OPEN ACCESS Freely available online

ISSN: 2153-0602
Journal of Data Mining in Genomics & Proteomics

Research Article

Comparing the Efficiency of Diagnostic Models of Breast Cancer, Using Genetic Algorithm and Multi-Layer Perceptron Models

Hamid Moghaddasi, Reza Rabie, Solmaz Sohrabie*

Department of Health Information Management and Medical Informatics, Shahid Beheshti University of Medical Sciences, Tehran, Iran

ABSTRACT

Background and objective: Breast cancer has become a common cancer in women. The early diagnosis of breast cancer has beneficial effects on the life patients. Due to difficulties in the disease, data mining techniques could help to facilitate the diagnosis, the current study amid to compare the efficiency of Genetic Algorithm (GA) and Multi-Layer Perceptron (MLP) in the diagnosis of the breast cancer.

Methods: The database used in this paper is provided by Tehran university of Motamed Cancer Institute (MCI) breast cancer research center. This database included 7,625 records; there were 4,008 patients (52.4%) with breast cancers (malignant) and the remaining 3,617 patients (47.6%) without breast cancers (benign). GA and MLP models were developed using 14 fields (risk factor) of the database. The present study divided the data into 10 folds where 1 fold for testing and 9 folds for training as a way of validating the 10-fold crossover validation. Ultimately, the comparison of the models was made based on sensitivity, specificity, accuracy and ROC indicators.

Findings: Sensitivity, specificity, accuracy and ROC under curve of the MLP model were 0.815, 76.27, 79.71 and 81.24 respectively. For the GA model, the note indicators respectively reported: 0.884, 86.32, 87.67 and 88.50. There was statistical significant difference between indicators of the two models (p-value<0.0001).

Conclusion: Both models had acceptable efficiencies in diagnosing breast cancer, that GA had better efficiency. The number of breast cancer risk factors and number of database records can cause different sensitivity, specificity, accuracy and ROC indicators. More breast cancer risk factors such as mutation types could help to developing more efficient GA and ANN models.

Keywords: Diagnosis model; Breast cancer; MLP; GA

INTRODUCTION

Breast cancer is an abnormal growth outside the control of breast cells that can be benign or malignant. Because it is a common cancer among women, it is the leading cause of cancer mortality among women worldwide. Symptoms of this disease include symptoms such as breast enlargement, skin changes, discharge, nipple changes, stretching or asymmetry of the breasts, redness, hard, irregular and fixed wounds and the presence of an axillary mass; therefore, people should be aware of the symptoms. Be aware of this disease to help with early diagnosis. The exact cause of breast cancer is not exactly known. However, scientists believe that a number of risk factors are involved in the development of this cancer [1]. These factors are: Gender (especially female), old age, early menstruation, delayed first pregnancy, lack of breastfeeding, old menopause, use of birth control pills, hormone therapy, family history of breast cancer, mutations in BRCA1, BRCA2, P53 genes, alcohol consumption, breast exposure to radiation, existence of dense breasts, previous history of invasive breast cancer and ductal or lobular cancer in situ. Malignant breast masses, due to their nature (ability to regrow, invade other parts of the same organ, invade other organs of the

Correspondence to: Solmaz Sohrabei, Department of Health Information Management and Medical Informatics, Shahid Beheshti University of Medical Sciences, Tehran, Iran; E-mail: solmazsohrabee1@gmail.com

Received: 02-Nov-2020, Manuscript No. JDMGP-24-7007; **Editor assigned:** 05-Nov-2020, PreQC No. JDMGP-24-7007 (PQ); **Reviewed:** 19-Nov-2020, QC No. JDMGP-24-7007; **Revised:** 01-Aug-2024, Manuscript No. JDMGP-24-7007 (R); **Published:** 29-Aug-2024, DOI: 10.4172/2153-0602.24.15.349

Citation: Moghaddasi H, Rabie R, Sohrabie S (2024) Comparing the Efficiency of Diagnostic Models of Breast Cancer, Using Genetic Algorithm and Multi-Layer Perceptron Models. J Data Mining Genomics Proteomics. 15:349.

body) are life-threatening and there are a small number of malignant breast masses that behave similarly in screening. They have benign breast masses, so it is important to distinguish malignant breast masses from their benign type.

