# Multilayer Perceptron Model for Breast Cancer Detection

Partha Chudal
*Department of Computer Science*
University of Nevada, Las Vegas
Las Vegas, USA
chudal@unlv.nevada.edu

Aashi Maharjan
Department of Computer Science
University of Nevada, Las Vegas
Las Vegas, USA
mahara2@unlv.nevada.edu

*Abstract*—**This paper focuses on the Multilayer Perceptron Algorithm for prediction of Nature of Breast Tumor i.e. Malignant or Benign as breast cancer has become one of the major reasons of death among the women today. Basically, detection of type of tumor is the prediction of presence of breast cancer. Timely diagnosis can help them get the gift of life. There are lot of data but information processing is still not that good leading to slow or late diagnosis. The Wisconsin Breast Cancer (Diagnostic) dataset is used in this paper for the detection of breast cancer. The data is normalized and used to train different MLP network and the best one is suggested based on accuracy and other factors. The accuracy of up to 97.66% is achieved.**

***Keywords*—*Multilayer Perceptron, Backpropagation, breast cancer, accuracy, cross-validation, ROC curve, sigmoid function***

## I. Introduction

Artificial Neural Network (ANN) is not an algorithm rather it is a paradigm for information processing. It tries to simulate the way human biological nervous system mainly the brain function to process the information and make the computer learn something. Generally, in ANN signal between any two nodes is a real number and the output of any node is computed by some non-linear function. Note that the activation function of ANN is not linear as this will result in a large linear regression model. Caution, ANN might sometimes perform incorrectly as it learns itself by example to solve the problem. Fig 1 shows an example of a neural network with one hidden layer.

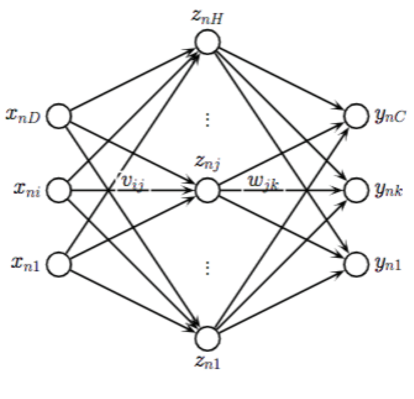

Fig. 1. A neural network with one hidden layer[1]

The backpropagation is an algorithm widely used to train ANN. Researchers became more interested in ANN after te backpropagation algorithm was published. Before backpropagation we did not have the mechanism to transfer the error back to the hidden layer which resulted in ANN to be a very complex thing. In backpropagation, the error is computed at the output layer and sent backward throughout the network's layers. The computation of the gradient vector of the Negative Log Likelihood by chain rule of calculus results in the backpropagation algorithm [1]. Steps involved in backpropagation algorithm are as follows:

a. Forward pass is done to compute pre-synaptic and post-synaptic hidden layers.

b. Error signals are computed for the output layer.

c. Pass error backward in order to compute error signals for the hidden layers.

d. Compute the gradient

e. Update the weights i.e gradient descent.

Multilayer Perceptron (MLP) is a feedforward ANN that utilizes the backpropagation algorithm. Feedforward ANN is the network where the interconnection of nodes does not form a cycle. According to Cybenko's Theorem [2], MLPs are Universal Function Approximators and hence mathematical models can be built by regression analysis using MLPs. The fact that classification is a special case of regression where the output variable is categorical leads us to the conclusion that MLPs are good for classification. MLPs have the capability to model the non-linear dataset as it uses some non-linear activation function. This helps us to solve more complex problems. According to [1], if we consider one hidden and one output layer then we can write the overall model as

$$x_n \xrightarrow{V} a_n \xrightarrow{g} z_n \xrightarrow{W} b_n \xrightarrow{h} \hat{y}_n$$

where

$x_n$ = input vector for the $n^{th}$ data instance

V = matrix of weights from input layer to hidden layer

$a_n$ = pre-synaptic hidden layer = V. $x_n$

g = some non-linear activation function at hidden layer

$z_n$ = post-synaptic hidden layer = g($a_n$)

W = matrix of weights from hidden layer to the output layer

$b_n$ = pre-synaptic output layer = W. $z_n$

h = some non-linear activation function at output layer

$\hat{y}_n$ = post-synaptic output layer = h($b_n$)

This model can be generalized for any number of hidden layers. All we need to be careful with is large number of calculations and weight matrix update. Murphy has not included bias term in his book but yes bias term can also be included while building the model.

