# Diagnosing Breast Cancer Type by Using Probabilistic Neural Network in Decision Support System

Asieh Khosravanian and Saeed Ayat

***Abstract*—In this research probabilistic neural network (PNN) was used to devise a decision support system (DSS) to diagnose the type of breast cancer in patients with this disorder. The proposed method was assessed by using a reservoir of data related to patients with breast cancer, which included 699 cases stored in UCI Machine Learning Repository. To implement the network, the applications and functions in Matlab (7.12.0) were utilized. Performance indices of this system were sensitivity, specificity, and accuracy. The performance of the proposed system based on the three indices, at the network testing phase, was found to be satisfactory (sensitivity, 1; specificity, 0.98; and accuracy, 0.99).**

***Index Terms*—Decision support system, Probabilistic neural network, breast cancer, neural network.**

## I. Introduction

Breast cancer is the most prevalent type of cancer in women and is one of the major causes of cancer morality in women aged 20-59 [1]. According to the statistics published by the Iranian Ministry of Health and Medical Education, in recent years, breast cancer has been the most serious disorder in Iranian women [1]. Timely diagnosis of cancer can increase the chances of a patient's life expectancy from 56% to 86% [2]. Breast cancer, similar to other cancers, starts with a rapid and uncontrolled outgrowth and multiplication of a part of the breast tissue, which depending on its potential harm, is divided into benign and malignant types [3].

Benign tumors represent an unnatural outgrowth but rarely lead to a patient's death; yet, some types of benign tumors, too, can increases the possibility of developing breast cancer. Furthermore, in some women with an experience of biopsy, even benign breast masses can increase the threat of breast cancer. On the other hand, malignant tumors are more serious and their timely diagnosis contributes to a successful treatment [4]. As a result, predication and diagnosis of cancer can boost the chances of treatment, decreasing the usually high costs of medical procedures for such patients.

The risks contributing to the development of breast cancer are old age, family record, consumption of alcoholic drinks and drugs, the experience of the first menstruation before the age of 12, the start of menopause after the age of 55, the time of pregnancy or non-pregnancy, overweight after menopause, and a physically inactive life [3]. Clinical diagnosis of breast cancer helps in predicting the malignant cases. A lump felt during the examination roughly give clues as to the size of tumor and its texture. The various common methods used for breast cancer diagnosis are Mammography, Biopsy, Positron Emission Tomography and Magnetic Resonance Imaging. The results obtained from these methods are used to recognise the patterns which are aiming to help the doctors for classifying the malignant and benign cases [5].

Research shows that different methods have been applied to cancer diagnosis such as Bayesian networks, neural networks, decision trees, support vector machines (SVM), and fuzzy methods [6], [7].

ANN is one of the best artificial intelligence techniques for common data mining tasks, such classification and regression problems. A lot of research showed that ANN delivered good accuracy in breast cancer diagnosis [8].

ANN is one of the best artificial intelligence techniques for common data mining tasks, such classification and regression problems. A lot of research showed that ANN delivered good accuracy in breast cancer diagnosis a neural network is a model that is designed by the way human nervous systems such as brain, that process the information. Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. Many neural network models, even biological neural networks assume many simplifications over actual biological neural networks. Such simplifications are necessary to understand the intended properties and to attempt any mathematical analysis. Even if all the properties of the neurons are known, simplification is still needed for analytical purpose [9].

Probabilistic neural networks (PNN) have been one of the most successful and widely used artificial neural networks which, according to experts' opinions, are powerful tools for identifying and classifying patterns with the maximum probability of accuracy [10].

PNNs were first introduced by Specht in 1988 and are among the most important supervised methods used for detecting and classifying patterns [11]. In recent years, in many studies, PNNs have served as prediction and detection of patterns, yielding acceptable results, compared to other methods [12]. In the present study, too, PNN is used to propose a support decision system (DSS) for diagnosing benignity or malignancy of breast cancer tumors.

## II. Related Work

Recently, expert systems have drawn the attention of many

Manuscript received April 9, 2015; revised November 26, 2015.