To diagnose the disease, physicians use diagnostic images of mammography, ultrasound and MRI and invasive biopsy, each of which has a limited sensitivity and a degree of error, which can lead to failure to diagnose the disease early or misdiagnosis. For example, in mammography, due to the limited sensitivity of this technique; 10% to 30% of cancerous masses are not detectable and remain hidden; therefore, it has been observed that patients with normal mammography have cancerous masses in the breast tissue. Visual error and misinterpretation of diagnostic images or pathology test results are other reasons for the delay in the diagnosis of this cancer, since the disease is asymptomatic in the early stage, which causes a delay in diagnosis. Also, if the mass is detected by imaging devices, it is reported by doctors due to inexperience and false self-confidence; without sampling, it was diagnosed as benign and the patient's life was endangered due to the delay in correct diagnosis; also, due to the inability to distinguish 100% benign from malignant mass in imaging, sampling is recommended to the patient which is a time consuming, painful and costly process and has a negative psychological effect on the patient, also has an error. In addition to the diagnostic problems of this disease, various sciences such as medical informatics through data mining and parametric and non-parametric modeling to diagnose and discover hidden patterns in order to diagnose, predict and treat diseases accurately help physicians and patients.

Modeling based on these methods can help diagnose malignant breast masses by increasing the sensitivity index compared to other imaging devices. Despite the use of data mining techniques to diagnose breast cancer, unnecessary sampling is reduced, reducing patients' anxiety and reducing diagnostic costs. Given that the correct diagnosis of breast cancer is a major problem in the field of medicine, these models lead to the best decision and reduce misdiagnosis [2]. Preliminary studies conducted by researchers indicate that different data mining methods have been used in the diagnosis of breast cancer, among which the multilayer perceptron neural network and genetic algorithms have been used to diagnose breast cancer. Therefore, this study was performed for the purpose of a comparative study to diagnose breast cancer based on the multilayer perceptron neural network and genetic algorithm. To determine the efficiency of these algorithms, indicators of accuracy, sensitivity, specificity and area under the ROC curve were used.

MATERIALS AND METHODS

The present study is of a fundamental analytical type. In this study, the breast cancer database of Tehran university Jihad was used, which contained the data of patients who referred to this center from 1986 to 1994. This database contained 3,736 information records. First, code columns and other columns that were not related to disease risk factors were removed from the database [3]. Records that lost 50 percent of their data were omitted on condition that three important risk factors, such as a personal history of breast cancer, a family history of breast cancer and hormone therapy, were also lost. Eventually, the number of database records reached 3,555. Records with lost data were then filled in with the median placement method in the SPSS26 tool; with this action, the number of records in the database reached 3,500 records. After this stage, the number of malignant records was 875 (25%) and the number of benign records was 2,625 (75%), which was used to balance the classes in the database using the smote technique. This technique is capable of balancing unbalanced databases. After use, the total number of records increased to 7,625; 3,996 records (52.4%) were data related to malignant breast cancer patients and 3,630 records (47.6%) were data related to benign breast cancer patients (Table 1).

Table 1: Factors risk cancer breast considered in the study.

Risk factors	Type	Range
Age diagnosis	Quantitative_ discrete	38-89
The age of 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 of history breast cancer	Qualitative _classified	0=No, 1=Yes
Infertility history	Qualitative _classified	0=No, 1=Yes
Smoking	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
Type of disease (Malignant or benign)	Qualitative _classified	Malignant=1, benign=0

Mat lab 2019 a software was used to implement the models of genetic algorithm and multilayer perceptron neural network. This software has more capabilities for data mining than its previous versions. Before implementing the models, using the cross-validation method (cross-validation 10-fold), 9 parts of the data were considered for model training and 1 part of them for model testing. It should be noted that the number of samples in each 10 parts is equal. The final results of the models were presented based on the indicators of accuracy, sensitivity, specificity and efficiency [4].

The categories in the problem of cancer diagnosis with four positive and negative categories and four numbers TP and FN, FP, TP are calculated according to the type of positive and negative categories. TP includes samples that are other than positive category samples and its algorithm correctly detected in the positive category. FP contains instances that are negative category instances and the algorithm has incorrectly detected it in the positive category. FN contains instances that are positive category instances and the algorithm has incorrectly detected it in the negative category. TN includes instances that are negative category instances, and the algorithm correctly detects it in the negative category. Accuracy (classification rate):

$$\text{Classification rate} = \frac{TN+TP}{TP+TN+FP+FN} \qquad (1)$$

Indicates the accuracy of the algorithm implemented in classifying the different categories in the problem of cancer diagnosis. This criterion is actually the percentage of correct classification of the algorithm. In other words, the classification rate is the number of samples that are correctly classified and the ratio of the number of samples that are correctly identified to the total number of samples [5]. The sensitivity rate or sensitivity level:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \qquad (2)$$

Which can be calculated for each of the available categories is intended to determine the accuracy of the classification for each category. In fact, this criterion indicates the success rate of the classification method in identifying samples related to each category. Feature index:

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

Which is calculated as the previous criterion for each of the available categories shows the percentage of reliability of the output of the classification method. In fact, it is possible for models to correctly predict the absence of the desired situation.

