Multidisciplinary Cancer Investigation

April 2018, Volume 2, Issue 2

Applying Two Computational Classification Methods to Predict the Risk of Breast Cancer: A Comparative Study

Alireza Atashi[1,2,*], Soolmaz Sohrabi[3], Ali Dadashi[4]

[1] Informatics Department, Breast Cancer Research Center, Motamed Cancer Institute, ACECR, Tehran, Iran
[2] Department of E-health, Virtual School, Tehran University of Medical Sciences, Tehran, Iran
[3] Department of Medical Informatics, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[4] Department Of Medical Informatics, Students` Research Committee, Mashhad University of Medical Sciences, Mashhad, Iran
**Corresponding author: Alireza Atashi, Informatics Department, E-health, Virtual School, Tehran University of Medical Sciences, Dolatshahi alley, Naderi St., Tehran, Iran. E-mail: smatashi@yahoo.com*

DOI: 10.30699/acadpub.mci.2.2.8

Submitted: 1 February 2018
Revised: 27 February 2018
Accepted: 2 March 2018
e-Published: 1 April 2018

Keywords:
Decision Tree
C5.5
Neural Network
Breast Cancer
Data Mining

Abstract

Introduction: Lack of a proper method for early detection and diagnostic errors in medicine are some fundamental problems in treating cancer. Data analysis techniques may significantly help early diagnosis. The current study aimed at applying and evaluating neural networks and decision tree algorithm on breast cancer patients' data for early cancer prediction.

Methods: In the current study, data from Breast Cancer Research Cancer (BCRC), ACECR (the Academic Center for Education, Culture and Research) were used consisting of data from 4004 patients with breast cancer risk factors. Of all records, 1642 (41%) were related to malignant changes and breast cancer and 2362 (59%) were related to benign tumors. Data were analyzed by neural networks perceptrons and decision tree algorithm and divided into two parts for training (70%) and testing (30%) using Rapid Miner 5.2.

Results: For decision tree, accuracy of 81.62%, specificity of 79.80%, sensitivity of 89.49%, and for neural network, accuracy of 81.62%, specificity of 89.99%, and sensitivity of 90.80% were reported. Results showed acceptable capabilities to analyze breast cancer data for both algorithms.

Conclusions: Although both models provided good results, neural network showed better diagnosis for positive cases. Database type and analysis method influenced the results. On the other hand, information about more powerful risk factors of breast cancer can provide models with high coverage.

INTRODUCTION

Breast cancer is a disease with various aspects, which inflicts huge expenses on the individual and the society. This is a disease in which malignant cells come from breast tissue and proliferate increasingly while they pass immune system without causing any defending and aggressive reaction against it [1, 2]. This disease usually initiates as a hard mass in superior lateral region of breast and may expand gradually to the whole body [3]. Although cancer is the result of combination of risk factors, the main cause of breast cancer is not clear, however, a number of risk factors are known for breast cancer [4, 5] including genetic and racial factors, hormones, diet and obesity, radiation, menopausal after age of 50, long time use of obstructive compulsive pills (OCPs), hormone therapy, cancer family history, and alcohol consumption [5, 6]. Thus, identification and right actions for the awareness of people (empowering people) about all

risk factors effective in breast cancer can help in its early detection or prevention.

Medical researchers are interested in statistical methods to develop prognostic models in various scientific fields such as medical informatics. Medical prediction models help physicians to overcome health care problems and decrease medical errors [7]. Since classification of medical issues is inherently non-linear, developing and improving a comprehensive model among data and independent variables, by statistical models, are not precise. Furthermore, conventional statistical techniques are not suitable to analyze large data sets [8]. Data analysis and its techniques, if used properly, can be more efficient in this regard. Since the application of modern technologies and software knowledge in medicine increased during the last two decades, studies show that early diagnosis has a significant role in decreasing cancer mortalities [9]. Therefore, application of data analysis techniques on breast cancer data sets and extraction of useful results to improve accuracy in medical diagnosis are crucial [8].

One of these new techniques is (medical) data mining, which can be defined as the non-trivial extraction of implicit previously unknown and potentially useful information or pattern about (medical) data [10, 11]. Using this technology, the risk of cancer can be predicted, which may play a pivotal role in the diagnosis process and an effective preventive strategy or play a role in cancer screening [12].

Several classification models are proposed over the years such as artificial neural networks (ANNs), statistical models, decision trees, and genetic algorithm [13, 14]. ANN is a set of connected input/output units in which each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights to be able to predict the correct class label of the input tuples [14]. These algorithms can be used to perform nonlinear statistical modeling and provide a new alternative to logistic regression. Neural networks offer a number of advantages such as requiring less formal statistical training, ability to detect all possible interactions between predictor variables, and the availability of multiple training algorithms [15].

