Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm

Conference Paper · August 2016

DOI: 10.1109/DeSE.2016.8

Mohannad Alhanahnah
University of Nebraska–Lincoln

2016 9th International Conference on Developments in eSystems Engineering

Breast Cancer Detection using K-nearest Neighbor Machine Learning Algorithm

[1]Moh'd Rasoul Al-hadidi, [2]Abdulsalam Alarabeyyat, [3]Mohannad Alhanahnah
[1]Computer Engineering Department, [2] Computer Science Department
[1,2]AlBalqa' Applied University, [3]University of Kent
[1,2]Salt, Jordan, [3]Kent, UK
[1]Mohammad_hadidi@bau.edu.jo

***Abstract*— Breast cancer is very popular between females all over the world. However, detecting this cancer in its first stages helps in saving lives. Radiologists have the ability to predict if the mammography images have cancer or not, but they may miss about 15% of them. In this paper, we proposed a new method to detect the breast cancer with high accuracy. This method consists of two main parts, in the first part the image processing techniques are used to prepare the mammography images for feature and pattern extraction process. The extracted features are utilized as an input for a two types of supervised learning models, which are Back Propagation Neural Network (BPNN) model and the Logistic Regression (LR) model with comparing the result and the accuracy for the both models.**

***Keywords*— Breast Cancer, Image Processing, Mammography, Machine Learning.**

I. Introduction

Breast cancer is the second most dangerous cancer after lung cancer. Early detection can survive the people lives because it is easier to treat and prevent the tumor from expanded. Tumor is the abnormal growth of cells.

For many years, the X-ray was the only method that was used to detect the breast cancer [1, 2]. However, many another methods have been generated and proposed for detecting process that are more efficient than x-ray procedure such as, neural networks [3], artificial intelligence, and data mining.

There is a self-test every woman can do it monthly using her hand to check for any abnormal growing cells, another way is going to a specialist doctor for mammography test. Mammography is "the process of using low-dose X-rays to examine the human breast and is used as a diagnostic as well as a screening tool" [4].

Image processing techniques are used to convert the image from one to another format and for feature extraction of the images that helps to get a more useful data set. There are a large number of applications that relates to the human activities use the image processing, from remotely position explanation to biomedical image interpretation [5].

Artificial neural networks (ANNs) are one of the most common in machine learning, it simulate the human body neural network that consists of many neurons in many layers.

A group of neurons in the neural network have separate functions at the same time [6]. In the neural networks there is a learning stage in which the weights are adjusted to get the desired output and testing stage in which the neural network is tested to see its accuracy in detecting process. Generally we have three types of learning that are supervised learning that need a teacher, unsupervised learning that works without teacher, and hybrid learning that is between supervised and unsupervised learning [7].

Another type of the supervised learning algorithm is the Logistic Regression (LR), where this algorithm is used for learning process; it is a type of statistical classification model. This model is used for predicting the outputs of many probable outcomes. The mathematical equation of the logistic function is shown in the following equation.

$$f(z) = \frac{1}{1 + e^{-}z} \qquad (1)$$

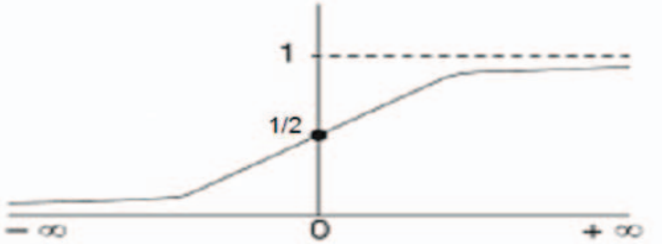

Fig. 1: Logistic Function Curve

By using the Logistic Regression (LR) model, it will estimate the logistic function to get a value between zero and one with making appealing to know the risk factor that mean a risk for disease.

