Performance Evaluation of Breast Cancer Diagnosis using Supervised Learning Algorithms

Shivani Jain[1], Neetu Sikarwar[2]
[1]Student, Department of Electronics, Institute of Engineering Jiwaji University, Gwalior, India
[2]Professor, Department of Electronics, Institute of Engineering Jiwaji University, Gwalior, India

***Abstract*: Breast cancer is one of the leading causes of death around the universe. It has got increased in such a way that its presence affects one in ten women. Here we develop a process for diagnosis and prediction of breast cancer using Artificial Neural Network (ANN) techniques that will assist the physicians in diagnosing the disease. Implementation is done using supervised learning algorithms such as perceptron, cascade-forward back propagation and feed- forward back propagation and thereby evaluates its performance by testing the dataset obtained from Wisconsin Breast Cancer Diagnosis (WBCD) database.**

***Keywords*: Breast Cancer, Artificial neural network, Wisconsin Breast Cancer Diagnosis, supervised Learning techniques**

1. Introduction

In today's era, early diagnosis and prediction of breast cancer is quite essential to safeguard women from the dreadful tumors. The advancements in the medical field have shown various achievements over the time, still there are circumstances that lead to failures. The reasons include lack of awareness owing to breast cancers, hesitations in approaching the physicians as a result of certain religious beliefs, lack of self-breast examinations, age factors and the list goes on. Therefore it is very much important to create caution among people especially in the rural areas through various awareness camps and providing services to them. Nowadays the government has rendered helping hands for conducting camps and rallies. Also it has declared October as the Breast cancer awareness month as a sign of support.

The 21st century has seen a transformation of technology in the healthcare industry. Computers now affect all spheres of medicine and new medical advancements based on Artificial Intelligence (AI) and Artificial Neural Networks (ANN) A neural network can be as defined as a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs. Has been created to improve efficiency and simplifying the testing and treatment processes. For this purpose medical records in the form of both images and numerical data are necessary and they have been digitized and stored in the repositories such that they are available at any time even for conducting research. Moreover these developments have led to a faster calculating power to create more intricate testing capabilities and have improved the diagnosis process.

Breast cancers are of different types based on the human body system and typically occur in post-menopausal women of age greater than 40. Primarily breast cancers are either invasive or non-invasive. Invasive breast cancers are likely to spread to other areas of the body. Non-invasive breast cancers do not have the ability to spread to other parts of the body but tumor's presence can be seen in the area. Symptoms of breast cancers include swelling of all or part of the breast, irritation of skin, abnormal breast pain, nipple pain and redness, a nipple discharge other than breast milk, a lump in the underarm area and so on.

Diagnoses of breast cancers include conducting various tests as follows:

- Mammogram
- Breast ultrasound
- Ductogram
- Magnetic Resonance Imaging (MRI) of the breasts
- Biopsy procedures

A. Motivation

The motivation behind the research reported in this paper is the results obtained from extensions of an ongoing research effort. The work reported here builds on the initial work by, first, using machine learning techniques to study and understand the accurate prediction of breast Cancer diseases and it helps physician to easily identify suggestive remedies based on the classification schemes or models.

2. Related work

Ryan Potter carried out the related work in preoperative patient classification. They have used Matlab for to classify the dataset [2]. So, far, a literature survey showed that there has been several studies on the survivability predict on problem using statistical approaches and artificial neural networks. However, we could only find a few studies related to medical diagnosis and survivability using data mining approaches like decision trees [3]. Here in this study, the latest tool Weka and SVM light is used to classify the breast cancer dataset with 10-fold cross validation method. In biomedicine, researchers try to calculate various outcomes. The aim of the study is to apply and analyze different machine-learning techniques for classification of Breast Cancer.

A. Breast cancer classification

Breast cancer happens when cells in the breast begin to grow out of control and can then invade nearby tissues or spread throughout the body [4]. Large collection of this out of control tissue called tumors. However, some tumors are not really cancer because they cannot spread or threaten someone's life. These are called benign tumors. The tumors that can spread throughout the body or invade nearby tissues are considered cancer and are called malignant tumors. The average incidence rate varies from 22 -28 per 100,000 women per year in urban settings to 6 per 100,000 women per year in rural areas.