Decision trees are powerful and popular both for classification and prediction. They are also useful to explore data to gain insight into the relationships of a large number of candidate input variables to a target variable. The model can be designed in a couple of layers with which the accuracy of the model is adjusted [13, 17]. C5.0 (after C5.0) is one of the decision tree algorithms (18) used/tested for different medical databases successfully or non-successfully. Application of this algorithm for medical purposes, especially in cancer, is still questionable [18, 19].

Regarding the importance of breast cancer and its early diagnosis as well as understanding the effective role of predicting models of data analysis type, it seems essential to evaluate the accuracy of these techniques in application fields and various sites to identify and introduce the most efficient and effective models. Thus, the current study aimed at investigating and identifying the accuracy of two different models, artificial neural network and decision tree (C5.0) to predict the risk of breast cancer and comparison of these two models. This may also lead to a proper result if C5.0 is comparable to neural network on cancer data as a successful method of classification.

METHODS

In a retrospective study, a merged database was used. The data consisted of information related to breast cancer patients admitted to Motamed Breast Diseases Clinic in Tehran, Iran from 2000 to 2015. Each record consisted of 14 fields including information related to breast cancer risk factors added with one field clarifying the type of main tumor (malignant or benign). The selection of risk factors for breast cancer was conducted according to their importance in various sources. All of the risk factors used in breast cancer were directly related to their significance in various study communities based on statistical tests. According to their priority levels in different sources and in consultation with surgeons and pathologists, 20 risk factors were selected, which were ultimately used by the researchers according to the data that the research center provided. The 14 priority factors were used in modeling by the centers clinicians' consensus. Table 1 presents the evaluated risk factors among data bases.

To preprocess data, columns unrelated to disease risk factors or related to patients' demographic information were omitted. Then, records with more than 20% missed information and records with irrelevant information were omitted to ensure more validity. Finally, missing values were imputed and replaced based on central criteria of mean 25 adjacent with SPSS version 21. By imputation, 3994 records remained (0.2% omitted). Then by random sampling, 70% of data were considered for models training and 30% for models testing and they were designed based on this classification [11]. In order to design neural network in Rapid Miner 5.2, the number of nodes was considered 14, learning rate of 0.01-0.5, hidden layer of 1-2; the number of nodes in hidden

layer was 10 and the number of iteration was considered 200-1000. Furthermore, to design the tree in Rapid Miner 5.2, data productivity criteria, minimum branch size of 2-8, minimum leaf size of 2, minimum productivity of 0.01-0.5, and confidence of 0.25 were used; then, models were evaluated using 30% of the data.

Finally, the confusion matrix and ROC (receiver operating characteristic) diagram were used. To interpret confusion matrix about classification and diagnosis of diseases and breast cancer patients, there were four states including true positive, true negative, false positive, and false negative [11]; and three main indices including sensitivity, specificity, and accuracy in classification were used [11].

As already mentioned, data were divided into two parts after being transferred to Rapid Miner software,: 70% for training and 30% for testing models and then, two multi-layer perceptron neural network (MLP) and decision tree (C5.0) were trained based on 70% of data and then, were tested based on 30% of data and results were provided by the three criteria of accuracy, sensitivity, and specificity.

RESULTS

The database consisted of 4004 records of female patients in which 1642 (41%) were related to breast cancer and 2362 (59%) to breast benign tumors.

After testing the relationship between the input parameters and breast cancer, these variables were used to create a decision tree and a neural network model. Correct prediction rates were greater in neural network guesses compared with those of the decision tree model. As described in previous section, after training and testing the models, the three indices of sensitivity, specificity, and accuracy were reported by the software. Table 2 represents the results of evaluation of models.

Based on Table 2 it maybe concluded that there was no significant difference between specificity and accuracy indices for the two algorithms. However, neural network was notably more sensitive than decision tree to identify the malignant tumors.

Table 1: Breast Cancer Risk Factors Considered in the Study

Risk Factor	Type	Range
The Age at the Time of Diagnosis	Quantitative–Discrete	38-89
The Age of the First Menstruation	Quantitative–Discrete	11-16
Menopausal Age	Quantitative–Discrete	48-62
The Age of the First Pregnancy	Quantitative–Discrete	18-45
History of Breastfeeding	Qualitative–Classified	Yes=1, No=0
OCPs Use	Qualitative–Classified	Yes=1, No=0
History of Hormone Therapy After Menopause	Qualitative–Classified	Yes=1, No=0
History of Breast Cancer	Qualitative–Classified	Yes=1, No=0
Family History of Breast Cancer	Qualitative–Classified	Yes=1, No=0
Infertility History	Qualitative–Classified	Yes=1, No=0
Smoking	Qualitative–Classified	Yes=1, No=0
Marriage Status	Qualitative–Classified	Yes=1, No=0
Education	Qualitative–Classified	Yes=1, No=0
Bad Events of Life	Qualitative–Classified	Yes=1, No=0
Type of Disease (Malignant or Benign)	Qualitative–Classified	Malignant=1, Benign=0

