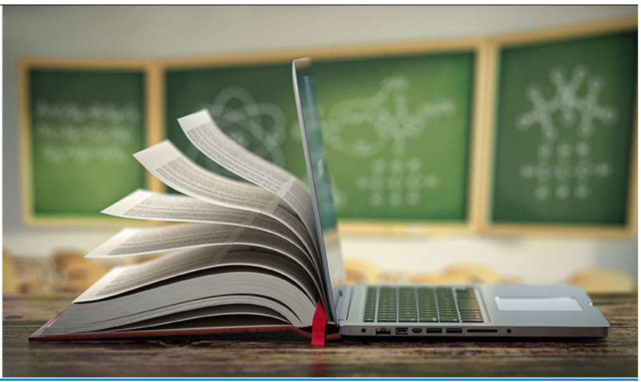
Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Reliable Early Breast Cancer Detection using Artificial Neural Network for Small Data Set

To cite this article: V. Vijayasarveswari *et al* 2021 *J. Phys.: Conf. Ser.* **1755** 012037



Reliable Early Breast Cancer Detection using Artificial Neural Network for Small Data Set

V.Vijayasarveswari[1]*, M.Jusoh[1], T.Sabapathy[1], R.A.A.Raof[2], S.Khatun[3] and I.Iszaidy[2]

[1]Advanced Communication Engineering (ACE) Centre of Excellence, Universiti Malaysia Perlis,No 15 & 17, Jalan Tiga, Pengkalan Jaya Business Centre, 01000 Kangar, Perlis, Malaysia.
[2]Embedded Network and Advanced Computing (ENAC), Universiti Malaysia Perlis.
[1,2]School of Computer and Communication Engineering, Universiti Malaysia Perlis, Kampus Alam UniMAP Pauh Putra, 02600 Arau, Perlis, Malaysia.
[3]Faculty of Electrical & Electronic Engineering, Universiti Malaysia Pahang, Pekan, Pahang.

*vijaya@unimap.edu.my

Abstract. This paper proposes a breast cancer detection module using Artificial Neural Network for small data set. The developed system consists of hardware and software. Hardware included UWB transceiver and a pair of home- made directional sensor/antenna. The software included a Graphical User Interface (GUI) and k-fold based feed-forward back propagation Neural Network module to detect the tumor existence, size and location along with soft interface between software and hardware. Forward scattering technique is used by placing two sensors diagonally opposite sides of a breast phantom. UWB pulses are transmitted from one side of phantom and received from other side, controlled by the software interface in PC environment. Firstly feed forward backpropagation neural network (FFBNN) is developed. Then, k-fold is combined with developed FFBNN for testing purpose. Four data sets are created where contains 125, 95, 65 and 30 data samples in 1st,2nd,3rd and 4th data set respectively. Collected received signals were then fed into the NN module for training, testing and validation. The process is done for all data sets separately. The system exhibits detection efficiency of tumor existence, location (x, y, z), and size were approximately 87.72%, 87.24%, 83.93% and 80.51% for 1st, 2nd, 3rd and 4th data set respectively. The proposed module is very practical with low-cost and user friendly. The developed breast cancer detection module can be used for large data samples as well as for minimum data samples.

1. Introduction

A breast cancer risk rate shows inclining trend in the developing countries. Symptoms such as changes in the breast can only be discovered at the late stage [1]. As a result, most of the detected cases lead to death. Breast cancer can affect one in 19 women within the age of 85 years in Malaysia. Based on National Cancer Society Malaysia [2], 4000 people are affected by breast cancer and 40% of them are under 50 years old. It happens due to lack of awareness of regular breast health check-up, lack of affordable self-monitoring devices and expensive breast health monitoring at the hospital. There are multiple methods to diagnose or detect the breast cancer such as mammography, magnetic resonance imaging (MRI) and ultrasound. However, they are unable to detect at an early stage, costly and impose a negative impact on health after or during the diagnostic process [3][4]. It is very important to detect breast cancer in the early stage [5][6]. As a health friendly method, Ultra wide-band (UWB) based breast cancer detection is a potential candidate and state-of-art research presently [7]. Researchers use either real time machine such as vector network analyzer or artificial neural network to analyze the received UWB signals [8-10].