Asieh Khosravanian is with Computer Engineering, Payame Noor University, Iran (e-mail: dr.ayat@pnu.ac.ir).

Saeed Ayat is with the Department of Computer Engineering and Information Technology, Payame Noor University, Iran (e-mail: dr.ayat@pnu.ac.ir).

researchers for application diagnosis of diseases. These systems, using data mining techniques, can discover patterns of medical data, improve the decision-making process, affect the costs, and increase the quality of health care [13], [14].

This section consists of the review of articles on Decision Support System techniques applied in breast cancer diagnosis.

Delen *et al*. used the artificial neural network, decision tree and logistic regression for prediction models development of breast cancer by analyzing large databases was compiled of known database Wisconsin. Research results showed that decision tree algorithms was precedence over other methods to extract knowledge from available data and the results of this research was closer to reality [15].

Also, Yi and Fvyang used of support vector machine alone to discover patterns in breast cancer in available data in Wisconsin Hospital. The results show that the SVM was a good way to detect breast cancer models and the results were matched with the available evidence and real [16].

Soltani *et al*. provided a comparison among the capabilities of various neural networks such as Multilayer Perceptron (MLP), Self Organizing Map(SOM), Radial Basis Function (RBF) and Probabilistic Neural Network(PNN) which are used to classify WBC and NHBCD data. The performance of these neural network structures was investigated for breast cancer diagnosis problem. RBF and PNN were proved as the best classifiers in the training set. But the PNN gave the best classification accuracy when the test set is considered. This work showed that statistical neural networks can be effectively used for breast cancer diagnosis as by applying several neural network structures a diagnostic system was constructed that performed quite well [5].

Orlando Anunciacao, Bruno C. Gomes, Susana Vinga, Jorge Gaspar, Arlindo L.Oliveira and Jose Rueff explored the applicability of decision trees for detection of high-risk breast cancer groups over the dataset produced by Department of Genetics of faculty of Medical Sciences of Universidade Nova de Lisboa with 164 controls and 94 cases in WEKA machine learning tool. To statistically validate the association found, permutation tests were used. They found a high-risk breast cancer group composed of 13 cases and only 1 control, with a Fisher Exact Test (for validation) value of $9.7\times10\text{-}6$ and a p-value of 0.017. These results showed that it is possible to find statistically significant associations with breast cancer by deriving a decision tree and selecting the best leaf [5].

Dr. Medhat Mohamed Ahmed Abdelaal and Muhamed Farouq investgated the capability of the classification SVM with Tree Boost and Tree Forest in analyzing the DDSM dataset for the extraction of the mammographic mass features along with age that discriminates true and false cases. Here, SVM techniques show promising results for increasing diagnostic accuracy of classifying the cases witnessed by the largest area under the ROC curve comparable to values for tree boost and tree forest [5].

Wei-pin Chang, Der-Ming and Liou explored that the genetic algorithm model yielded better results than other data mining models for the analysis of the data of breast cancer patients in terms of the overall accuracy of the patient classification, the expression and complexity of the classification rule. The artificial neural network, decision tree, logistic regression, and genetic algorithm were used for the comparative studies and the accuracy and positive predictive value of each algorithm were used as the evaluation indicators. WBC database was incorporated for the data analysis followed by the 10-fold cross-validation. The results showed that the genetic algorithm described in the study was able to produce accurate results in the classification of breast cancer data and the classification rule identified was more acceptable and comprehensible [5].

K. Rajiv Gandhi, Marcus Karnan and S. Kannan in their paper constructed classification rules using the Particle Swarm Optimization Algorithm for breast cancer datasets. In this study to cope with heavy computational efforts, the problem of feature subset selection as a pre-processing step was used which learns fuzzy rules bases using GA implementing the Pittsburgh approach. It was used to produce a smaller fuzzy rule bases system with higher accuracy . The resulted datasets after feature selection were used for classification using particle swarm optimization algorithm. The rules developed were with rate of accuracy defining the underlying attributes effectively [5].

