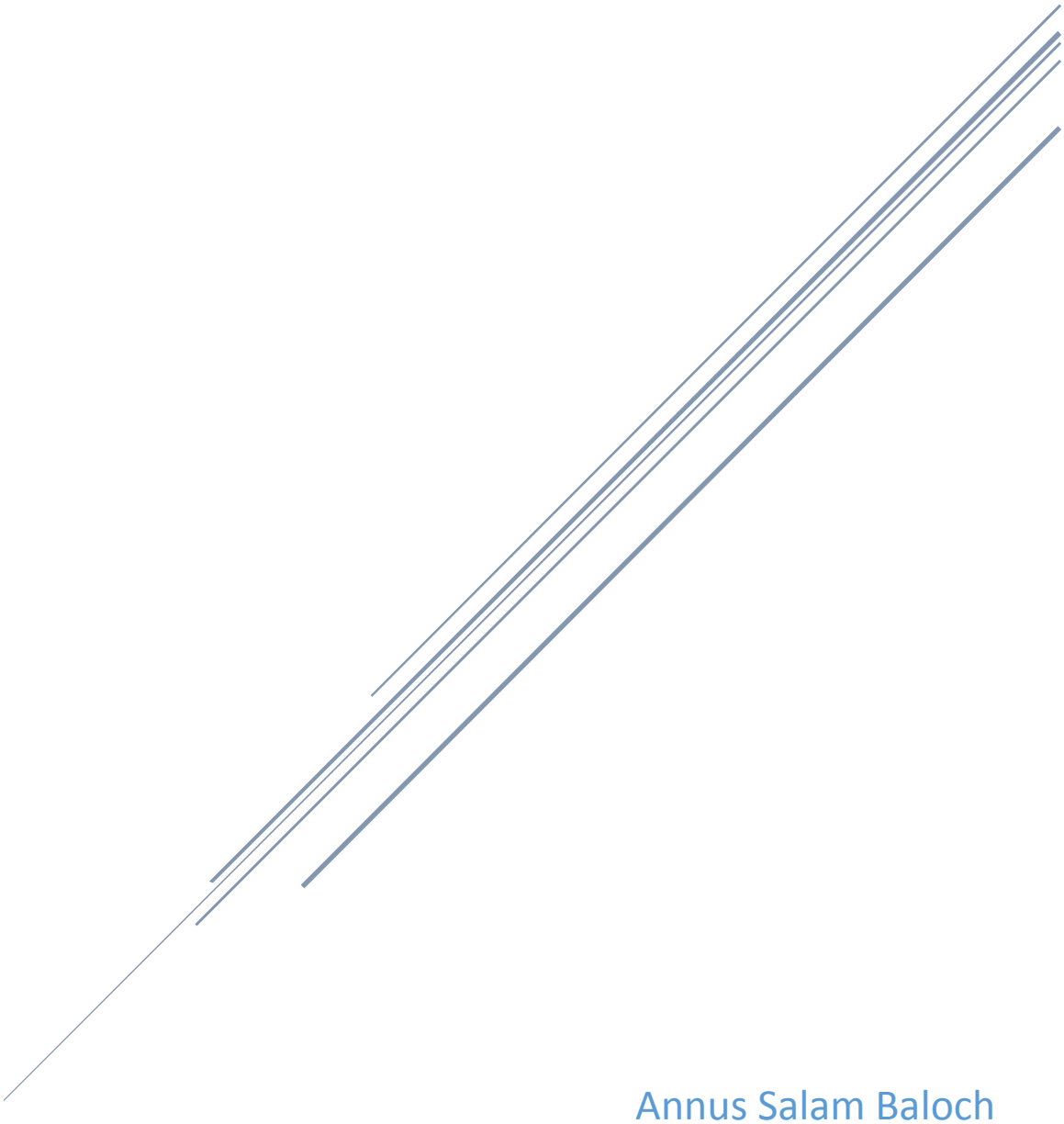
DIOGNASING THE BREAST CANCER

Using Neural Network



Annus Salam Baloch
14031204

- **Introduction:**

Breast cancer is a harmful tumor (an accumulation of malignancy cells) emerging from the cells of the breast. In spite of the fact that breast cancer mostly happens in ladies, it can likewise also influence men. According to a survey done by Medicine net, it is second highest cause of death in women. This extremely harmful disease is spreading with the passage of time. Mostly, women from rural areas are affected widely due to lack of awareness. It is the greatest threat to all the humans as it has started striking the men as well.
In order to stop this increasing danger, many ways had been discovered by scientists. A lot of methods were proposed by using which one can easily check the symptoms of breast