The first section of this paper will be the introduction then the literature review of the breast cancer is presented where the next section will be the explanation of the

978-1-5090-5487-9/17 $31.00 © 2017 IEEE
DOI 10.1109/DeSE.2016.8

experiment and finally the conclusion and the future works are illustrated.

II. RELATED WORKS

In Breast cancer detection field there are many studies with many concepts and methods were used to be a useful methods. Many researchers present many methods and algorithms to detect the breast cancer disease; here we briefly discussed some of them.

Using Ultra-Wide Band (UWB) antenna to get microwave images of the breast with cancerous information was presented by Zhang et al. they used the gelatin-oil technology to get the experimental results to manifest the efficiency of these microwave images in breast cancer detection. By using UWB the detection process was more accurate and the signal-to-noise ratios showed that the noise was attracted around the tumor [8].

In [9], the authors proposed a method that used a gaussian pulse generator in UWB application to make the image detection of the breast cancer diagnosis occurred in fast manner. That needs to use static inverter with phase detector and NMOS pulse shaping circuit to accelerate the system detection process.

A logistic Generalized Additive Model (GAM) was proposed by Roca-Pardinas et al.. They used linear kernel smoothers with this logistic GAM,they speed up their system using many techniques. In this simulation model they used odds-ratio curves. Using this model help in early detection process for the breast cancer [10].

In another study, the Back propagation Neural Network was used for breast cancer detection and the authors compared the result with another model that used the radial basis function network, where the best result were gotten by the BPNN [11].

For breast cancer detection the researchers used the direct subtraction beam forming imager, where Jin et al. used a numirical simulation and electromagnetic in their model and they found that the model have a high resolution and robustness in detecting breast cancer process [12].

Another model that was designed for breast cancer detection was proposed by Sajjadieh et al. They used an electromagnetic model that based on Finite Difference Time Domain (FDTD). This model gives a higher accuracy in tumor detection process than another signal processing algorithms [13].

III. EXPERIMENT AND METHODOLOGY

The proposed method for breast cancer detection consists of two main parts: image processing techniques and the machine learning algorithms where applying these algorithms were done by using Matlab software. In this work we extracted 209 images for 50 patients cases who have breast cancer and the testing stage was applied on many people either they have a breast cancer or they have not.

A. Image processing

For the mammography images we applied a sequence of image processing functions to generate a utilized image to be input for the machine learning algorithms.

The first process was image cropping where a specific part of the image (margin) was deleted and the remained part was extracted. In our dataset, the images have the same size of the margins so it is easy to have a static cropping process. As shown in Fig.2.

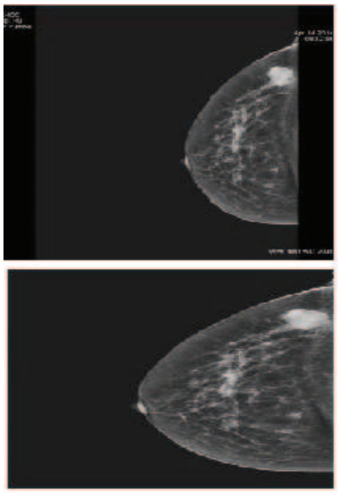

Fig.2. Cropping Process

To enhance and modify the images we used the image filtering process. Our data set have blur effects and to remove it we used Weiner filter to eliminate this noise. In figure 3 we cane note that the white lines are less than the original image.

In many works, the researchers preferred to convert the images into binary images (black and white), regarding to the useful information they can get for detection process. But, in our work, we kept the images in grey scale format because when we tried to convert them we noticed that the white dot in the images that presents the tumor was vanished from the image this means the useful information will be omitted as shown in Fig.3.

The filtered images were transformed from time domain into frequency domain using Discrete Wavelet Transformation (DWT). The output of the wavelet transform is known as wavelet decomposition that consists of four matrices: the approximation coefficients matrix, the horizontal detailed coefficient matrix, the vertical detailed coefficient matrix, and the diagonal detailed coefficient matrix. In our work, we used the last three matrices where the first matrix was not used, see Fig 4. The values in these matrices were used as input for our learning algorithm; we can observe that the shown values of these plots were distributed cross zero value.