J. Padmavati performed a comparative study on WBC dataset for breast cancer prediction using RBF and MLP along with logistic regression. Logistic regression was performed using logistic regression in SPSS package and MLP and RBF were constructed using MATLAB. It was observed that neural networks took slightly higher time than logistic regression but the sensitivity and specificity of both neural network models had a better predictive power over logistic regression. When comparing RBF and MLP neural network models, it was found that RBF had good predictive capabilities and also time taken by RBF was less than MLP [5].

Chul-Heui Lee, Soen-Hak Soc and Sang-Chul Choi in their study proposed a new classification method based on the hierarchical granulation structure using the rough set theory. The hierarchical granulation structure was adopted to find the classification rules effectively. The classification rules had minimal attributes and the knowledge reduction was accomplished by using the upper and lower approximations of rough sets. A simulation was performed on WBC dataset to show the effectiveness of the proposed method. The simulation result showed that the proposed classification method generated minimal classification rules and made the analysis of information system easy [5].

Aboul Ella Hassanien, and Jafar M.H.Ali in their paper presented a rough set method for generating classification rules from a set of observed 360 samples of theWBC data. The attributes were selected, normalized and then the rough set dependency rules were generated directly from the real value attribute vector. Then the rough set reduction technique was applied to find all reducts of the data which contains the minimal subset of attributes that are associated with a class label for classification. They showed that the total number of generated rules was reduced from 472 to 30 rules after applying the proposed simplification algorithm. They also made a comparison between the obtained results of rough sets with the well known ID3 decision tree and concluded rough sets showed higher accuracy and generated more compact rules [5].

Sudhir D. Sawarkar et al applied SVM and ANN on the WBC data .The results of SVM and ANN prediction models were found comparatively more accurate than the human being. The 97% high accuracy of these prediction models can be used to take decision to avoid biopsy [5].

Sepehr M. H. Jamarani et al presented an approach for early breast cancer diagnosis by applying combination of ANN and multiwavelet based sub band image decomposition. The proposed approach was tested using the MIAS mammographic databases and images collected from local hospitals. The best performance was achieved by BiGHM2 multiwavelet with areas ranging around 0.96 under ROC curve. The proposed approach could assist the radiologists in mammogram analysis and diagnostic decision making [5].

Hybrid machine learning method was applied by Sahan in diagnosing breast cancer. The method hybridized a fuzzy-artificial immune system with k-nearest neighbour algorithm. The hybrid method delivered good accuray in Wisconsin Breast Cancer Dataset (WBCD). They believe it can also be tested in other breast cancer diagnosis problems [8].

Comprehensive view of automated diagnostic systems implementation for breast cancer detection was provided by Ubeyli [10]. It compared the performances of multilayer perceptron neural network (MLPNN), combined neural network (CNN), probabilistic neural network (PNN), recurrent neural network (RNN) and support vector machine (SVM). The aim of that works was to be a guide for a reader who wants to develop this kind of systems [8].

Afzan Adam et al. have developed a computerized breast cancer diagnosis by combining genetic algorithm and Back propagation neural network which was developed as faster classifier model to reduce the diagnose time as well as increasing the accuracy in classifying mass in breast to either benign or malignant. In these two different cleaning processes was carried out on the dataset. In Set A, it only eliminated records with missing values, while set B was trained with normal statistical cleaning process to identify any noisy or missing values. At last Set A gave 100% of highest accuracy percentage and set B gave 83.36% of accuracy. Hence the author has concluded that medical data are best kept in its original value as it gives high accuracy percentage as compared to altered data [9].

Val´erie Bourd`es et al have submitted the article by comparing artificial neural network with logistic regression. The author has compared multilayer perceptron Neural Networks (NNs) with Standard Logistic Regression (SLR) to identify key covariates impacting on mortality from cancer causes, Disease-Free Survival (DFS), and Disease Recurrence using Area Under Receiver-Operating Characteristics (AUROC) in breast cancer patients [9].