According to the World Cancer Research Fund, there were over 2 million new patients diagnosed with breast cancer. A person is said to have breast cancer if she develops a malignant tumor in the breast cells. A lump or any benign tumors are not regarded as cancer. The tumors that cannot invade the neighbouring tissues or spreadover by metastasizing are the benign tumors whereas the metastatic tumors are known as malignant [3]. Benign tumors are not life threatning until it is in the brain but any malignant tumors can be diagnosed as a terminal disease. If we can distinguish between the benign or malignant breast tumor then we wont just be diagnosing early but also saving lot of persons time and money. Breast cancer is one of the major reasons for the death of women of age 35 – 55. However, this can be reduced by proper and timely diagnosis. We have worked on Wisconsin Breast Cancer (Diagnostic) Dataset from the UCI Machine Learning Repository. Some detail about the dataset is mentioned in section 2.

## II. Review of Literature

### A. About the Dataset

The Wisconsin Breast Cancer Diagnostic Dataset has 32 attributes including one binary class for classification. The original dataset has 76 attributes but most of the research uses the diagnostic dataset with relevant 32 attributes. The attributes of the dataset and their meaning are summarized in Table I. The dataset has 569 instances which are of different patients. According to the data donors these describe the characterstics of the cell nuclei present in the digitized image of a fine needle aspirate of a breast mass. There are 10 real-valued features are computed for each cell nusleus. The mean, standard error and worst of these features were computed for each image which geve 30 features. Data donors claim that the data contains no any missing values. Its target variable diagnosis is binary class (B= Benign and M = Malignant). Class distribution is as follows

B: 357 instances

M: 212 instances

TABLE I. ATTRIBUTES AND THEIR MEANING

| Attribute | Meaning |
|---|---|
| id | Id of patient |
| radius_mean | Mean of average of distances from center to points on the perimeter |
| Texture_mean | Mean of standard deviation of gray-scale values |
| Perimeter_mean | Mean of perimeter |
| Area_mean | Mean of area |
| Smoothness_mean | Mean of smoothnes (local variation in radius lengths) |
| Compactness_mean | Mean of compactness (perimeter ^ 2 / area-1) |
| Concavity_mean | Mean of concavity (severity of concave portions of the contour) |
| Concave points_mean | Mean of concave points (number of concave portions of the contour) |
| Symmetry_mean | Mean of symmetry |
| Fractal_dimension_mean | Mean of fractal dimension ( “ coastline approximation” – 1) |
| Radius_se | Standard Error of average of distances from center to points on the perimeter |
| Texture_se | Standard Error of standard deviation of gray-scale values |
| Perimeter_se | Standard Error of perimeter |
| Area_se | Standard Error of area |
| Smoothness_se | Standard Error of smoothnes (local variation in radius lengths) |
| Compactness_se | Standard Error of compactness (perimeter ^ 2 / area-1) |
| Concavity_se | Standard Error of concavity (severity of concave portions of the contour) |
| Concave points_se | Standard Error of concave points (number of concave portions of the contour) |
| Symmetry_se | Standard Error of symmetry |
| Fractal_dimension_se | Standard Error of fractal dimension ( “coastline approximation” – 1) |
| Radius_worst | Mean of 3 largest average of distances from center to points on the perimeter |
| Texture_worst | Mean of 3 largest standard deviation of gray-scale values |
| Perimeter_worst | Mean 3 largest of perimeter |
| Area_worst | Mean 3 largest of area |
| Smoothness_worst | Mean of 3 largest smoothnes (local variation in radius lengths) |
| Compactness_worst | Mean of 3 largest compactness (perimeter ^ 2 / area-1) |
| Concavity_worst | Mean of 3 largest concavity (severity of concave portions of the contour) |
| Concave points_worst | Mean of 3 largest concave points (number of concave portions of the contour) |
| Symmetry_worst | Mean of 3 largest symmetry |
| Fractal_dimension_worst | Mean of 3 largest fractal dimension ( “coastline approximation” – 1) |
| Diagnosis | Class (B = Benign, M= Malignant) |

### B. Related Works

Many researchers have worked to produce a better result on the dataset using various data mining techniques and have reached an accuracy of up to 95.96%. G. Ravi Kumar et al. [4] have shown SVM gave 94.5% test accuracy and MLP had 91.5% test accuracy. D. Lavanya and Dr. K. Usha Rani [5] have achieved 95.96% accuracy using Ensemble Decision Tree Classifier.

