Prediction of Breast Cancer using Artificial Neural Networks

Mihir Borkar
Computer Department
DJSCE
Mumbai, India

Abhinav Garg
Computer Department
DJSCE
Mumbai, India

Prof. Khushali Deulkar
Computer Department
DJSCE
Mumbai, India

Abstract - **In this study, we have developed an artificial neural network that will help determine if patients have breast cancer or not. Whether patients have a cancerous tumour can be determined by analysing certain attributes of the cells. We use these attributes to screen benign tumours from those which could be malignant. Data obtained from 699 patients has been used to train the network. The results have been accordingly investigated using ROC analysis. The analysis performed proves to be successful. Hence an artificial neural network can be used to diagnose breast cancer in a rapid manner without causing a great deal of trouble to patients.**

Keywords— Artificial neural network; breast cancer; breast cancer prediction; diagnosis; diagnostic aid

I. INTRODUCTION

Breast cancer is a major health concern that is plaguing the modern world. American Cancer Society estimates that in 2015 about 231,840 new cases of breast cancer will be diagnosed in women. Around 40,290 women are expected to die from breast cancer this year [11].

Currently, breast cancer can be diagnosed using Mammograms, Breast Ultrasound and MRI of the breast. These techniques are expensive and can also pose health risks to the patients with benign tumours. In view of this, an alternative solution is proposed that is safer, and can also be used as a primitive differentiating tool to separate completely benign patients from those that may or may not have breast cancer. Hence, we have developed an artificial neural network (ANN) that takes cell-based attributes as input and detects if the tumour is benign or not.

ANN is a mathematical learning model deployed on a computer that is designed in the same way the biological neural network of the brain functions. The ANN designed, learns how to predict an outcome from a set of attributes. It gives the answer in a simple "yes" or "no" format. The algorithm learns to predict from the training samples, not with any particular programming logic written into code.

Artificial neural networks are currently widely used in the fields of function approximation, regression analysis, finance and robotics. In information technology, a neural network is a system of programs and data structures that approximates the operation of the human brain. A neural network usually involves a large number of processors operating in parallel, each with its own small sphere of knowledge and access to data in its local memory [7]. Artificial neural networks have been successfully used in medical science. Especially for cardiology, pharmacoepidemiology and medical data mining. It is also used in EKG patterns, lesion detection in SPECT images and texture analysis in ultrasound [10].

II. METHODS

A. Artificial Neural Network

The ANN was implemented using a three-layer back propagation architecture as in [6]. The three layers are namely an input, hidden and output layer. Each layer consists of certain elements called neurons. Each neuron is connected to the other via links or weights, input layer consisted of nine nodes representing the nine different characteristics utilised. The output layer consists of two nodes which classifies the output as either benign or malignant, where benign tumour is represented by the digit "1" and malignant is represented by the digit "2". The network was initially trained by random initialization. After the training is over, the weights remain unchanged. When a set of input attributes is presented to the network, it classifies the tumour as benign or malignant. This classification is done on the basis of the knowledge gained by the network during the training phase, where it was exposed to a data set of 699 entries.

The network was trained using back propagation algorithm. In this algorithm initially we perform a "forward pass" in which we calculate all the activations throughout the network including the output value of hypothesis term h (x) then for each node j in layer l it computer the error term $_j$ that measures how responsible the node was for any errors in our output .

For the output node we can directly measure the $_j$ for that target node by comparing it directly with the required output of the training set. We then use that $_j$ in the output layer to compute the error terms for the l-1 layer where l is the total number of layers present.

B. Case Selection

The training data used was features obtained from a digitized image of a fine needle aspirate (FNA) [1] of a breast mass as in [1] . The data consisted of a total of 699 instances of patients from the University Of Wisconsin Hospitals, Madison which was obtained as a total of summing up of 8 groups which were verified over the period from January 1989 to January 1992. The total set of 699 patients consisted of 458 "benign" (65.5 %) and 241 "malignant" (34.5%) instances as in [2] . There were no patients with missing data and the output for malignant or benign tumours was recorded as in [4].