Presently, 75,000 new cases occur in Indian women every year over the course of a lifetime, 1 in 22 women will be diagnosed with breast cancer. Early detection is your best protection. Close to 90% of breast cancer can be detected early, when they are most treatable. About 12 -13% of women develop breast cancer in their lifetime. Experts estimate that about 178,480 women will be newly diagnosed with invasive breast cancer in the United States in 2007. Another 2,030 men will be diagnosed with breast cancer during the year. Although breast cancer in men is rare, the incidence has been increasing, and men are diagnosed at a later stage than women [5]. An estimated 40,460 women and 450 men will die from breast cancer in 2007. The earlier breast cancer is diagnosed, the earlier the opportunity for treatment.

The performance criterion of supervised learning classifiers such as Naïve Bayes, SVM-RBF kernel, RBF neural networks, Decision trees (J48) and simple CART are compared, to find the best classifier in breast cancer datasets (WBC and Breast tissue). The experimental result shows that SVM-RBF kernel is more accurate than other classifiers; it scores accuracy of 96.84% in WBC and 99.00% in Breast tissue. In [6], the performance of C4.5, Naïve Bayes, Support Vector Machine (SVM) and K- Nearest Neighbor (K-NN) are compared to find the best classifier in WBC. SVM proves to be the most accurate classifier with accuracy of 96.99%. In [7], the performance of decision tree classifier (CART) with or without feature selection in breast cancer datasets Breast

Cancer, WBC and WDBC. CART achieves accuracy of 69.23% in Breast Cancer dataset without using feature selection, 94.84% in WBC dataset and 92.97% in WDBC dataset. When using CART with feature selection (Principal Components Attribute Eval), it scores accuracy of 70.63% in Breast Cancer dataset, 96.99 in WBC dataset and 92.09 in WDBC dataset. When CART is used with feature selection (Chi Squared Attribute Eval), it scores accuracy of 69.23% in Breast Cancer dataset, 94.56 in WBC dataset and 92.61 in WDBC dataset. In [8], the performance of C4.5 decision tree method obtained 94.74% accuracy by using 10-fold cross validation with WDBC dataset. In [9], the neural network classifier is used on WPBC dataset. It achieves accuracy of 70.725%. In [10], a hybrid method is proposed to Enhance the classification accuracy of WDBC dataset (95.96) with 10 fold cross validation. In [11], the performance of linear discreet analysis method obtained 96.8% accuracy with WDBC dataset. In [12], the accuracy obtained 95.06% with neuron- fuzzy techniques when using WDBC dataset in [13].

3. Previous implementations

Breast cancer spreads when the cancer grows into other parts of the body or when breast cancer cells move to other parts of the body through the blood vessels or lymph vessels. This is called metastasis. Breast cancer most commonly spreads to the regional lymph nodes. The regional lymph nodes are located under the arm, in the neck, under the chest bone, or just above the collarbone. When the cancer spreads further through the body, it most commonly reaches the bones, lungs and liver. Less often, breast cancer may spread to the brain. If cancer comes back after initial treatment, it can recur locally in the breast or regional lymph nodes. It can also recur elsewhere in the body which is called distant metastases.

A literature review showed that there have been several studies on the survival prediction problem using statistical approaches and artificial neural networks. However, we could only find a few studies related to medical diagnosis and recurrence using data mining approaches such as decision trees [5], [6]. Delen et al. used artificial neural networks, decision trees and logistic regression to develop prediction models for breast cancer survival by analyzing a large dataset, the SEER cancer incidence database [6]. Lundin et al. used ANN and logistic regression models to predict 5, 10, and 15 -year breast cancer survival. They studied 951 breast cancer patients and used tumor size, axillary nodal status, histological type, mitotic count, nuclear pleomorphism, tubule formation, tumor necrosis, and age as input variables [7]. Pendharker et al. used several data mining techniques for exploring interesting patterns in breast cancer. In this study, they showed that data mining could be a valuable tool in identifying similarities (patterns) in breast cancer cases, which can be used for diagnosis, prognosis, and treatment purposes [4]. These studies are some examples of researches that apply data mining to medical fields for prediction of diseases.

4. System implementation

A machine learning technique that uses Bayesian inference to obtain parsimonious solutions for regression and classification it has an identical functional form to the support vector machine, but provides probabilistic classification. It is actually equivalent to a Gaussian process model with covariance function:

$$k(x,x') = \sum_{j=1}^{n} \frac{1}{\alpha j} \partial(x,xj)\partial(x',xj)$$

Where φ is the kernel function (usually Gaussian), and x1,...,xN are the input vectors of the training set. Multilayer Perceptron (MLP) network is the most widely used neural network classifier. MLPs are universal approximates. MLPs are valuable tools in problems when one has little or no knowledge about the form of the relationship between input vectors and their corresponding outputs.