Artificial neural network (ANN) is computed system processes the similar way human brain process the information. ANN is included feed forward neural network, radial basis function neural network, Kohonen neural network and recurrent neural network [11]. ANN is widely used for various applications especially in medical environment because it gives better performance [12]. For example, classifying type of breast cancer [13], early breast cancer detection [14-16]. It also can solve complicated problems as well as simplify the model. However, the size of data sample used to train affects the performance of the developed ANN. The performance increases if the size of data sample is sufficient and large [17]. Limited number of data samples normally occurs in medical line where limited number of patients resulted to minimum available data. In the other hands, the way of collecting the data samples are very complicated as need to take account multiple items if dealing with phantoms. Researchers use multiple types of neural network to solve the insufficient data samples for different types of applications. Sun M. et.al. proposes a new approach called Multiple Instance Learning Convolutional Neural Network (CNN) to solve insufficient data sample problem which ANN suffering with [18]. CNN is proved is the best way to solve the insufficient data sample compared to Support Vector Machine and K-Fold [19]. In typical ANN, data samples are divided into three groups: training, validation and testing while in Bayesian Regularization (BR) algorithm, data samples are divided into only two groups: training and testing. So, performance can be more efficient when dealing with limited data samples [20]. Wang L. et.al. stated Grey Neural Network (GNN) can deal with insufficient data samples more effectively since GNN is the combination of gray theory and ANN [21]. By using k-fold for insufficient data samples, it can maximize the usage of data samples. This is done by splitting the data samples into folds and train and test each fold [22- 23].).

However, these types of neural network have its own drawback. CNN need more information in each data sample in order to detect the tumor effectively. But this makes network architecture more complex. By using GNN and BR, the performance of the network may cannot be highly achieved. Furthermore, GNN is unable to perform with larger number of data sample. Therefore, in this paper, a breast cancer detection module using the combination of k-fold and feed forward backpropagation neural network is proposed for small data set. By using this method, the performance of the system can be maximized with limited data samples.

In this section (Section 1), introduction, problem statement and objective of this paper is discussed. In the following section, Section II, the procedure to perform the experiment is discussed in detail. In Section 3, the produced result from the proposed module in discussed. Finally, in Section 4, the conclusion from this experiment and future work is stated.

2. Material and Methods

The proposed breast cancer detection system contains of hardware and software. Hardware consists of two pair of antennas, breast phantom and Ultra wideband (UWB) transceiver with PC interfacing. Software consists of a Graphical User Interface (GUI) and a neural network module. The workflow of the overall system is as shown in Figure 1. Data is collected in first stage, in the left box of Figure 1. Data pre-processing, tumor detection and 2D and 3D visualizing is done in second stage, in the right box of Figure 1. A pair of antenna is attached to UWB transceiver (interfaced with PC environment). The antennas are used to transmit and received the UWB signal. The UWB signal is fed into neural network module for pre-processing, tumor detecting and tumor visualizing in 2D and 3D environment.