K. Sivakami [6] has reached 91% using DT (ID3) + SVM to predict Breast Cancer which however was done on the Wisconsin Breast Cancer Dataset with 11 attributes. Zehra Karapinar Senturk and Resul Kara [7] in their paper have said, among various algorithms MLP and SVM gave 96.16% and 96.49% accuracy respectively which work is also however done on the same dataset that K. Sivakami [6] worked on. Murat Karabatak and M. Cevdet Ince [8] have used Association Rule Mining for dimension reduction and neural network for classification and gained accuracy of 95.6%. This work was also done in the Wisconsin Breast cancer Original dataset and association rule reduced dimension from 10 to 4. Chintan Shah and Anjali G. Jivani [9] used WEKA on Wisconsin Breast Cancer dataset to compare the performance of Decision tree (Ramndom Forest), Bayesian Network and K-Nearest Neighbor algorithms and obtained the accuracy of 95.99%, 95.99% and 94.99% respectively.

Several researches are going on in the field to detect the breas cancer in an efficient way. J A Baker et al. [10] used Breast Imaging Recording and Data System of American College of Radiology to train neural network with 206 cases and concluded that ANN could improve the prediction accuracy of biopsy to 61% from 35% where 35% accuracy was of the radiologist. Dursun Delen et al. [11] used decision tree (C5) and ANN for prediction of breast cancer survivability on the dataset with more than 200,000 instacnces and reported accuracy of 93.6% and 91.2% respectively. Shelly Gupta et al. [12] have presented the overview of the research being carried out for diagnosis and prognosis of the dreast cancer.

## III. Methodology

### A. Data Collection and Preprocessing

We focused on MLP to predict the presence of Breast Cancer in the dataset. We Collected the dataset from UCI Machine Learning Repository. The link to the dataset is here: <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)>. We then did the following before training the model.

a. Looked at the dataset attributes and removed the id as it is not at all relevant to the mining process. Id in the dataset is the patient’s id.

b. As duplicate instances in the dataset will degrade the performance of the model when it is to be generalized, we looked for any duplicates and found none.

c. Checked for any missing values even after the donors have mentioned the dataset has no missing values and found they were correct. Data imputation would be required if we had found any missing values. Data imputation is the process of replacing the missing data with the substituted values. We should try not to run our experiments on data with missing values as this can reduce the efficiency of the network. Missing values can also increase the bias and can make the analysis of data more difficult and tiring.

d. Normalized the attributes using the equation (a). Doing so helped us to improve the performance of the network. This brought all the data values in the range [0,1]. This can also be refered to as rescaling data so it has the value between 0 and 1 inclusive. Normalization, if not done then the network might not get stable but normalization guarantees the stable convergence of weight and biases.

$$x_{new} = \frac{x\text{-}x_{min}}{x_{max}\text{-}x_{min}} \qquad (a)$$

### B. Model Building

As we were interested in building MLP model which could best estimate the type of breast tumor (Benign or Malignant) we started with a simple configuration of MLP. We then modified it until we got the best result. We used the sigmoid function as the activation function at all the nodes of the network. We started with the following model.

- Hidden layers: 16
- Learning rate: 0.3
- Momentum: 0.2
- Activation function: Sigmoid shown in equation (b) and the nature is illustrated in figure 3

$$sigm(\eta) = \frac{1}{1 + e^{-\eta}} = \frac{e^{\eta}}{e^{\eta} + 1} \qquad (b)$$