cancer. . Another reason for using neural network in diagnosing the breast cancer is that they are faster in generating results and thus require lesser time to manipulate symptoms.

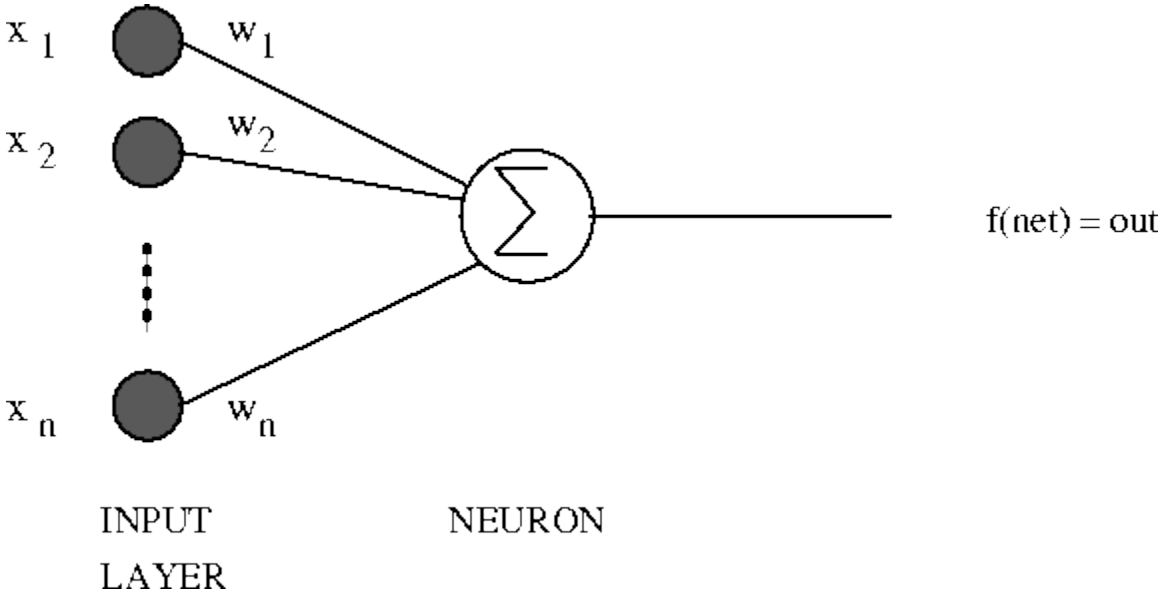
This report is based on a problem in which sample data is given and a neural network is trained on that data so that we can check either the suspect has breast cancer or not. This cover all the stages starting from designing networking to analyzing result and carrying out the experimentation.

- **Background:**

The design and architecture of neural network is inspired from human brain. Neural network research is motivated by two desires: to obtain a better understanding of the human brain, and to develop computers that can deal with abstract and poorly defined problems. The basic building block of neural networks are neurons.