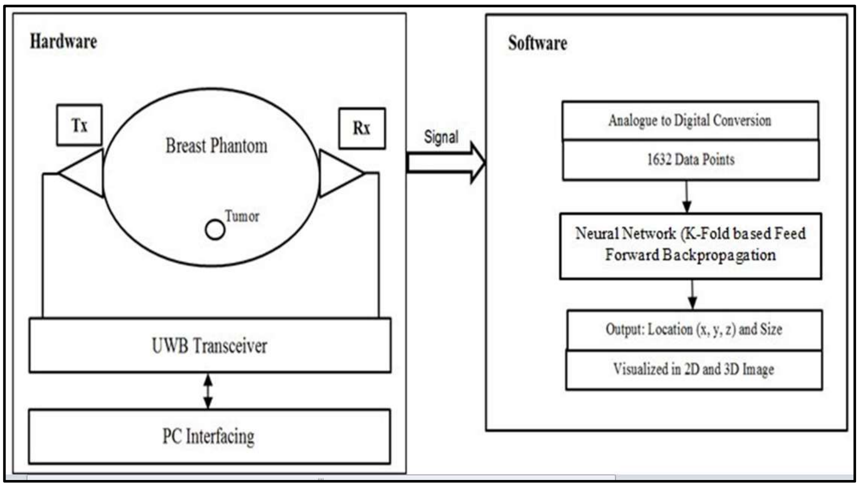
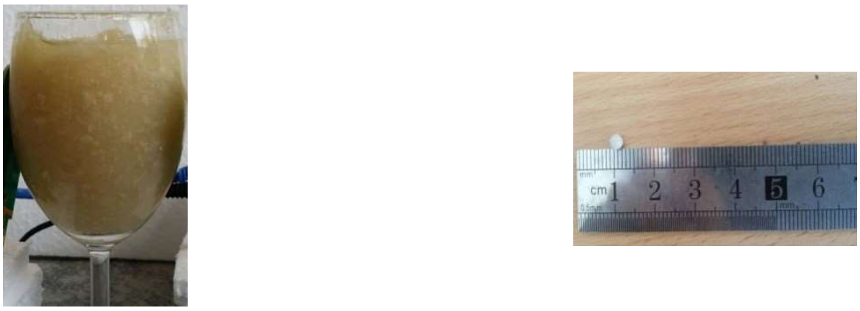

Hardware
Tx
Breast Phantom
Rx
Tumor
UWB Transceiver
PC Interfacing
Signal
Software
Analogue to Digital Conversion
1632 Data Points
Neural Network (K-Fold based Feed Forward Backpropagation
Output: Location (x, y, z) and Size
Visualized in 2D and 3D Image

Figure 1. Overall System Architecture.

2.1. *Breast Phantom*

Several types of breast phantom are proposed to investigate the detection of the breast cancer [24 – 26]. Most of the existing breast phantoms are developed based on the dielectric properties of real breast and tumors. Alshehri et. al. proposes low cost and non-chemical breast phantom. For this research, a heterogeneous breast phantom is developed based on [24] using petroleum jelly, mixture of flour and water and soy oil. The developed breast phantom is in hemisphere shape and met the size requirement as shown in Table 1 and Figure 2.



a) Heterogeneous Breast Phantom b) Tumor

Figure 2. Developed Breast Phantom and Tumor

Table 1. Breast Model Features.

Model Part	Size (cm)
Breast diameter	6.5
Breast height	5
Skin thickness	0.2

Tumor is developed with the mixture of flour and water (55%) [24]. Tumor is developed with the smallest tumor as possible. The developed tumor sizes are in the range of diameter 0.2 cm to 6 cm. Table 2 shows the dielectric properties (permittivity and conductivity) of proposed breast phantom.

Table 2. Dielectric properties of Considered Breast Phantom [24,30].

	Permittivity	Conductivity (S/M)
Skin	37.9	1.49
Fat	5.14	0.14
Tumor	50.0	1.20

2.2. *Data Collection*

The experiment is set up as in Figure 3, where the breast phantom is placed in between a pair of transmit- and receive- UWB antennas [7, 10]. The antennas are connected to the UWB transceiver through feeding cable. The UWB transceiver generates UWB pulses and transmitted through transmit antenna while the other antenna received the signal concurrently.

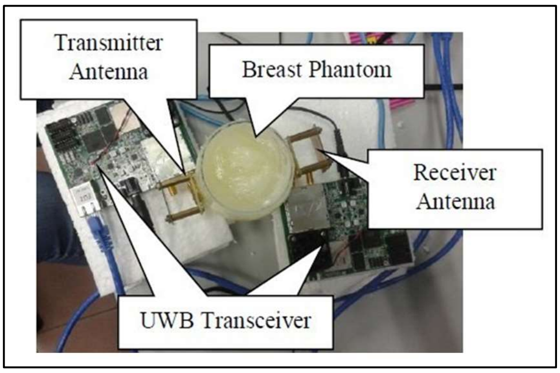


Figure 3. Experimental set-up [7, 10, 30].

