performance evaluation of machine learning methods for breast cancer prediction. *Applied and Computational Mathematics.* 2018; 7(4): 212-6.
 43. Nourelahi M, Zamani A, Talei A, Tahmasebi S. A model to predict breast cancer survivability using logistic regression. *Middle East Journal of Cancer.* 2019; 10(2): 132-8.
 44. Lotfinezhad Afshar H, Jabbari N, Khalkhali HR, Esnaashari O. Prediction of breast cancer survival by machine learning methods: An application of multiple imputation. *Iran J Public Health.* 2021; 50(3): 598-605. PMID: 34178808 DOI: 10.18502/ijph.v50i3.5606 [[PubMed](#)]
 45. Sohrabi S, Atashi A, Dadashi A, Marashi S. A comparative study of multilayer neural network and C4.5 decision tree models for predicting the risk of breast cancer. *Archives of Breast Cancer.* 2018; 29: 11-4.
 46. Tanha J, Salarabadi H, Aznab M, Farahi A, Zoberi M. Relationship among prognostic indices of breast cancer using classification techniques. *Informatics in Medicine Unlocked.* 2020; 18: 100265.
 47. Zand HK. A comparative survey on data mining techniques for breast cancer diagnosis and prediction. *Indian Journal of Fundamental and Applied Life Sciences.* 2015; 5(s1): 4330-9.
 48. Mehri Dehnavi A, Sehhati MR, Rabbani H. Hybrid method for prediction of metastasis in breast cancer patients using gene expression signals. *J Med Signals Sens.* 2013; 3(2): 79-86. PMID: 24098861 PMCID: PMC3788197 [[PubMed](#)]