A. *Feed-Forward Back propagation*

A Feed-Forward network consists of a series of layers. The first layer has a connection from the network input. Each subsequent layer has a connection from the previous layer. The final layer produces the network's output. Feed-forward networks can be used for any kind of input to output mapping. Specialized versions of the feed-forward network include fitting (fitnet) and pattern recognition (patternnet) networks. Feed-forward backpropagation network is simply the application of backpropagation procedure into the feed -forward networks such that every time the output vector is presented, it is compared with the desired value and the error is computed. The error value tells us how far the network is from the desired value for a particular input and the backpropagation procedure is to minimize the sum of error for all the training samples.

The error is computed by,

Error = (desired value – actual value)^2

The syntax of Feed-Forward Backpropagation takes the following arguments:

net = feedforwardnet (hidden-Sizes, training-function)
where,
hidden-sizes – Row vector of one or more hidden layer sizes (Default = 10)
training-function – Training function (Default = 'trainlm')
The functions return a new feed-forward back propagation network.

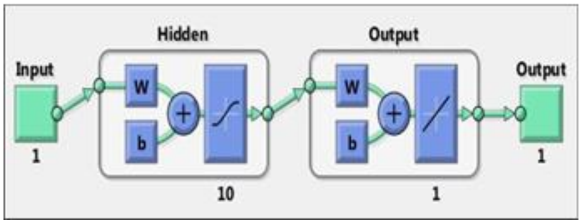

Fig. 1. Feed-Forward Network

This example shows how to use feedforward neural network to solve a simple problem.

```
[x,t] = simplefit_dataset;
net = feedforwardnet(10);
net = train(net,x,t);
view(net)
y = net(x);
perf = perform(net,y,t)
```

B. *Cascade-Forward Back propagation*

Cascade-Forward networks are similar to feed-forward networks, but include a connection from the input and every previous layer to following layers. As with feed-forward networks, two-or more layer cascade-network can learn any finite input-output relationship arbitrarily well given enough hidden neurons.

The syntax of Cascade-Forward Backpropagation takes the following arguments:

net = cascadeforwardnet (hidden-Sizes, training-function) where,
hidden-sizes – Row vector of one or more hidden layer sizes (Default = 10)
training-function – Training function (Default = 'trainlm')
The functions return a new Cascade-Forward Back propagation network.

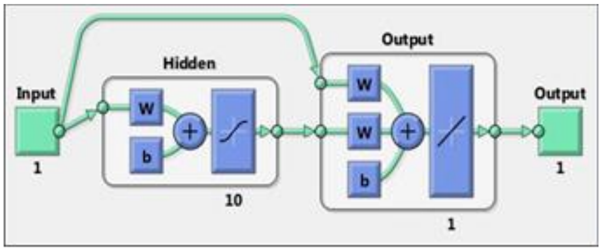

Fig. 2. Cascade-Forward Back propagation

Certain points are noteworthy while developing either a feed-forward or cascade-forward networks as follows:

- The transfer functions can be any differentiable transfer function such as tansig, logsig or purelin.
- The training function can be any of the backpropagation training functions such as trainlm, trainbfg, trainrp, traingd, traingdx etc.
- The learning function can be either of the following functions such as learngd or learngdm

Cascade-forward networks are similar to feed-forward networks, but include a connection from the input and every previous layer to following layers. As with feed-forward networks, a two-or more layer cascade-network can learn any finite input-output relationship arbitrarily well given enough hidden neurons.

Here a cascade network is created and trained on a simple fitting problem.

```
[x,t] = simplefit_dataset;
net = cascadeforwardnet(10);
net = train(net,x,t);
view(net)
y = net(x);
perf = perform(net,y,t)
```

C. *Perceptron*

Perceptron's are simple single-layer binary classifiers, which divide the input space with a linear decision boundary. Perceptron's can learn to solve a narrow range of classification problems. They were one of the first neural networks to reliably solve a given class of problem and their advantage is a simple learning rule.