The antennas are low-costs [7, 10] and place diagonally opposite of the breast phantom for forward scattering. They are placed close to the breast phantom to avoid the loss of signal and noise. The UWB transceiver with frequency range 3 to 10GHz is used and connected to PC through Ethernet cross connectors. Receive antennas capture the forward and backward scattered signals at 4.3GHz centered between 3 to 10 GHz shown in Figure 3. This process is repeated several times for various tumor sizes and locations (x, y, z). Tumor is placed along x, y and z in 134 different locations and size of tumors are 2mm, 3mm, 4mm, 5mm and 6mm. Also, breast phantom without tumor is used to get healthy signal. A total of 136 signals are collected. In this system, one transmit – one receive signal is utilized with a pair of antennas. This is done to make the system simple and low-cost in contrary with other systems with many antennas [7-9] [26] and/or collection of repeated signals (in rotation) [27-28].

2.3. *Feature Extraction*

A sample forward scattered transmitted and received signals is shown in Figure 4. The received signal is slightly different than transmitted signal due to some signals are scattered backward. These received scattered signals contain the signature of the tumor. They are processed to convert from analogue to digital and obtain 1632 data points for each signal. A large number of data points increases the processing time and builds a complex network structure. Four data points are extracted from 1632 data points. Four data points consist of mean, median, maximum number and minimum number [10].

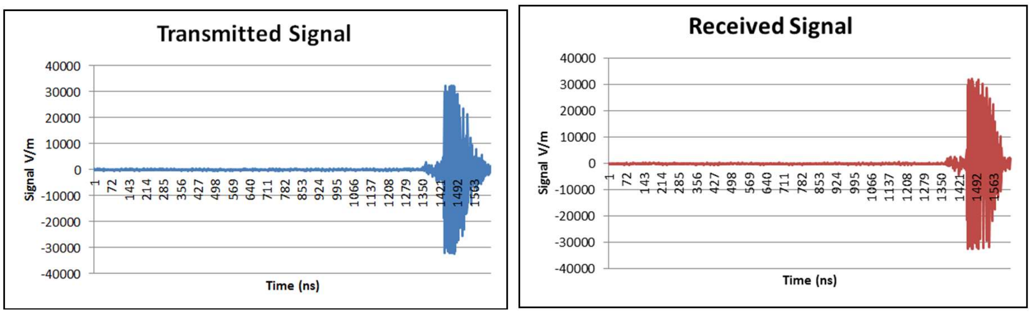


Figure 4. Transmitted and Received Signals.

2.4. Breast Cancer Detection Module

Feed forward backpropagation neural network (FFBPNN) is a type of ANN that flow in one way and does not have feedback cycle as shown in Figure 5. There are three layers which are input layer, hidden layer and output layer. Input layer is the first layer of the network and does not compute or process anything. It will just pass the information to the next layer called hidden layer. Output layer is the last layer of the network and receives the processed information from the hidden layer.

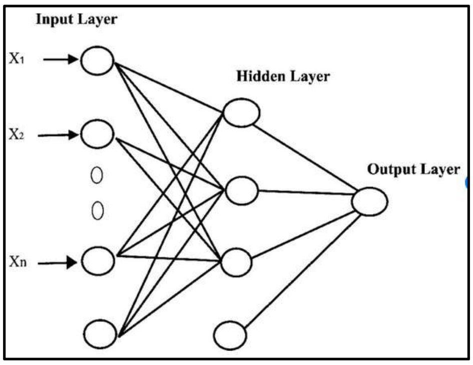


Figure 5. Architecture of the FFBPN.

Firstly, feed forward backpropagation neural network (FFBNN) module is developed. FFBNN can be expressed as below:

net=netff(x,t [n1,n2],{“tf”}

where x is input, t is target, n1 is hidden neuron in first layer, n2 is hidden neuron in second layer and tf is training function.