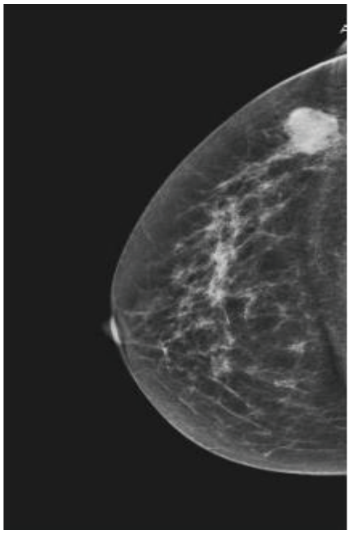

Fig. 3. Filtered Image

This distribution cross zero value was utilized in our model to extract information. After generating a new matrix depending on our algorithm with the count of zero crossing we got the input values for the learning model. The values of the algorithm outputs were 0's for the normal images and 1's for images that have tumor, so we need to normalize the data by dividing these data on the maximum value for each column of these output data.

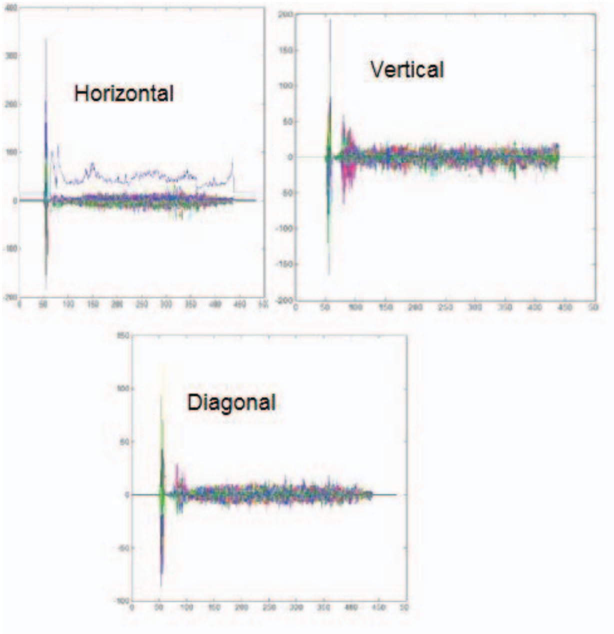

Fig.4. Plot of coefficient matrices

B. *Machine Learning Algorithm*

In our work, supervised learning algorithm has been used. Indeed, we used two types of supervised machine learning algorithms which were the Logistic Regression and the Backpropagation neural Network and we compared the results from both of them.

Logistic Regression (LR) was used for classification process for the mammography images. As any machine learning algorithm, LR needs a hypothesis and a cost function. The following equations show the hypothesis and the cost function equations.

$$h_{\emptyset}(x) = \frac{1}{1+e^{-\emptyset T_x}} \quad (2)$$

$$J(\emptyset) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^i log h_{\emptyset}(x^i) + (1-y^i) log(1-h_{\emptyset}(x^i))\right] \quad (3)$$

Where we have the values of weights of the hypothesis, x's are the input values and the y's are the output values. Now, our purpose to optimize the cost function where this is can be achieved by repeating equation 4 many times until reach the desired cost function.

$$\emptyset_j = \emptyset_j + \alpha \frac{d}{d\emptyset_j} J(\emptyset) \quad (4)$$

Many features and numbers were required to get the optimal value with utilization such as the following parameter:

- Value of A = 0.45
- Number of Iterations = 1000
- Number of Features = 750

Fig. 5 shows the cost function with different values of alpha during the training process.