The syntax of a perceptron takes the following arguments:
Perceptron(hardlimitTF, perceptronLF)
Where,
hardlimitTF- hard limit Transfer function (Default=‘hardlim’)
perceptronLF- perceptron Learning rule (Default = ‘learnp’)
The functions return a new perceptron network.

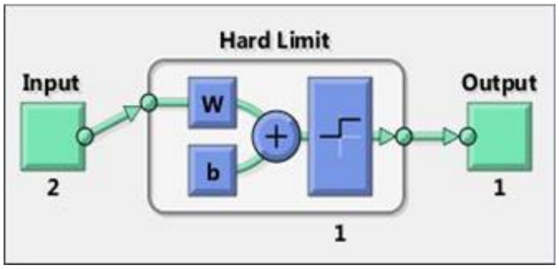

Fig. 3. Perceptron network

Input: Training Data, Testing Data
Output: Decision Value **Method:**

```
Step 1:  Load Dataset
Step 2: Classify Features (Attributes) based on class
        labels
Step 3: Estimate Candidate Support Value
        While (instances! =null)
        Do
Step 4: Support Value=Similarity between each instance in
        the attribute Find Total Error Value
Step 5: If any instance < 0
Estimate
        Decision value = Support Value/Total Error
        Repeat for all points until it will empty
End If
```

D. Classification Tree Algorithm

Algorithm: Generate a Classification from the training tuples of data partition D.
Input:

- Data partition D, which is a set of training tuples and their associated class labels;
- Attribute list, the set of can didate attributes;
- Attribute selection method, a procedure to determine the splitting criterion that “best” Partitions the data tuples into individual classes. These criterions consist of a splitting Attribute and, possibly, either a split point or splitting subset.

Output: A decision tree
Method:

```
1. Create a node N;
2. If tuples in D are all of the same class, C then
3. Return N as a leaf node labeled with the class C.
4. If attribute list is empty then
5. Return N as a leaf node labeled with the majority
   class in D
6. Apply Attribute selection method (D, attribute list) to
   find the “best” splitting criterion
7. Label node N with splitting criterion
8. If splitting attribute is discrete-valued and multiway
   splits allowed then
9. Attribute list ← attribute list − splitting attribute
10. For each outcome j of splitting criterion
11. Let Dj be the set of data tuples in D satisfying
    outcome j
12. If Dj is empty then
13. Attach a leaf labeled with the majority class in D to
    node N
14. Else attach the node returned by Generate decision
    tree (Dj, attribute list) to node N
15. End for
16. Return N
```

Table 1
Parameters

S.no	Variable name	Definition
1	Local Recurrence	Yes or No
2	Age at Diagnosis	≤ 35, 35 to 44, 44-55, 55 ≥ years old
3	Age at Menarche	≤ 12 to ≥ 12 years old
4	Age at Menopause	≤ 50 to ≥ 50 years old
5	Side	Left, right, Bilateral
6	Tumor size	≤ 2cm to ≥ 5cm years old
7	Type of chemotherapy	Adjuvant or Neo Adjuvant
8	Hormone Therapy	Tamoxifen, Raloxifen, Femara, Megance
9	Her2	Negative or Positive

5. Evaluation result

The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository is used to differentiate benign (non-cancerous) from malignant (cancerous) samples. To evaluate the effectiveness of our method, experiments on WDBC is conducted. This database was obtained from the university of Wisconsin hospital, Madison from Dr. William H. Wolberg. This is publicly available dataset in the Internet.

Table 2
Data set

Data Set	No. of attributes	No. of Instances	No. of Classes
Wisconsin Breast Cancer (WBC)	11	699	2
Wisconsin Diagnosis Breast cancer (WDBC)	11	699	2
Wisconsin Prognosis Breast Cancer (WPBC)	11	699	2

Table 1 shows a brief description of the dataset that is being considered. The dataset for this study are collected from the Wisconsin breast cancer diagnosis database available in the UCI repository. All the data that have been collected are the results of diagnosis made through Biopsy procedures. There are

569 instances and 10 attributes which also contains patient's ID number as a separate attribute. All the values are encoded with four significant digits. The ten real-valued attributes are as shown in Table 2.