A.Punitha [15] et al have discussed the genetic algorithm and adaptive resonance theory neural network for breast cancer diagnosis using Wisconsin Breast Cancer Data (WBCD). They trained 699 samples which was taken from Fine Needle Aspirates (FNA) with 16 missing data, and 683 samples with breast tumors are used in this work of which 65% was proved to be benign and 35% malignant. The author has also compared the result of Adaptive Resonance

Theory (ART) with Radial Basis Function (RBF), Probabilistic Neural Network (PNN), Multi Layer Perceptron (MLP), in which the performance of these combined approach has not only improved the accuracy but also reduced the time taken to train the network [9].

## III. Method and Materials

In this research, a PNN was devised which, based on input variables, helps predict the type of breast cancers. The proposed system relied on the available data of 699 cases of patients with breast cancer that were stored in UCI Machine Learning Repository [17]. To implement the network, the applications and functions in Matlab were used, and 65% of the data were used for the network training phase, whereas the remaining 35% were used for network testing phase.

Nine clinical variables were considered as the network inputs: clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses.

The existing data in the reservoir were preprocessed and the rows involving lost data were eliminated. After this stage, 683 rows remained, and the data were then normalized through the linear method in such a way that numerical values could be expressed in terms of 0-1 binary sets to be used in the PNN.

### *A. Neural Network Training Phase*

At this phase, 65% of the data (444 cases) were used for training the neural network. To implement the PNN in Matlab, an input matrix was created including 9 rows (the 9 clinical variables) and 444 columns, and another matrix was created as the target matrix with 2 rows (2 types of benign and malignant tumors) and 444 columns.

The data entering the network were normalized through the linear method to represent 0-1 binary values. The target matrix included 2 classes: benign and malignant. In cases where the cancer type matched the class of the column in question, the value corresponding the row would be 1 and the other row would be 0. In the proposed PNN, only one epoch was used for training the network, which is an advantage of neural networks compared to other networks.

### *B. Neural Network Testing Phase*

At this phase, 35% of the data (239 cases), which had not been used in the training phase, were applied, as a vector, to the implemented artificial neural network in the software. Table I shows the outputs of the network in comparison with true outputs in the testing phase.

TABLE I: Prediction of Breast Cancer in PNN Testing Phase

| Classification | PNN prediction | Actual results |
|---|---|---|
| Benign class | 182 | 184 |
| Malignant class | 57 | 55 |

As Table I shows, from among 239 cases, which were used in the testing phase, 184 cases suffered from benign breast cancer and 55 ones from malignant cancer. The PNN diagnosed 182 patients to have benign cancer and 57 ones to have malignant cancer. This showed that the network made wrong diagnosis in case of 2 types which were truly benign but were wrongly represented as malignant.

Generally speaking, to investigate the degree of successfulness and applicability of diseases diagnosis and classification systems, the confusion matrix is used. The analyses in confusion matrices yield four possible results of diseases diagnosis and classification: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

The confusion matrix provides three indices which are used to assess the performance of classification:

- Sensitivity: the system's precision in diagnosing the malignant type;
- Specificity: the system's precision in diagnosing the benign type;
- Accuracy: the proportion of all cases truly diagnosed.

$$Sencitivity = \frac{TP}{TP + FN} \quad (1)$$

$$Specificity = \frac{TN}{FP + TN} \quad (2)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

For the data set used in the study, the values of sensitivity, specificity, and accuracy, were found 1, 0.98, and 0.99, respectively.

## IV. Conclusion

In this study, PNN was used to classify the type of breast cancer: benign or malignant. First, through training samples, the network was trained, and then tested through testing samples. In this research, the parameters of sensitivity, specificity, and accuracy were found by the system to be 1, 0.98, and 0.99, respectively. This implies that the network involved an acceptable level of reliability in classifying the cases.

In fact, the network implemented in this study, due to its high processing speed and good generalizability, seems to be more efficient than other artificial neural systems. In this network, the training process involves only one epoch and no other replication is needed to modify the weights. One of the reasons for the high sensitivity and specificity of the network in this study could be imputed to the normalization of the input vector and the appropriate selection of the network for the functional purposes of the research.