Neurons:

The following figure show the structure of neuron:



Neuron is also known as the node. The neuron receives one or more input and sums them to give an output. Each input is separately weighted and the sum is passed through a nonlinear function known

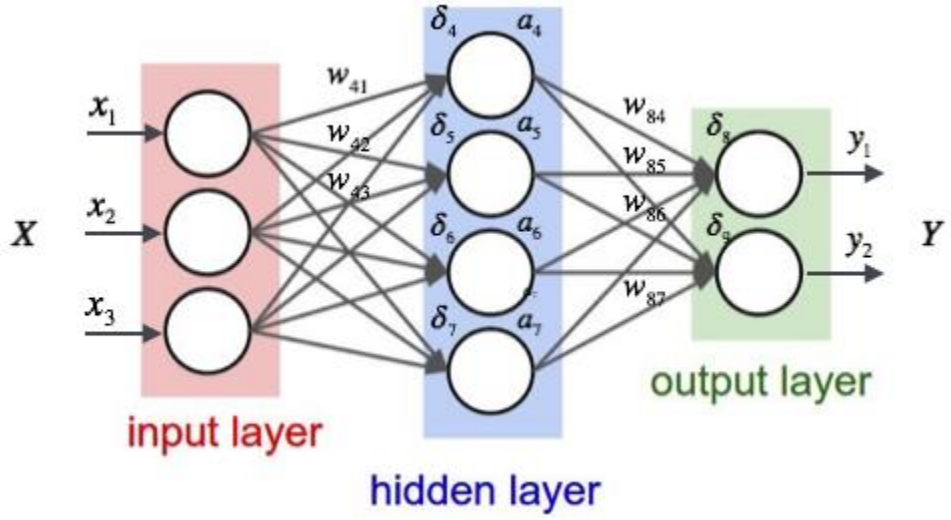
as transfer function or activation function. The activation function may take form of another nonlinear function but mostly they have sigmoid shape.

Neural Network:

It is an information processing paradigm in which the neurons are interconnected in a specific order. It contains three major layers. An input layer. An output layer, and many hidden layers. The main feature of paradigm is its architecture. The large number of neurons are connected with each other to solve a problem. This network can be trained through learning data. Following image shows the working of a neural network. There are two major types of Neural Networks:

- Feed Forward Network:
  A unidirectional flow of data from one node to another node is carried out in this type of neural network.

- Feed Back Network:
  The feedback loops are allowed, processed and used in the addressable memory to train the neural network.



Previous Work:

Classification of breast cancer has been a noticeable problem in soft computing. Previous solutions have been using the Wisconsin Diagnostic Breast Cancer (WBDC) dataset as an input for various types of neural networks. Senapati et. Al(2013) used a linear wavelet neural network in combination with 'Firefly algorithm' on the previously mentioned data set.

- **Main Part:**

This part provide the details about technical methods we followed in order to carry out the whole scenario. How the neural network was trained and how the data was gathered, processed and how the neural network was developed.

Collection of Data:

The data which we had used was taken from UCI Machine dataset repository. Data has 699 rows. The data has one key (Id) of patients. There are 9 columns in dataset which contains the results of different tests performed on the patients in order to diagnose breast cancer. It also has one attribute which holds the overall result in the form of 2 and 4 in which 2 means non maligned and 4 means maligned.

Pre Processing:

As it was a raw data so it need to be preprocessed before using it for training. The data was manually entered to Matlab. The txt file was switched to .m file. The coding was carried out in .m file. As the first column that is of id was of no use, so it was deleted from the data set. There were some values which were missing and rows have some of the data which was not complete. This kind of data need to be corrected. There were several ways to correct this kind of data. But we don't have any slandered to handle this data so it was the better option to delete such rows which had missing items in them. The dataset now contain 683 columns instead of 699 columns. The data which was unsorted initially was sorted on the basis of two and four. Now there are 444 columns with the output 2 and 239 columns with the result of 4. Now the last attribute required some sort of setting because the activation function tansig was used which has the range from -1 to 1 while the output has the values of 2 and 4. This issue was sorted out by representing 2 as -1 and 4 as 1.

The Creation of NN and its Training:

After processing all the data, the neural network was created using newff function. The different attributes used for the learning of this neural network like goal, rate, activation function etc. were executed. By using the train function, the neural network was trained over the data upon the neural architecture. Once the training had been completed, the SIM function is now used to test the network.

Post Processing:

As soon as the testing and training is done, now the accuracy of the neural network Is measured by using the following formula.

Total matched/Total testing * 100.