6. Experimental result

The experiment was carried out in MATLAB work space (version number–MATLAB 2016a). MATLAB's Neural Network Toolbox (NN Tool) provided various features to carry out the implementation part of the three algorithms chosen as mentioned in the previous sections of this document. Initially all the input, sample and the target data have been imported into the Matlab workspace to create and train the networks using the NN Tool. The discussion is as follows:

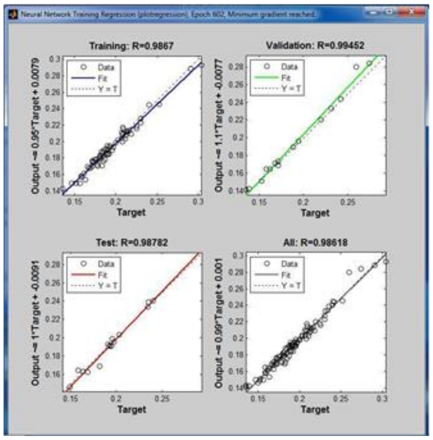

Fig. 4. Regression Plot of Feed-Forward Back propagation

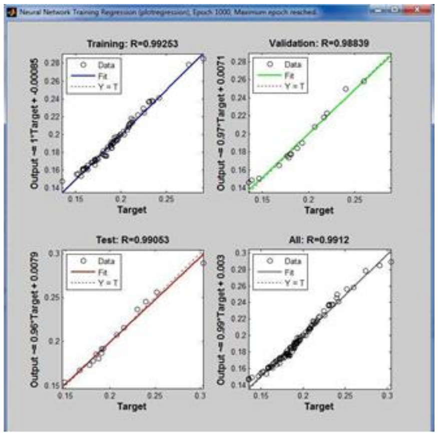

Fig. 5. Regression Plot of Cascade-Forward Back propagation

The number of neurons and layers in the network are initialized at random and the transfer function used is LOGSIG. For perceptron, network is created using HARDLIM training function and LEARNP learning function. The network is thus created and trained number of times until satisfactory performance is met. The following figure shows the working analysis of feed-forward network using the tool.

Fig. 4. and Fig. 5, shows the plot regression analysis of a feed-forward network and a cascade-forward network with respect to the data computed for the analysis of the breast cancer diagnosis.

Fig. 6, shows the Performance of a perceptron network. The Regression plot displays three divisions of output namely Training, Validation and Testing. The Regression R values measure the correlation between outputs and targets. An R value of 1 means a close relationship and 0 a random relationship.

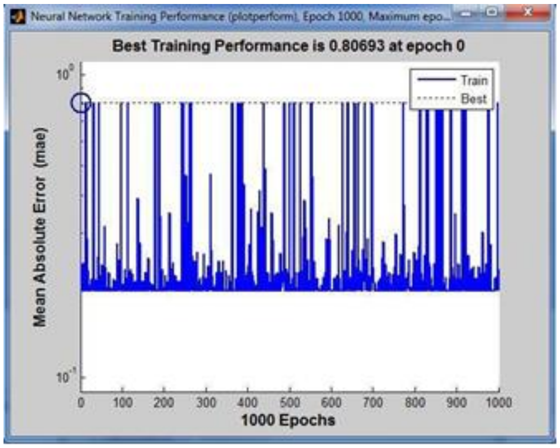

Fig. 6. Performance Plot of Perceptron Network

Based on the experimental analysis of the algorithms on the breast cancer diagnosis data, it is shown that the cascade-forward back propagation algorithm works better than the feed-forward back propagation and perceptron networks by producing the network output close to its target, supplying the value 0.99 which has a linear close relationship with that of the target data. The performance graph of the cascade-forward back propagation network is given as follows:

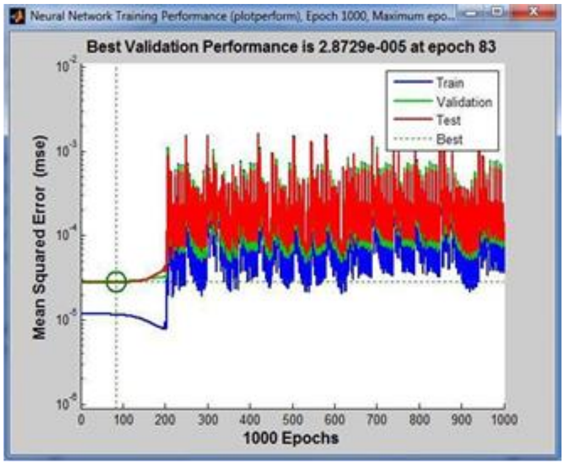

Fig. 7. Performance graph of Cascade-Forward Back propagation Network

Clustering is one of the most important unsupervised learning problems in that it deals with finding a structure in a collection of unlabeled data. In other words, clustering is the process of organizing objects into groups whose members are similar in some way.