Table 2: The Results of Testing Models by Sensitivity, Specificity, and Accuracy of Models

Model	Sensitivity (Optimized)	Specificity (Optimized)	Accuracy
Multi-layer Perceptron Neural Network	90.80%	89.99%	81.62%
Decision Tree C5.0	89.49%	79.80%	80.01%

DISCUSSION

In the current study, a multi-layer neural network and decision tree (C5.0) were developed to diagnose breast cancer, trained and tested using data analysis algorithms based on a real database of Iranian patients. In comparison, despite the similar results in the accuracy measure and the specificity, one of the models (neural network) was significantly better in diagnosing the positive cases. Early diagnosis of breast cancer is important from different viewpoints and can enhance patients' survival. Due to the importance of risk factors in breast cancer incidence, efficacy of data analysis techniques to achieve an effective model in diagnosis and predicting diseases is undeniable [7].

Two neural network and decision tree models were used by other researchers on other breast cancer databases, the results were different from those of the current study. Especially, it was found that C5.0 as the new version for C4.5, may be used properly on the current study real breast cancer data. Also, various results may be found for the assessments with some other databases. For instance, in a study by Senturk and Kara [20] on neural network and decision tree models in Wisconsin sampling database, the accuracy of both models were more than that of the current study. The reason for this difference can be attributed to the difference of databases, methods, and missing data management. But, using a real local database used in the current study makes the results more reliable for local use. Moreover, in a study by Anand [21] to diagnose breast cancer on SEER database, C4.5 algorithm was used with a higher accuracy than that of the current study. In addition to the different applications of the algorithm, reason can be attributed to differences in databases or classifications, and data selection methods. In a study by Lakshmi et al., to evaluate efficacy of data analysis algorithms, C4.5 algorithm was used for Wisconsin sampling database. The accuracy of this model was significantly higher than that of the current study, due to the difference in the evaluation method. Therefore, differences in types of databases can cause different results in data analysis. Kiani & Atashi [10] used decision tree to model breast cancer data in early prediction of cancer recurrence. Similarities of these two studies were the application of decision tree, a real collection of patients, and similar results to evaluate models. These researchers showed 75% accuracy for decision tree model, which was lower than that of the current study. Maybe the most important cause of this difference was lower number of records and choosing dependent variables in the study by Kiani. Also, higher number of training data to a specific level increased the probability of accuracy improvement [9]. Furthermore, Tolooi et al., used C5.0 decision tree to model breast cancer data using data from the current study and obtained 95% accuracy [17].

The most important limitation of the current study was large volume of missing data. Replacing methods with data estimation were used due to independency among variables and lack of specific order in them, this approach may have affected the results. Another limitation was the application of one of the various methods in neural network and different kinds of decision tree; hence, it was not possible to evaluate the most efficient algorithm among these algorithms. The main strength of the current study was the application of a real data collection from patients with high number of records that improved system training and was relatively better than those of regional studies in this context.

According to one of the main purposes of medical data analysis, achieving the best algorithm for data description, results from models` analysis on databases are unique based on the applied method in the same study. Therefore, results are only valid for that method. On the other hand, a more complete list of risk factors can provide a model with more extensive coverage. Moreover, the type of method used, its identity for data preprocessing, replacing missing data, and the method considered for data evaluation can affect different results of the models.

Researchers can use results of the current study for their future studies on breast cancer risk factors database and provide models with higher efficacy and accuracy. It is suggested to investigate breast cancer data about diagnosis, compare separate modeling results, in decision tree, variation of iterations and investigate results with variation of neural network indices and compare more algorithms, especially support vector machine (SVM), due to its promising results in medicine.

In the current breast cancer database, although the two models showed a similar accuracy value, ANN method had a better diagnosis for positive cases. In addition, C5.0 performed properly in medical (cancer) data, particularly in positive case detection, which was comparable to neural network as a powerful successful algorithm. This method can help early diagnosis, especially in breast cancer (pre-) screening when other tests seem expensive in long-term planning. Such informatics methods are easy to apply, inexpensive, and fast return. More studies should be designed for this purpose.

ACKNOWLEDGEMENTS

The authors thank MCI editors for their kind help. Also, they would like to express their appreciation to the staff of Medical Informatics Department, Motamed Cancer Institute for their tremendous support in data analysis.

CONFLICT OF INTEREST

The authors declared no conflict of interests.

ETHICS APPROVAL

Not applicable.