The accuracy was measured in percentage.

- **Experimentation and Examination:**

Some hypotheses were assumed and then these hypotheses were experimented and the results were noticed down. Some of the hypothesis are discussed below.

First Hypotheses:

The very first hypotheses is if the value of epochs, which is initially 5000, is reduced, the accuracy will be decreased. We will check the validity of hypotheses by performing number of experiments.

Value of Epochs	Number of nodes	Accuracy
5000	40	93.690
4000	40	96.367
3000	40	95.98470
2000	40	96.74
1000	40	96.3671

It is clear from the experimental result that there is no consistency in the accuracy of the result with the change of values. It shows that my hypothesis which I assumed earlier was totally wrong. If the value of epochs is reduced, the change in accuracy is random and it does not produce any kind of decrement in the accuracy. If we see it from the other way around, we can also write it as the accuracy does not depend on the value of epochs if the number of nodes are kept constant.

Second Hypothesis:

My second hypothesis is based on the assumption that if the neural network is trained on the large set of data, the accuracy will be decreased. Or in other words the result will be more accurate if the data set is small, bearing in mind that every other aspect is kept constant. We will check this hypothesis through experimentation and its results are as follow:

Number of Input data	Ratio of input data	Accuracy
80	50% Y. 50% N.	95.688
160	50% Y. 50% N.	95.602

240	50% Y. 50% N.	96.839
320	50% Y. 50% N.	95.592
400	50% Y. 50% N.	96.4664

The ratio of input data suggests that in the 80 values, 40 of them are 2, and the remaining 40 values are 4. This means 50% are malignant and 50% are not malignant. Above table suggest that the accuracy is not by affected by the size of data. In my experiment, if the data set is increased, the accuracy is still random. As every other aspect is kept constant, the randomness in accuracy clearly indicate that it is only the size of data set upon which the accuracy is changing. But it is neither increasing nor decreasing. The randomness proves that my hypothesis is also wrong.

Third hypothesis

The logsig activation function is unipolar in nature as it has range from zero to one. While the tansig function is opposite of logsig that is it is not unipolar and it ranges from -1 to 1. My third hypothesis says that if we use logsig instead of tansig activation function at hidden layer, we will get get maximum accuracy because tansig returns insignificant values for some of the cases. (By convention). The result generated are given as follow.

Activation function (Hidden layer, Output layer)	Accuracy
(Tansig, Tansig)	96.466
(logsig, tansig)	98.2332
Tansig, logsig	93.780918727915195
Logsig,Logsig	93.592

The above result show that if we use logsig function at hidden layer, and tansig function at output layer, we will get maximum accuracy. This means that my proposed hypothesis was right this time. The result was according to the prediction.

- **Conclusion**

Similarly there are many more hypothesis which can be carried out in order to increase performance and accuracy of this neural network. There are many parameters which can be focused to improve the work. Wrong hypothesis gives us the inspiration and a new direction to think. According to Claude Bernard (a French physiologist) "*Even mistaken hypothesis and theories are of use in leading to discoveries. The chemists founded chemistry by perusing chemical reaction and theories which were false.*"

- **References**

"Neural Networks." https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html. Accessed 1 Dec. 2017.

"Neural Network Architecture." http://www.dspguide.com/ch26/2.htm. Accessed 2 Dec. 2017.

"Artificial Neural Networks - UpDog." 1 Jun. 1997, https://tywdphlhh.updog.co/dHl3ZHBobGhoMDA3MDU3MTE4WA.pdf. Accessed 2 Dec. 2017.

“An Artificial Neuron - TecO." https://www.teco.edu/~albrecht/neuro/html/node16.html. Accessed 2 Dec. 2017

"What Is Breast Cancer? - American Cancer Society." 21 Sep. 2017, https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html. Accessed 2 Dec. 2017.

"Hypothesis Quotes - 227 quotes on Hypothesis Science Quotes ...." https://todayinsci.com/QuotationsCategories/H_Cat/Hypothesis-Quotations.htm. Accessed 2 Dec. 2017.