These variables are needed to be taken into account in order to develop the FFBNN module. The collected 136 data samples are divided into training and testing group as below:

Group (1): 125 data samples. 4 data sets are developed as in Table 3. Different number of data sample is set into each data set. These data samples are used for training, validating and testing.
Group (2): 11 data samples. These untrained data samples are used for real time testing to ensure the system detection performance efficiency for all data set.

Table 3. Data Set.

Data Set	Data samples		
1	100%	$\frac{100}{100} \times 125 = 125$	125
2	75%	$\frac{75}{100} \times 125 = 93.75$	95
3	50%	$\frac{50}{100} \times 125 = 62.5$	65
4	25%	$\frac{25}{100} \times 125 = 31.25$	30

The data samples in Group (1) are divided into 3 groups: training, validating and testing as shown in Table 4.

Table 4. Data sample for Training, Validating and Testing.

Data Set	Data samples			
	Training (70% from the total data sample)	Validating (15% from the total data sample)	Testing (15% from the total data sample)	Total Data Sample
1st	89	18	18	125
2nd	67	14	14	95
3rd	47	9	9	65
4th	22	4	4	30

Training session is done only for 1st data set repeated until the performance of the network is satisfied. Performance is optimized by increasing/decreasing number of neurons or inputs. Large data points led to more processing time and overfitting. Overfitting can be solved by decreasing the number neurons. The proposed FFBNN architecture has two hidden layers with 21 neurons and 4 nodes in the input and output layer respectively as shown in Figure 6. Table 5 shows the used FFBNN Matlab training parameters. Here to mention, the NN module, the network architecture and training parameters are exactly same for each data set in our experiment. The architecture is non-complex and could be easily executed as the training runtime for all datasets is within few seconds.

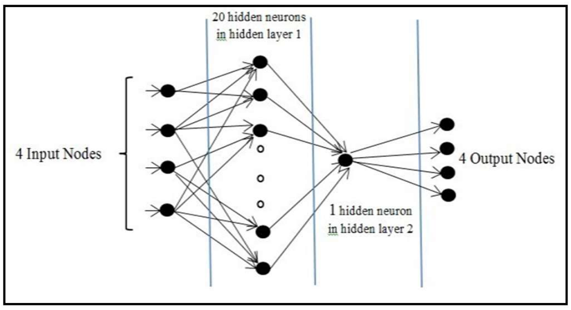


Figure 6. Proposed FFBNN Architecture.

Table 5. NN Parameter for Matlab Training.

NN Parameters	Value
Number of nodes in Input layer	4
Number of nodes in Hidden layer 1	20
Number of nodes in Hidden layer 2	1
Number of nodes in Output layer	4
Training function	trainlm
Learning rate	0.009
Momentum constant	0.6
Maximum number of Epochs	100000
Minimum performance gradient	1e-25

The proposed optimize FFBNN is expressed as below:

net = netff(x,t [20,1],{“trainlm”})

This expression is used to developed the new approach; the combination of k-fold and FFBNN. 5-fold is considered here. The smaller fold decreases the performance while the larger fold makes complex network architecture [29]. Each fold contains 25 data samples for 1st data set, 19 data samples for 2nd data set, 13 data samples for 3rd data set and 6 data samples for 4th data set. The folds are divided as in Table 6.

Table 6. Training and Testing Fold.

K-	Training				Testing
1	1	2	3	4	5
2	1	2	3	5	4
3	1	2	4	5	3
4	1	3	4	5	2

The training process is repeated by using 2nd, 3rd and 4th data set. Here to mention, k-fold and FFBNN module give better performance result compare to FFBNN [30]. After each training (for four cases), the error (E) of the NN module is calculated by using the proposed Equation (1).

$$MSE = \frac{i}{j} \sum j \, (t_j - y_j)^2 \qquad (1)$$

where j is number of input, t is actual target, and y is NN output. The data samples in Group (2) are used to test the developed K-fold and FFBNN module. This is to ensure the efficiency of the proposed system using untrained data samples. The accuracy of the proposed system is calculated by using the Equation (2):

$$Accuracy = \left(\frac{M-A}{A}\right) \times 100 \qquad (2)$$

where M is the K-fold and FFBNN‟s output and A is actual target.