TABLE 1   FEATURES DEFINTION

Sr. No.	Attribute Name	Attribute Definition
1	Clump Thickness	Malignant cells are multi-layered while benign cells are layered in a single band.
2	Uniformity of cell size	Cancer cells tend to vary in size - That is why this feature is valuable in saying whether the cells are cancerous or not.
3	Uniformity of cell shape	Cancer cells tend to vary in shape - That is why this feature is valuable in saying whether the cells are cancerous or not.
4	Marginal adhesion	Normal blood cells adhere to each other .In cancer cells this ability diminishes. So loss of adhesion is a sign of malignancy.
5	Single epithelial cell size	Epithelial cells that are enlarged significantly may turn out to be malignant.
6	Bare nuclei	Term for nuclei not surrounded by cytoplasm. Those are typically seen in harmless tumors.
7	Bland Chromatin	In malignant cells the chromatin is found to be rough in texture, while the nucleus of non-cancerous cells have a unvarying texture.
8	Normal nucleoli	In malignant cells the nucleoli are largely visible, and are sometimes present in a greater quantity. In contrast, benign cells have a microscopic nucleoli.
9	Mitosis	Mitosis is a process that is inherently managed by the genes inside every cell. If this control fails, a single cell can multiply to make new cells that also lose control and are cancerous .

C. Constructing the Neural Network

A three layered back propagation based architecture of a neural network was proposed as outlined in Fig 1.The input layer consisted of 9 nodes, that specify certain attributes of cell size critical in detecting tumours. The hidden layer consists of 15 nodes. The number of nodes in the hidden layer was varied between 10 and 25, the most accurate prediction was observed with 15 hidden layer nodes. A series of trials had to be made to achieve optimization of results.
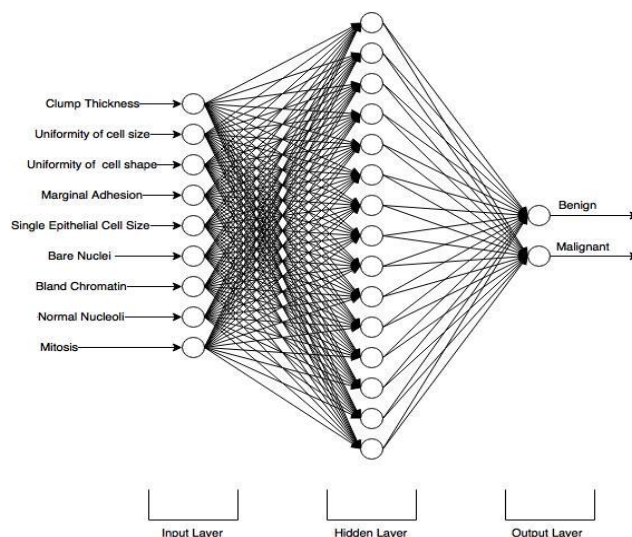

Fig. 1.   Artificial Neural Network Architecture

Learning rate = 0.2
Number of training samples = 599
Number of testing samples = 100
Training iterations = 50
Every set of attributes was presented to the network on each iteration. The network was implemented on a notebook computer running on the Mac operating system using MATLAB as the neural network training software.

III. RESULTS

A total of 599 instances were used to train the network which trained the network with an accuracy of 96.49 %. After these samples were utilized to train the network a testing set of 100 samples from the original data set that consisted of 699 samples was utilized in order to test the trained
network.