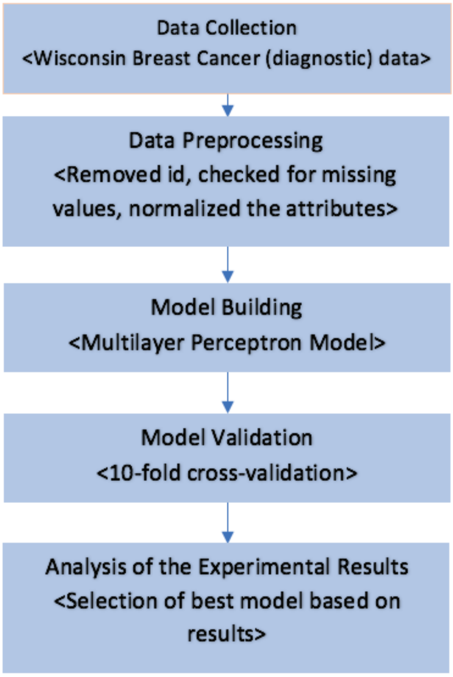
Data Collection
<Wisconsin Breast Cancer (diagnostic) data>
Data Preprocessing
<Removed id, checked for missing values, normalized the attributes>
Model Building
<Multilayer Perceptron Model>
Model Validation
<10-fold cross-validation>
Analysis of the Experimental Results
<Selection of best model based on results>

Fig. 2. Overview of the methodology of experiments of this paper.

This model gave us the accuracy of 96.66%. Then we modified the MLP network by changing the number of hidden layers and the nuber of nodes. Changing the learning rate did not show any good result as it could increase the accuracy by approximately 0.01 at the cost of high CPU cycles i.e. time.

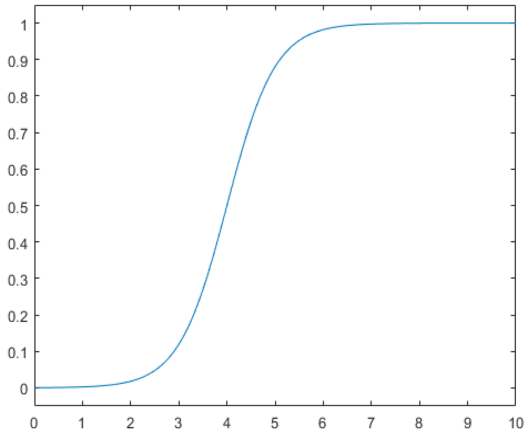

Fig. 3. Example of Sigmoid function.

After several modification of the network we finally got ended up with two models with accuracy of 97.66% and 97.31%. Details about this result is discussed in section IV.

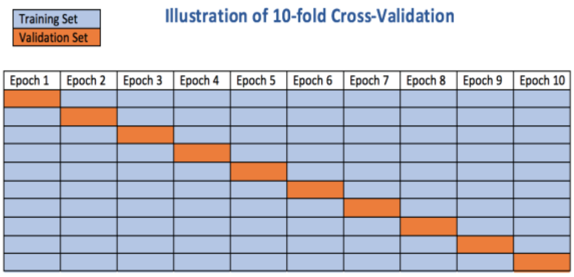

Fig. 4. Illustration of 10-fold Cross-Validation technique

## C. Validation

There are several ways to check the validity of the model. We could use a separate training and test set of simply split the dataset into training and test set or the cross fold validation. We checked the validity of the models by means of the Cross -Validation Technique.

Kohav [13] says cross-validation, also referred to as out-of-sample testing is a technique of validating a model for assessing the generalization of the results of the statistical analysis to an independent dataset. The cross-validation procedure has a parameter k which is the number of groups that the given dataset is to be split into. So the method is now referred to as K-fold cross-validation and K is replace by its value i.e. if k = 10 then it is referred to as 10-fold cross-validation. Figure 4 illustrates the 10-fold Cross-validation technique. The general procedure of this validation technique is as follows:

1. Randomly shuffle the dataset
2. Split the dataset into k groups
3. For each group
   a) Hold that group for testing purpose
   b) Use remaining data to train the model
   c) Evaluate the trained model on the test group
   d) Keep the evaluation score and discard the model
4. Use the sample of model evaluation scores to summarize the skill of the trained model

We used 10-fold Cross-Validation instead of a simple train-test split, as simple train-test split might not validate the model in a generalized manner resulting in misinterpretation which can be dangerous regarding the domain that we were working on i.e. Breast Cancer Diagnosis. This can be explained with one simple example. Consider the test split was last 20% of the data set and the M class is contained only in those 20% of the data then the trained model wont have any instance from class M while being trained and the validation would not be correct. This might result in the model which looks good but is not properly generalized for the unseen dataset.