2.5. Graphical User Interface

For our system, complicated program is replaced with graphical icon to make user's work easy with enhanced productivity. A Graphical User Interface is developed as part of software system in Matlab. Then, the developed GUI is compiled to an exe file to make it as independent all expensive software (including Matlab licence), hardware (e.g., VNA, etc.) and manually done UWB based soft system (signal processing + Neural Network, etc.). The home window of the GUI is as shown in Figure 7.

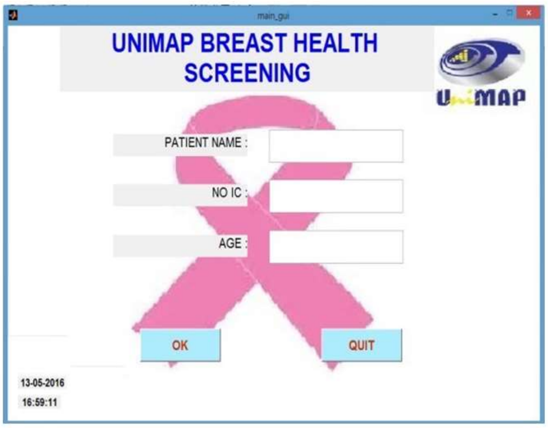


Figure 7. Home Window of GUI.

3. Results and Discussion

Figure 8 shows the performance of the network in terms of training, validation and testing. The best validation performance is achieved at epoch 11 with mean square error (MSE) value of 3.1209. The data of trained, validated and tested are analyzed as MSE and number of epoch. The performance is good based on the Figure 8 as the validation and testing curves are consistent and agreeable after epoch 11. The small number of iteration number/epoch used which are only 17 since few numbers of data points are used (four data points after extract important features from 1632 data points.

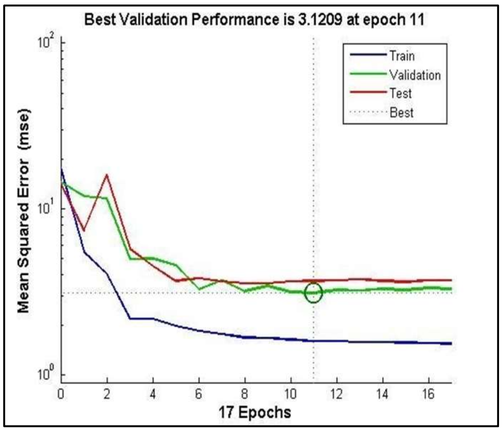


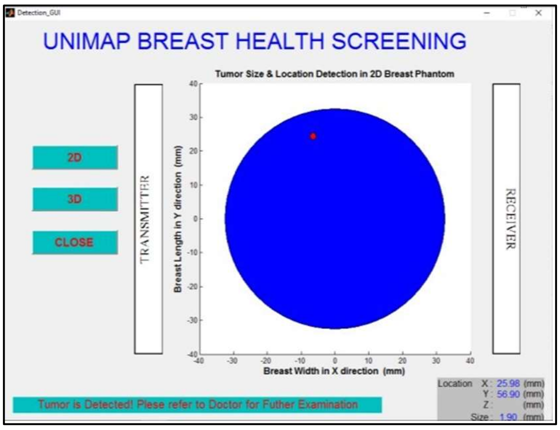
Figure 8. Performance of FFBPN Module.

Locations (x,y,z) and sizes of the tumor are randomly selected to test the accuracy of the proposed FFBPNN module. Table 7 shows the comparison between data sets. All data sets show the 100% detection efficiency in terms of tumor existence (i.e., whether there is any tumor or not). To summarize Table 7, the proposed FFBPN module is able to detect 3D location (x,y,z) and size of a tumor in the heterogeneous breast phantom with approximate average accuracy of 87.72%, 87.24%, 83.93% and 80.51% for 1st, 2nd, 3rd and 4th data set respectively.