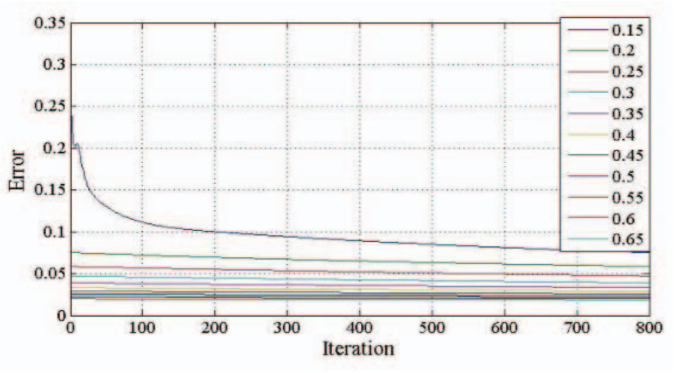

Fig. 5. Error values of the cost function

The second learning algorithm that was used in our work was the Backpropagation Neural Network (BPNN).

Our purpose focused on still finding optimization method for detection and classification. The BPNN is easy to implement and has been used widely for classification purposes. For our neural network the configuration parameters are shown in table 1. However, we could not use the same number of features as in LR model, this related to the limitation of the Matlab memory. We used 264 features with LR.

Table. 1. The Configuration Parameters of Neural Network

Parameter	***Value***
Number of Hidden Layers	1
Number of Neurons in the second layer	10
Number of Neurals in the first layer	240
Used function	Triangular
Epoch	1000

IV. Result and Discussion

In our proposed system we used 209 images that were extracted from 50 patient's cases. These images were used for training, testing, and validating processes.

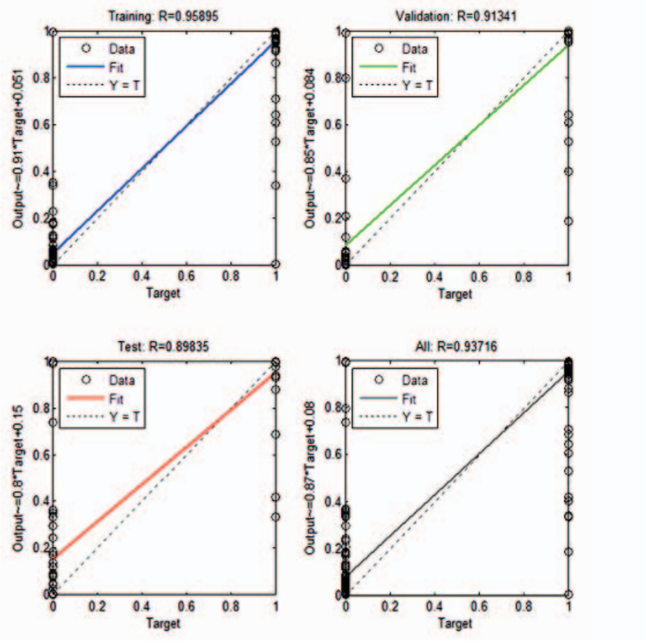

Fig.6. Regression of Neural Network

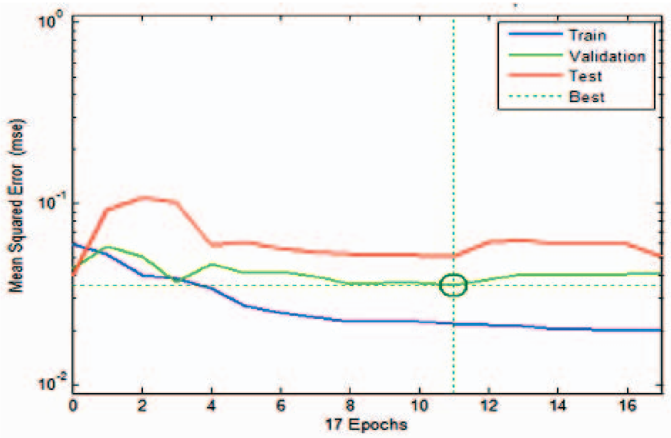

Fig. 7. MSE of Neural Network

The utilization percentage for training, testing, and validation was 70%, 20%, and 10%.