In order to create a genetic algorithm model, first its parameters were determined to achieve the goal of the problem; to create this model: Population size of 30 chromosomes, length of each chromosome 15 genes (number of fields in a database record was 15), selection based on roulette wheel, single point combination method and its rate was 0.5, mutation rate was considered 0.01 and the condition to stop production time was 100 generations. For the fit function (FITNESS), the sensitivity factor multiplication was used:

$$\text{Fitness}=SE \times SP$$

After designing the genetic algorithm models, training data were used to train the models and the generation with the highest fitness level was used to evaluate the test data. In neural networks, the number of generations should be selected in such a way that they are optimally trained based on the accuracy of training and test data. The number of neural network generations (epoch) in this study was 200 to 1,000, the number of hidden layers of this network was 2, weights between the input layer and the hidden layer was between (-0.01, 0.01) and learning rates tried was (0.01, 0.1, 1), the number of hidden layer nodes was less than twice the number of input nodes (10 in this study) [6]. Prior to network training, it was necessary to normalize the amount of input and output of the target; at to be within a certain range. Therefore, the values of all target inputs and outputs were placed in the range (1 and -1) using sigmoid functions (trainscg). For better network flexibility, trainscg transmission function was used. After designing the models, training data were used to train the models and the amount of accuracy, sensitivity, specificity and level below the ROC curve was calculated for all models; then, test data were used to test the models and the model that provided the highest ROC

curvature level was considered as the final model. For each model, 10 tests were performed and then the average values of the indicators were obtained.

RESULTS

In the genetic algorithm models, after achieving the highest fit in the last generation, 30 models were built; after calculating the sensitivity, specificity, accuracy and level of subscripts, the model with the highest area under the curve was selected as the final model for comparison with the multilayer neural network model. This output was achieved with a fit of 0.89 in the 100th generation (Table 2) [7].

Table 2: Final model of the multilayer perceptron neural network.

ROC	Accuracy	Specificity	Sensitivity
0.884	88.50%	87.67%	86.32%

In the multilayer neural network, 30 models were designed and the model that presented the highest level below the ROC curve was selected to compare with the genetic algorithm model. The final model of the multilayer perceptron neural network, which had a higher ROC subsurface than other models (Table 3).

Table 3: Final model of the multilayer perceptron neural network.

ROC	Accuracy	Specificity	Sensitivity
0.815	81.24%	79.71%	76.28%

This output was obtained in 2 hidden layers and 600 rounds and 0.1 learning rate (Figure 1).

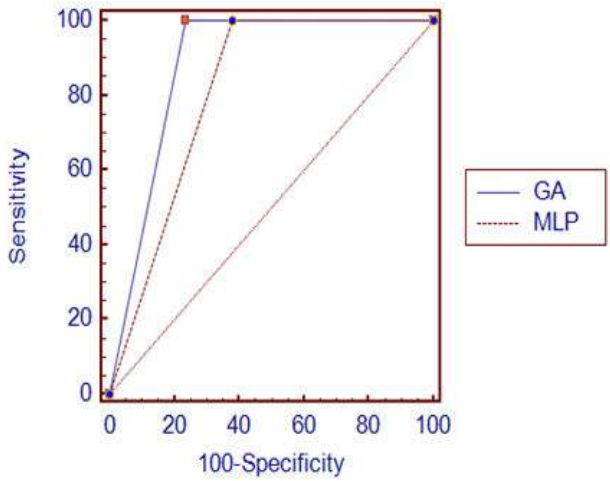

Figure 1: Comparison of 2 ROC charts.

To compare the indicators of accuracy, sensitivity and specificity of the models, McNemar statistical test was used in SPSS26 tool and p-value<0.005 was obtained. Sensitivity, accuracy and specificity indices of the two models are significantly different. MedCalc tool was used to compare the area under the two-model curve.

The results showed that there is a significant difference between the areas under the curve of the two models (efficiency); the p-value was less than 0.0001. The ROC sub-curve presents two neural network models and a genetic algorithm in the Medcalc tool [8].

DISCUSSION

Liu et al., conducted a study that used a genetic algorithm to categorize breast cancer data into a breast cancer and skin cancer risk database. The number of risk factors for breast cancer was 8, the number of chords was 286, the function of the sensitivity of the sensitivity index multiplied by the specificity index and the random assessment method was 65 to 35. The accuracy of using genetic algorithm in this study was 0.518. In order to classify breast cancer data, in this study, 8 risk factors for breast cancer with 85 records were used to implement the genetic algorithm.