7. Conclusion

ANNs have very efficient tools for classification, analysis and prediction of data based on the given problem and they can be successfully employed especially in breast cancer applications. They have the ability to increase the survival rate of the patients by enhancing its diagnosing capabilities using powerful machine learning and neural network algorithms. Therefore based on the results computed the physicians can decide what to do further in the treatment processes. The data produced to the networks are useful to predict which network has the better and faster performing capability. Supervised learning algorithms such as Feed-Forward Back propagation, Cascade-Forward Back propagation and Perceptron networks participated in the process and Matlab simulation results proved that the trophy was taken by the Cascade-Forward Back propagation algorithm for the diagnosis of breast cancer.

A large dataset with added complicated attributes and many instances can be selected to enhance the diagnosis process of breast cancer. Importing a very large dataset may also open paths to issues such as loading time, stability of the data values, prediction timing of the algorithm and so on. Also, experiments can be conducted by using powerful algorithms such as Radial Basis Function networks (RBF) or any hybrid combination of the algorithms that serves both in a supervised and unsupervised way. This challenge can be taken by the future researchers so that if implemented successfully it paves way for a better diagnosis algorithm for breast cancer.

References

[1] Htet Thazin Tike Thein and Khin Mo Mo Tun, “An Approach for Breast Cancer Diagnosis Classification Using Neural Network”. Advanced Computing: An International Journal (ACIJ), Volume 6, No.1, January 2018.

[2] R. DelshiHowsalya Devi and P. Deepika, “Performance Comparison of Various Clustering Techniques for Diagnosis of Breast Cancer”. IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), December 2017.

[3] R. R. Janghel, Anupam Shukla, Ritu Tiwari and Rahul Kala, “Breast Cancer Diagnosis using Artificial Neural Network Models”. IEEE International Conference on Information Sciences and Interaction Sciences (ICIS), August 2014.

[4] L. Alvarez Menendez, F. J. de Cos Juez, F. Sanchez Lasheras and J.A. Alvarez Riesgo, “Artificial Neural Networks Applied to Cancer Detection in a Breast Screening Programme”. Elsevier Mathematical and Computer Modelling, March 2018.

[5] AbeerAlzubaidi, Georgina Cosma, David Brown and A. Graham Pockley, “Breast Cancer Diagnosis using a Hybrid Genetic Algorithm for Feature Selection based on Mutual Information”. IEEE International Conference on Interactive Technologies and Games, December 2016.

[6] Chandra Prasetyo Utomo, Aan Kardiana and Rika Yuli wulandari, “Breast Cancer Diagnosis using Artificial Neural Netwroks with Extreme Learning Techniques”. International Journal of Advanced Research in Artificial Intelligence (IJARAI), Volume 3, No.7, 2014.

[7] Ritika Bewal, Aneecia Ghosh and Apoorva Chaudhary, “Detection of Breast Cancer using Neural Networks-A Review”. Journal of Clinical and Bio- Medical Sciences, December 2015.

[8] E. Venkatesan and T. Velmurugan, “Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification”. Indian Journal of Science and Technology, Volume 8, November 2015.

[9] Seema Singh, Sushmitha H, Harini J and Surabhi B R, “An Efficient Neural Network Based System for Diagnosis of Breast Cancer”. International Journal of Computer Science and Information Technologies (IJCSIT), Volume 5, No.3, 2014.

[10] Luqman Mahmood Mina and Nor Ashidi Mat Isa, “Breast Abnormality Detection in Mammograms using Artificial Neural Network”. IEEE International Conference on Computer, Communication and Control Technology (I4CT), April 2015.

[11] Dishant Mittal, Dev Gaurav and Sanjiban Sekhar Roy, “An Effective Hybridized Classifier for Breast Cancer Diagnosis”. IEEE International Conference on Advanced Intelligent Mechatronics (AIM), July 2015.

[12] Gouda I. Salama, M.B. Abdelhalim and Magdy Abd-elghany Zeid, “Breast Cancer Diagnosis on Three Different Datasets using Multi-Classifiers”. International Journal of Computer and Information Technology, Volume 1, Issue 01, September 2015.