## References


[1] T. Fathi, S. Khbaz-Zadeh, and M. Mazloum, "Evaluation of risk factors for breast cancer in women of childbearing age in Mashhad on 2002-2003 years," *Journal of IranUniversity of Medial Siences*, vol. 4, pp. 568-577.

[2] L. Zhaohui, W. Xiaoming, G. Shengwen, and Y. Binggang, "Diagnosis of breast cancer tumor based on manifold learning and support vectormachine," *IEEE Trans. International Information and Automation*, pp. 703-707, 2008.

[3] National Center for Chronic Disease Prevention and Health Promotion, Breast Cancer Division, Breast Cancer and You Fact Sheet. (Dec. 2010). [Online]. Available: www.cdc.gov/cancer/breast/pdf/BreastCancerFS_Dec2010.pdf

[4] American Cancer Society. (2006). [Online]. Available: URL http://www.cancer.org

[5] S. Gupta, D. Kumar, and A. Sharma, "Data mining classification techniques applied for breast cancer diagnosis and prognosis," *Indian Journal of Computer Science and Engineering (IJCSE)*, vol. 2, no. 2, pp. 188-195, 2011.

[6] H. A. Abbass, "An evolutionary artificial neural networks approach for breast cancer diagnosis," *Artificial Intelligence in Medicine*, vol. 3, pp. 265-281, 2002.

[7] C. P. Utomo, A. Kardiana, and R. Yuliwulandari, "Breast cancer diagnosis using artificial neural networks with extreme learning techniques," *International Journal of Advanced Research in Artificial Intelligence*, vol. 3, no. 7, pp. 10-14, 2014.

[8] B. M. Gayathri, C. P. Sumathi, and T. Santhanam, "Breast cancer diagnosis using machine learning algorithms — A survey," *International Journal of Distributed and Parallel Systems (IJDPS)*, vol.4, no.3, pp. 105-112, 2013.

[9] L. Jiang, D. Wang, Z. Cai, and X. Yan, "Survey of improving naive Bayes for classification," *Advanced Data Mining and Applications*, pp. 134-145, 2007.

[10] P. Wasserman, *Advanced Methods in Neural Computing*, New York: Van Nostrand Reinhold, 1993.

[11] D. Specht, "Probabilistic neural networks for classification, mapping, or associative memory," *IEEE International Conference on Neural Networks*, vol. 1, pp. 525–532, 1988.

[12] H. C. Koh and G. Tan, "Data mining applications in healthcare," *J Healthcare Info Manag*, vol. 19, no. 2, pp. 64-72, 2005.

[13] Y. Chae, H. Kim, K. Tark, H. Park, and S. Ho, "Analysis of healthcare quality indicator using data mining and decision support system," *Expert System Application*, vol. 24, no. 2, pp. 167-72, 2003.

[14] D. Delen, G. Walker, and A. Kadam. "Predicting breast cancer survivability: a comparison of three data mining methods," *J. Artificial Intelligence in Medicine*, vol. 34, pp. 113-127, 2010.

[15] W. Yi and W. Fuyong, "Breast cancer diagnosis via support vector machines" in *Proc. the Twenty*.

[16] I. Kalatzis and I. Liappas, "Design and implementation of a multi-pnn structure for discriminating one-month abstinent heroin addicts from healthy controls using the p600 component of erp signals," *Pattern Recognition Letters*, vol. 26, pp. 1691–1700, 2005.

[17] UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)



**Asieh Khosravanian** was born in Shiraz, Iran. She received her master degree certificate in computer engineering from Payame Noor University. Now she teaches at Payame Noor University and University of Applied Science and Technology. Her research interests include soft computing, neural network and medical expert system under supervision of her counselor professo.

**Saeed Ayat** was born in Najafabad, Iran. Currently, he is an associate professor in Department of Computer Engineering and Information Technology at Payame Noor University. He received his PhD degree in computer engineering from Sharif University of Technology, in 2006. His research interests include speech processing, signal processing, wavelet and its applications, information technology and its applications and fuzzy logic and its applications.