Were 70 to 30, the accuracy of the genetic algorithm classification rules on this dataset was 85%. In 2000, Fidelis et al., conducted a study entitled "discovery of intelligible classification rules using genetic algorithms". In this study, the breast cancer database was used, which contained 286 records and included 9 risk factors for breast cancer [9]. The evaluation method was a random selection of 70 to 30, the accuracy of the rules obtained from the genetic algorithm on the breast cancer data was 67%. Chang et al., conducted a study in 1999, whose database had 444 records and contained three cancer risk factors and the random sampling method was 60 to 40. The results of comparing the two methods based on ROC indicated that in the test phase, the ROC subscript level for the algorithm genetics and network release was 0.83. Comparing the results of previous research with the present study, it can be concluded that the difference in the number of hazards, the size of database records, the balance of classes in a database, the type of fit function for the problem and its difference can provide indicators whose sensitivity, specificity and accuracy vary. According to the output, both models had good performance in diagnosing breast cancer and among them, the genetic algorithm has a higher efficiency than the neural network.

It was different from the present study (the main reason being the selection of the researcher and the purpose of the problem) and the method of evaluating random selection. Having data on important risk factors for breast cancer, such as gene mutations, we will have higher performance and more accurate models. Finally, in the diagnosis of breast cancer, based on the results obtained in the correct diagnosis of the number of malignant patients, the correct diagnosis of the number of benign patients and the result obtained for the correct ratio to the total diagnoses, it can be hoped that with more work and research, these models can help diagnosis of malignant breast cancer from benign, used in the health system [10].

CONCLUSION

Artificial neural networks and GA are modern disease diagnosis methods that have excited the attention of researchers in recent years. Diagnosis models can be helpful and beneficial in this regard. But it should be noted that in the field of evaluation of medical prediction models, at least two features of the model and sensitivity of the model has to be considered because considering one of them alone can be misleading. Besides, special attention should be paid to the value of false-negative. It is essential because the patient is mistakenly considered healthy and it can have hazardous consequences. Both models had acceptable efficiencies in diagnosing breast cancer, that GA had better efficiency. The number of breast cancer risk factors and number of database records can cause different sensitivity, specificity, accuracy and ROC indicators. More breast cancer risk factors such as mutation types could help to developing more efficient GA and ANN models.

ACKNOWLEDGMENT

We would like to thank Mr. Alireza Atashi, a researcher at the Jihad university breast cancer research center in Tehran, for his cooperation in providing the data needed for the study.

REFERENCES

1. Gupta S, Kumar D, Sharma A. Data mining classification techniques applied for breast cancer diagnosis and prognosis. Indian J Comput Sci Eng. 2011;2(2):188-195.
2. Li J, Fine J. On sample size for sensitivity and specificity in prospective diagnostic accuracy studies. Stat Med. 2004;23(16): 2537-2550.
3. Ahmad F, Mat Isa NA, Hussain Z, Osman MK. Intelligent medical disease diagnosis using improved hybrid genetic algorithm-multilayer perceptron network. J Med Syst. 2013;37:1-8.
4. Karakıs R, Tez M, Kilic YA, Kuru Y, Giuler I. A genetic algorithm model based on artificial neural network for prediction of the axillary lymph node status in breastcancer. Eng Appl Artif Intell. 2013;26(3):945-950.
5. Guo H, Nandi AK. Breast cancer diagnosis using genetic programming generated feature. Pat Recog. 2006;39(5):980-987.
6. Sharifi A, Alizadeh K. Comparison of the particle swarm optimization with the genetic algorithms as a training for multilayer perceptron technique to diagnose thyroid functional disease. Shiraz E-Med J. 2021;22(1).
7. Rojas MG, Olivera AC, Vidal PJ. Optimising multilayer perceptron weights and biases through a cellular genetic algorithm for medical data classification. Array. 2022;14:100173.
8. Guo Z, Xu L, Ali Asgharzadeholiaee N. A homogeneous ensemble classifier for breast cancer detection using parameters tuning of MLP neural network. App Artif Intell. 2022;36(1):2031820.
9. Turkoglu B, Kaya E. Training multi-layer perceptron with artificial algae algorithm. Eng Sci Technol Int J. 2020;23(6):1342-1350.
10. Punitha S, Stephan T, Gandomi AH. A novel breast cancer diagnosis scheme with intelligent feature and parameter selections. Comput Methods Programs Biomed. 2022;214:106432.