Table 7. Performance Accuracy for Tumor Detection.

Data Sample	Performance Accuracy					Average (%)
	Existence	Location			Size	
		x	y	z		
125	100	77.83	74.77	95.29	90.69	87.72
95	100	79.99	74.76	93.33	88.14	87.24
65	100	76.56	73.23	87.83	82.02	83.93
30	100	73.34	71.14	81.02	77.06	80.51

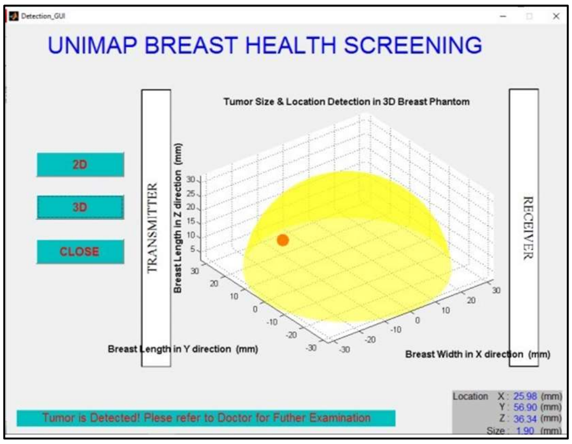
One of the detected 2 mm tumor is visualized in 2D and 3D environment through GUI as shown in Figures 9 and 10 respectively. This is to help end-user to check any breast abnormalities easily. The output's location is determined by the distance from center of the tumor to the outer skin line.


Detection_GUI
UNIMAP BREAST HEALTH SCREENING
2D
3D
CLOSE
TRANSMITTER
RECEIVER
Tumor Size & Location Detection in 2D Breast Phantom
Breast Length in Y direction (mm)
Breast Width in X direction (mm)
40
30
20
10
0
-10
-20
-30
-40

Tumor is Detected! Plese refer to Doctor for Futher Examination

Location X : 25.98 (mm)
Y : 56.90 (mm)
Z : (mm)
Size : 1.90 (mm)

Figure 9. 2D Imaging.


Detection_GUI
UNIMAP BREAST HEALTH SCREENING
2D
3D
CLOSE
TRANSMITTER
RECEIVER
Tumor Size & Location Detection in 3D Breast Phantom
Breast Length in Z direction (mm)
Breast Length in Y direction (mm)
Breast Width in X direction (mm)
30
25
20
15
10
5
-10
-20
-30

Tumor is Detected! Plese refer to Doctor for Futher Examination

Location X : 25.98 (mm)
Y : 56.90 (mm)
Z : 36.34 (mm)
Size : 1.90 (mm)

Figure 10. 3D Imaging.

4. Conclusion

In this paper, a breast cancer detection module based ANN is proposed. The module is developed by combining k-fold crossvalidation method with feed forward backpropagation neural network. The developed module is tested for breast cancer detection application. Four data sets with different number of data samples ie: 125, 95, 65 and 30 data samples respectively. All data sets are able to identify the presence of tumor with 100%. The average detection performance efficiency is approximately more than 80% for all four data sets. This concludes that the developed module can be used either for large data sample or small data sample in different environment without any difficulties. Therefore, this module is able to solve the insufficient data sample problem for all application especially for medical application.

Acknowledgement
This work is supported by Ministry of Higher Education, Malaysia, Grant FRGS – 9003-00418.