## D. Evaluation Parameters

We checked for the different parameters of the model and also plotted the Reciever Operating Characterstic (ROC) curve which is presented in section IV. List pf parameters we calculated are following:

- Accuracy: Correctly classified instance

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- Kappa Statistic: Takes into account the possibility of correct classification occuring by chance

$$\kappa = 1\text{-} \frac{1\text{-}p_0}{1\text{-}\,p_e}$$

$p_0$ *is the relative observed agreement among raters and*

$p_e$ *is the hypothetical probability of chance agreement*

- True Positive Rate / Recall: Proportion of actual positive classified correctly

$$TPR = \frac{TP}{TP + FN}$$

- False Positive Rate: Proportion of negative events wrongly classified

$$FPR = \frac{FP}{FP + TN}$$

- Precision: How many classified items are relevant

$$precision = \frac{TP}{TP + FP}$$

- F-Measure: Harmonic mean of precision and recall

$$F\text{-}Measure = \frac{2*recall*precision}{recall + precision}$$

Area under ROC curve: In RoC curve TPR is plotted in function of FPR for different cut-off pints. Higher the Area under ROC curve better is the performance. Something above 0.5 is better than random guessing and below 0.5 is not even as good as random guessing. If area is 0.5 then the performance is exactly as the random guessing.

## IV. Experimental Results

We trained the MLP with one configuration and modified that MLP’s configuration several times to get new MLPs. We then selected the MLP with the best performance. best one. Configuration of MLP and their respective accuracy are shown in Table II. We used learning rate of 0.3 and momentum of 0.2 for all the models.

TABLE II. SUMMARY OF VARIOUS NETWORK ACCURACY

| SN | Hidden Layers | Epoch | Accuracy % |
|---|---|---|---|
| 1 | 16 | 500 | 96.66 |
| 2 | 16 | 1000 | 96.13 |
| 3 | 30 | 1000 | 97.66 |
| 4 | 15,15 (2 layers with 15 nodes each) | 500 | 96.66 |
| 5 | 30 | 500 | 96.13 |
| 6 | 15,15 (2 layers with 15 nodes each) | 1000 | 97.36 |

TABLE III. DETAILED EXPERIMENTAL RESULT OF PROPOSED MODEL

| Properties | Values |
|---|---|
| Hidden Layers | 15,15 (2 layers with 15 nodes each) |
| Learning Rate | 0.3 |
| Momentum | 0.2 |
| Activation Function at all nodes | Sigmoid |
| Epoch | 1000 |
| Validation Technique | 10 Fold Cross Validation |
| Accuracy | 97.36 |
| Kappa Stastic | 0.936 |
| True Positive Rate/ Recall* | 0.97 |
| False Positive Rate* | 0.035 |
| Precision* | 0.97 |
| F-Measure* | 0.97 |
| Area Under ROC* | 0.992 |

* means the values are weighted average of two class

The model at number 3 (one having one hidden layer with 30 nodes) of Table II outperformed all other models in terms of accuracy. However, if we see the accuracy of the model at number 6 (one having 2 hidden layers with 15 nodes each) it is just less by 0.3% but when we compared the area under ROC curve found that it is better for the latter one and also the training time taken for latter is 8.9 seconds and that for first one is 9.2 seconds. As the accuracy alone doesnot determine the performance of a model we need to rethink on selecting the final model. Area under the ROC curve is higher of the model at number 6 (one having 2 hidden layers with 15 nodes each) in Table II which signifies that its prediction is better than the other if we compare it with the probability of correct prediction by random guessing. Today we have less data so the performance might look similar but when the data size increases this thing will matter. Not only the area under the ROC but the time taken to build the model is also another key parameter as slower systems are not prefered. It is true that accuracy cannot be compromised for the shake of time but here at this case we would choose the system that performs faster over the system that gives us the higher accuracy just by 0.3%. So, keeping this in mind we suggest the latter model has better performance. The detailed output and configuration of the proposed model is shown in Table III. Fig 5 shows the ROC curve of class M (Malignant) and class B (Benign) for the model we suggested.