An ROC (Relative operating characteristic curve) analysis [5] was done for the training and the testing data . ROC analysis helps to determine the performance of a system where there are just two possible outputs. It's a result of plotting the true positive rate against the false positive rate. The true positive rate is also known as sensitivity and the false positive rate is known as specificity. A prediction would be the best if specificity is 0 and sensitivity is 1. This would denote a successful test. Thus, the accuracy of the study can be determined by understanding these curves.
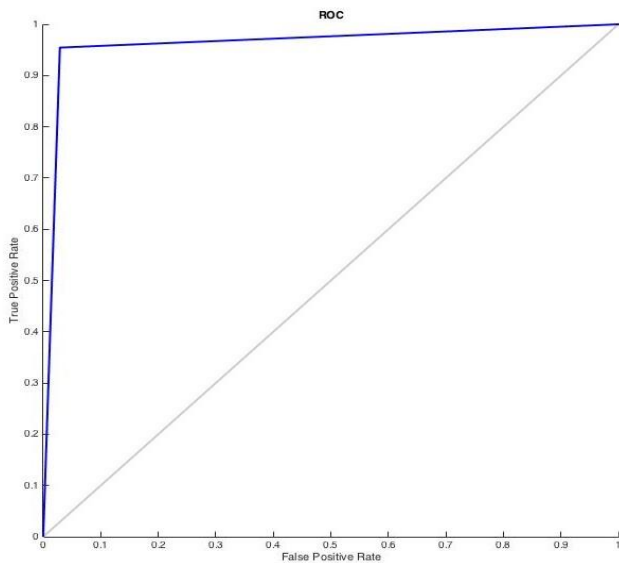
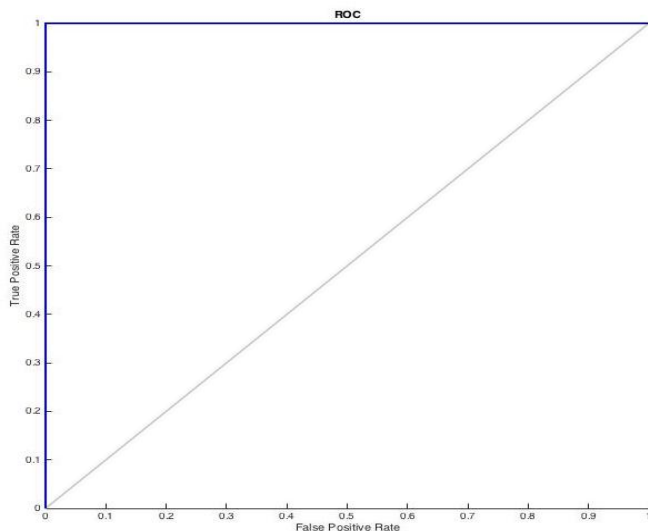Fig.2.   ROC Analysis for the training set

Fig. 3.   ROC Analysis for the testing set

For a test to be successful the area beneath the curve has to be larger. Fig 2 represents the ROC for the data that had been used during training. Fig 3 represents the data used for testing .As we can see from Fig 2 and Fig 3 the area under the ROC for the testing set is greater than the area under the ROC for the training set and is large (closer to 1). Using this analysis we can conclude that the study is successful in prediction of breast cancer.

IV. CONCLUSION AND FUTURE SCOPE

The ANN has been described solely on the basis of certain cell attributes. These attributes are obtained from the data of 699 women [3] and may not be a characteristic of the entire population.

Instead of being the sole basis for detection of malignancy, the ANN could in fact be used as a screening tool , to send to potentially more serious and probable cases of cancer for further diagnosis like Mammography or MRI. The objective of the ANN is to help the radiologist discern the definitely benign tumor's from the rest in a manner as accurate as possible.

If the number of benign patients being subjected to unnecessary and potentially harmful tests of mammography and breast MRI falls down as a result of this computer based screening then the ANN would have served as a useful tool. This also helps the patients avoid unnecessary emotional trauma involved in undergoing procedures such as MRI and biopsies.

Current research suggests that test based on Cyclin E or microarray analysis [9] can outperform the conventional method of diagnosis. This method could be clubbed with the current ANN to increase the degree of accuracy or evaluate the patient samples from a different and possibly a more inclusive standpoint.

Scientists from the University of Copenhagen [8] have devised a method to predict if women will get breast cancer within two to five years using a metabolic blood profile. Needless to say, such a system of prediction may be implemented using a similar neural network architecture, with a variation in the input nodes on the basis of the attributes of the metabolic blood profile and the hidden nodes in order to achieve maximum possible level of accuracy.

ACKNOWLEDGMENT

REFERENCES

[1]   O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.
[2]   William H. Wolberg and O.L. Mangasarian : "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A. , Volume 87, December 1990, pp 9193-9196.
[3   O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognitionVia linear programming : Theory and application to medical diagnosis", in : "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li , editors, SIAM Publications, Philadelphia 1990, pp 22-30
[4]   K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers)
[5]   Saritas, Ismail."Prediction of breast cancer using artificial neural networks."Journal of medical systems 36.5 (2012): 2901-2907.
[6]   Floyd, Carey E., et al. "Prediction of breast cancer malignancy using an artificial neural network." Cancer 74.11 (1994): 2944-2948.
[7]   http://searchnetworking.techtarget.com/definition/neural-network
[8]   http://news.ku.dk/all_news/2015/04/metabolicbloodprofile/
[9]   http://www.nature.com/nm/journal/v9/n1/full/nm0103-16.html
[10] ww.researchgate.net/publication/2315977_Appications_of_Artifici al_Neural_Networks_in_Medical_Science
[11] http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-   key-statistics.