References

[1] Kshetrimayum RS 2009 *IEEE Potentials* **28** 9-13.
[2] National Cancer Sociery Malaysia (NCSM) 2020 [Online]. Available http://www.cancer.org.my/quick-facts/types-cancer/.
[3] Huynh PT, Jarolimek AM and Daye S 1998 *Radiographics* **18** 1137-1154.
[4] Mjphmorgmy 2020 [Online]. Available http://www.mjphm.org.my/mjphm/journals/2015 - Volume 15 (1)/State Level Variation Of Breast Cancer Cases In 2007 Among Malaysian Women.pdf.
[5] National Research Council 2005 *National Academies Press* Joy JE, Penhoet EE, & Petitti DB (N.W. Washington, DC).
[6] Tabar L, Yen MF, Vitak B, Chen HHT, Smith RA and Duffy SW 2003 *The Lancet* **361** 1405-1410.
[7] Reza KJ, Khatun S, Faizal M, Ikram E, Ishwar Z and Khalib A 2013 *Int. Journal of Engineering and Technology* **5**.
[8] Salleh M, Hasmah S, Othman MA, Ali N, Sulaiman HA, Misran MH, Aziz A and Abidin MZ 2015 *ARPN Journal of Engineering and Applied Sciences* **10** 723-727.
[9] Xia X, Hang S, Zong-Jie W and Liang W 2014 *Chin. Phys. B.* **23** 1-5.
[10] Reza KJ, Khatun S, Jamlos MF, Fakir MM and Mostafa SS 2014 *ARPN Journal of Engineering and Applied Sciences* **9** 329-335.
[11] Jha GK 2007 *IARI, New Dehli* V42-V49.
[12] Khan IY, Zope PH and Suralkar SR 2003 *Int. Journal of Engineering Science and Innovative Technology (IJESIT)* **9** 210-217.
[13] Abdel-Ilah L, Šahinbegović H 2017 *Proc. of the Int. Conf. on Medical and Biological Engineering 2017* vol 62 (Sarajevo, Bosnia and Herzegovina/ Springer) p 3-8.
[14] Mazen F, AbulSeoud RA and Gody AM 2016 *Int. Journal of Computer Trends and Technology (IJCTT)* **32**.
[15] Mandal S, Saha G, Pal RK 2015 *Global Journal on Advancement in Engineering and Science (GJAES)* **1**.
[16] Saini S and Vijay R 2015 *2015 Fifth Int. Conf. on Communication Systems and Network Technologies* (Gwalior, India) p 1177- 1180.
[17] Roohi F 2013 *The Int. Journal of Engineering and Science (IJES)* **2** 33-38.
[18] Sun M, Han TX, Liu MC, Khodayari-Rostamabad A 2016 *2016 23rd Int. Conf. on Pattern Recognition (ICPR)* (Cancun, Mexico) 3270-3275.
[19] Nguyen PM 2016.
[20] Chaipimonplin T 2016 *KSCE Journal of Civil Engineering* **200** 478-84.
[21] Wang L, Jiang H, He D 2014 *Open Civil Engineering Journal* 416-419.
[22] Lever J, Krzywinski M, Altman N 2016 *Nature Methods* **13** 703.
[23] Vanwinckelen G and Blockee lH 2012 *InBeneLearn 2012: Proc. of the 21st Belgian-Dutch Conference on Machine Learning* (Ghent, Belgium) p 39-44.
[24] AlShehri SA and Khatun S 2009 *Progress In Electromagnetics Research C* **7** 79–93.
[25] Porter E, Fakhoury J, Oprisor R, Coates M and Popović M 2010 *Antennas and Propagation*

(EuCAP) 1-5.
[26] Lazebnik M, Popovic D, McCartney L, Watkins CB, Lindstrom MJ, Harter J, Sewall S, Ogilvie T, Magliocco A, Breslin TM and Temple W 2007 *Phys.Med. Biol.* **52** 6093–6115.
[27] AlShehri SA, Khatun S, Jantan AB, Raja Abdullah RSA, Mahmud R and Awang Z 2011 *Progress In Electromagnetics Research* **111** 447- 465.
[28] AlShehri SA, Khatun S, Jantan AB, Raja Abdullah RSA, Mahmud R and Awang Z 2011 *Progress In Electromagnetics Research* **116** 221-237.
[29] Raschka S 2016 [Online]. Available https://sebastianraschka.com/blog/2016/model-evaluation-selection-part3.html.
[30] Vijayasarveswari V, Khatun S and Jusoh M, Fakir M M 2017 *Indian Journal of Science and Technology* **10** 1-6.