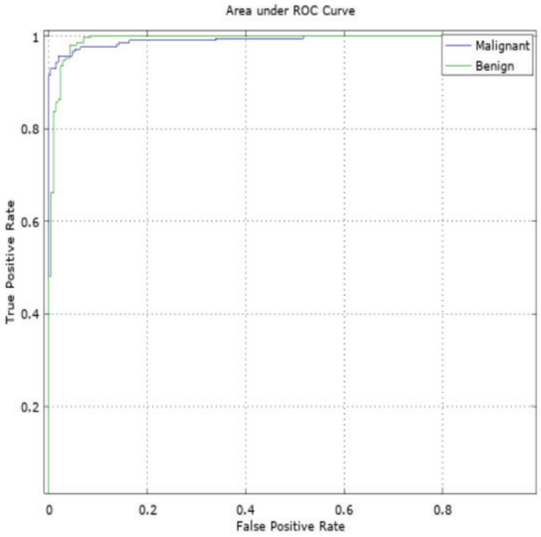

Fig. 5. Graph showing the Area under the ROC curve of the sugested model

## V. Conclusion

The best accuracy only is not always best sometimes we need to consider tradeoff between accuracy and time. Also, the higher accuracy does not always mean better performance as area under ROC might be higher for the one with lower accuracy. The performance of MLP in predicting nature of breast tumor has outperformed other algorithms published after certain modifications in the network structure. The proposed model can be used to predict whether the patient has Benign or Malignant Breast Tumor and the timely and proper treatment can save some life.

## References

[1] Kevin P. Murphy, and Francis Bach “Machine learning a probabilistic approach,” 1st ed., MIT Press, 2012.

[2] G. Cybenko, “Approximation by superpositions of a sigmoidal function,” Math Control Signals Systems, vol. 2, pp. 303-314, 1989.

[3] Kathleen Wilson, Anne Waugh, Graeme Chambers, Allison Grant, and Janet Ross, “Ross and Wilson anatomy and physiology in health and illness,” 10th ed., Edinburgh: Churchill Livingstone, 2006, pp. 53-54.

[4] G Ravi Kumar, Dr. G. A. Ramachandra, and K. Nagamani, “An efficient prediction of breast cancer

data using data mining techniques," IJIET, vol. 2, no. 4, pp. 139-144, August 2013.

[5] D. Lavanya, and Dr. K. Usha Rani, " Ensemble decision tree classifier for breast cancer data, " IJITCS, vol. 2, no. 1, pp. 17-24, February 2012.

[6] K Sivakami, " Mining big data: breast cancer prediction using DT - SVM Hybrid Model," IJSEAS, vol. 1, no. 5, pp. 418-429, August 2015.

[7] Zehra Karapinar Sentruk, and Resul Kara, "Breast cancer diagnosis via data mining: performance analysis of seven different algorithms," CSEIJ, vol. 4, no. 1, pp. 17-24, February 2014.

[8] Murat Karabatak, and Cevdet Ince, " An expert system for detection of breast cancer based on association rules and neural network, " Expert Systems with Applications, vol. 36, no. 2, pp. 3465-3469, March 2009.

[9] Chintan Shah, and Anjali G. Jivani, "Comparison of data mining classification algorithms for breast cancer, " 4th ICCCNT, July 2013.

[10] J A Baker, P J Kornguth, J Y Lo, M E Williford, and C E Floyd, Jr, " Breast cancer: prediction with artificial neural network based on BI-RADS sandardized lexicon," Radiology, vol. 196, no. 3, September 1995.

[11] Dursun Delen, Glenn Walker, and Amit Kadam, " Predicting breast cancer survivability: a comparison of three data mining methods," Artificial Intelligence in Medicine, vol. 32, no. 2, pp. 113-127, Jiune 2005.

[12] Shelly Gupta, Dharminder Kumar, and Anand Sharma, "Data mining classification techniques applied for breast cancer diagnosis and prognosis," IJCSE, vol. 2, no. 2, pp. 188-195, April-May 2011.

[13] Ron Kohavi, " A study of cross-validation and bootstrap for accuracy estimation and model selection, " 14th International Joint Conference on Artificial Intelligence, 1995.