REFERENCES

1. Rezvani B. The relationship betweenTRU9I polymorphism in vitamin D receptor (VDR) gene and breast cancer [dissertation]. Damghan Iran: Islamic Azad Uni. Damghan branch; 2012.
2. Setayeshi S, Akbari ME, Darghahi R, Haghighat khah HR. Breast Cancer and Technical Analysis of its Diagnostics. Tehran: Bitarafan; 2011.
3. American Cancer Society. Breast cancer facts & figures 2009-2010. [Cited 2006 Feb 11].
4. National Breast Cancer Foundation. Early Detection [Internet]. National Breast Cancer Foundation, Inc. 2012.
5. Khoury- Collado F, Bombard AT. Hereditary breast and ovarian cancer: what the primary care physician should know. Obstet Gynecol Surv. 2004;59(7):537-42. PMID:15199272
6. Claus EB, Risch N, Thompson WD. Autosomal dominant inheritance of early-onset breast cancer Implications for risk prediction. Cancer. 1994;73:643-51. https://doi.org/10.1002/1097-0142(19940201)73:3<643::AID-CNCR2820730323>3.0.CO;2-5
7. Azimian F, Tadaion-T GH, Jalali M. Breast Cancer Detection Using Data Mining Techniques. In proc. 4th Iranian Data Mining Conference 2010.
8. Hota H. Diagnosis of breast cancer using intelligent techniques. International Journal of Emerging Science and Engineering (IJESE). 2013;1(3):45-53.
9. WP Chang, DM Liou. A Comparison of Several Approaches to Missing Attribute Values in Data Mining. J Telemed Telecare. 2008.
10. Kiani B, Atashi A. A Prognostic Model Based on Data Mining Technics to Predict Breast Cancer Recurrence. Journal of Health and Biomedical Informatics 2014;1(1):26-31.
11. Larose DT. Data mining methods & models. New Jersey: John Wiley & Sons; 2006.
12. Islam MS, Akhter S, Salahuddin M, Sah JP, Rahman MR, Asaduzzaman S, Ahmed K, Mohiuddin AK, Shibly AZ. Early Prevention and Detection of Cancer Risk for Low Income Country using Data Mining Technology: Bangladesh Perspective. Biochem Physiol. 2016;5:4. https://doi.org/10.4172/2168-9652.1000e155
13. Islam MS, Akhter S, Salahuddin M, Sah JP, Rahman MR, Asaduzzaman S, Ahmed K, Mohiuddin AK, Shibly AZ. Early Prevention and Detection of Cancer Risk for Low Income Country using Data Mining Technology: Bangladesh Perspective.
14. Ahmed K, Habib MA, Jesmin T, Rahman MZ, Miah MBA. Prediction of breast cancer risk level with risk factors in perspective to bangladeshi women using data Mining. Int J Comput Appl. 2013;82(4):36-41. https://doi.org/10.5120/14107-2147
15. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol. 1996;49(11):1225-31. https://doi.org/10.1016/S0895-4356(96)00002-9
16. M. Kuhn and K. Johnson, Applied Predictive Modeling. New York: Springer; 2013. https://doi.org/10.1007/978-1-4614-6849-3
17. Toloie Ashlaqi A, PourEbrahimi A, Ebrahimi M, GhasemAhmad L. Using Data Mining Techniques for Prediction 11. Breast Cancer Recurrence. Iran J Breast Dis. 2012; 5(4):23-34.
18. Mostovich LA, Prudnikova TY, Kondratov AG, Gubanova NV, Kharchenko OA, Kutsenko OS, et al. The TCF4/β-catenin pathway and chromatin structure cooperate to regulate D-glucuronyl C5-epimerase expression in breast cancer. Epigenetics. 2012;7(8):930-9. https://doi.org/10.4161/epi.21199 PMID:22805760 PMCID:PMC3427288
19. Fernández-Lao C, Cantarero-Villanueva I, Fernández-de-las-Peñas C, Del-Moral-Ávila R, Menjón-Beltrán S, Arroyo-Morales M. Widespread mechanical pain hypersensitivity as a sign of central sensitization after breast cancer

surgery: comparison between mastectomy and lumpectomy. Pain Medicine. 2011;12(1):72-8. https://doi.org/10.1111/j.1526-4637.2010.01027.x PMID:21143767

20. Senturk ZK, Kara R. Breast cancer diagnosis via data mining: performance analysis of seven different algorithms. Computer Science & Engineering.2014;4(1):35.
21. Rajesh K, Anand S. Analysis of SEER dataset for breast cancer diagnosis using C4. 5 classification algorithm. International Journal of Advanced Research in Computer and Communication Engineering. 2012;1(2):2278-1021.
22. Lakshmi K, Krishna MV, Kumar SP. Performance comparison of data mining techniques for prediction and diagnosis of breast cancer disease survivability. Asian Journal Of Computer Science And Information Technology. 2013;3(5):81-7.