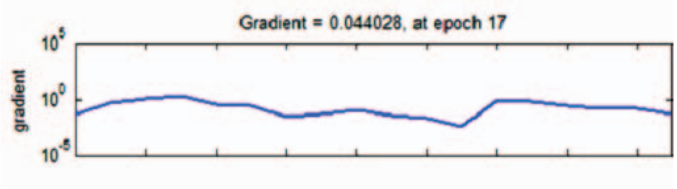

Fig. 8. Gradient of Neural Network

Each image in our dataset has resolution of 1024x1024. Our dataset consist of 96 normal images and 110 effected images, we arranged them as vector which consist 96 zeros and 110 ones, and this vector was used in LR and BPNN models.

The output matrices with our training vector that we prepared for machine learning models is shown in Fig. 6 for the BPNN where the regression value of the neural network model exceeded 93.7%.Moreover, we can see that the Mean Square Root (MSE) value is less than 0.07 where the performance of the gradient of neural network is shown in Fig.8.

V. Conclusions

In this paper, we focused on a very dangerous disease that causes death for many women over the world which is the breast cancer and we proposed a contributed method to diagnosis this disease and give information about the patient status. Our proposed model consists mainly of two parts, the first one is using image processing techniques for feature extraction where the second part was the machine learning algorithms in two types of supervised learning algorithms, LR and BPNN. We observed that, the number of features utilized in LR model was much higher than with the BPNN. However, we also have a good regression value using BPNN that exceeded 93% with only 240 features.

REFERENCES

[1] M. Brown, F. Houn, E. Sickles, and L. Kessler. "Screening mammography in community practice". Amer.J.Roentgen, vol. 165, 1995.

[2] M. Alhadidi, M. Al-Gawagzeh , B. Alsaaidah, "Solving A Mammography Problems of Breast Cancer Detection Using Artificial Neural Networks And Image Processing techniques", Indian Journal of Science and Technology, Vol.5, No.4, 2012.

[3] E. D. Ubeyli,"Implementing automated diagnostic systems for breast cancer detection", Elsevier, Expert systems with applications, vol.33, (2007).

[4] D. Kulkarni,S. Bhagyashree, G. Udupi, "Texture analysis of mammographic images",International Journal of Computer Applications, vol.5 , (2010).

[5] T. Acharya, A. Ray,"Image processing: principles and applications",Wiley-Interscience,Hoboken new jersey, ISBN 0471719986,(2005).

[6] A. Hopgood, Intelligent systems for engineers and scientists, Library of Congress Cataloging in Publication Data, (2000).

[7] H. Zhang, T. Arslan, B. Flynn,"A Single Antenna Based microwave System for Breast Cancer Detection: Experimental Results", IEEE, (2013).

[8] E. Y. K. NG, E. C. KEE,"Advanced integrated technique in breast cancer thermography", Journal of Medical Engineering and Technology, Vol. 32, No. 2,(2008).

[9] S. H. Barboza,J. A. Palacio, E. Pontes, S. Kofuji, "Fifth Derivative Gaussian Pulse Generator for UWB Breast Cancer Detection System",IEEE,(2014).

[10] J. Roca-Pardiasa,C. Cadarso-Surezb, P. Tahocesc, M. Ladod, "ssessing continuous bivariate effects among different groups through nonparametric regression models: An application to breast cancer detection", Elsevier, Computational Statistics and Data Analysis, Vol. 52,(2008).

[11] P. Pawar, D. Patil,"Breast Cancer Detection Using Neural Network Models", IEEE, International Conference on communication Systems and Network Technologies,(2013).

[12] Y. Jin, J. Moura, Y. Jiang, "Breast Cancer Detection By Time Reversal Imaging", IEEE,(2008).

[13] M. Sajjadieh, F. Foroohar, A. Asif, "Breast Cancer Detection using Time Reversal Signal Processing",IEEE, (2009).