THROMBO - EMBOLIC STROKE PREDICTION AND DIAGNOSIS USING ARTIFICIAL NEURAL NETWORK AND GENETIC ALGORITHM

L. T. Popoola[1], Gutti Babagana[1] & A.A. Susu[3]
[1]Department of Petroleum and Chemical Engineering, Afe Babalola University, Ado-Ekiti, Ekiti State, Nigeria
poplektaoy@yahoo.co.uk
[2]Department of Chemical Engineering, University of Maiduguri, Maiduguri, Borno State, Nigeria
babaganagutti@yahoo.com
[3]Department of Chemical Engineering, University of Lagos, Lagos, Nigeria
alfredasusu222@hotmail.com

ABSTRACT
The Back Propagation Neural Network algorithm was validated using hypothetical data from fifty patients with symptoms of Stroke. The data set was divided into training set and test set while the validation data set were chosen randomly from the testing data. Forty-two (42) data set were used for the training set while eight data set were used for the test. Four data were chosen from the test set and used for the validation. A MATLAB program was written for training, testing and validation of the neural network. Three different architectures with 5, 10 and 20 hidden neurons in the network architecture were tested to avoid overfitting and inaccuracy after which neural network with 10 hidden neurons was chosen as the best architecture. The training error converged to 0 after 50 iterations with architecture of 10 hidden neurons while convergence was almost achieved after 100 and 1000 iteration steps with 5 and 10 hidden neurons respectively. The ANN was trained and tested after optimizing the input parameters using Genetic Algorithm, the overall predictive accuracy obtained for the thrombo-embolic stroke was 90%.

Keywords: *Thrombo-Embolic Stroke, Artificial Neural Network, Architecture, Genetic algorithm, Training, Validation, Testing and Back-propagation Algorithm.*

1. INTRODUCTION

Brain is the control centre of the body. It controls thoughts, memory, speech and movement. It regulates the function of many organs. When the brain is healthy, it works efficiently and automatically. However, when disorders occur, the results can be devastating. One of the common disorders of brain is ***stroke*** also called ***brain attack*** (a medical emergency). Stroke is a major cause of death and disability in both the developed and the less developed countries [1,2]. Stroke is a life-threatening event in which part of the brain is not getting enough oxygen. Stroke consumes an important part of the total healthcare costs (i.e. excluding social care and indirect costs) in Europe and USA [3]. Stroke is characterized by the sudden loss of blood circulation to a region of the brain, resulting in a corresponding loss of neurological function [4].

There are different types of stroke namely Brain Attack, Embolic Stroke, Thrombotic Stroke, Ischemic Stroke, Cerebrovascular Accident (CVA). Medical personnel treating a stroke are challenged to treat the patient as quickly as possible to avoid permanent tissue damage or death. Strokes were responsible for more deaths and nearly half of those deaths occurred outside of a hospital. Stroke is the third leading cause of death, behind heart disease and cancer. Most recovery occurs during the first few months following a stroke. According to the National Institute of Health, the risk of stroke is greater and the recovery process is slower [5].

Thrombo-embolic strokes are caused by fatty deposits (plaques) that have built up in the arteries carrying blood to the brain. This slows blood flow and can cause clots to form on the plaques that narrow or block the flow of oxygen and nutrients to the brain. It is also caused by a blood clot formed in another part of the body that breaks loose, travels through the bloodstream, and blocks artery carrying oxygen and nutrients to the brain. When travelling through the body the blood clot is called an embolus [6].

The symptoms of stroke usually come on suddenly. The suddenness of onset distinguishes stroke from other conditions such as migraine or brain tumours. Treatments are available to minimize the potentially devastating effects of stroke, but to receive them; one must recognize the warning symptoms and what are the risk factors that increase the probability of brain attack. The symptoms and risk factors of brain diseases vary widely depending on the specific problem. In some cases, damage is permanent. In other cases, treatments such as surgery, medicines or physical therapy can correct the source of the problem or improve symptoms [7].

Neural Networks are one of many data mining analytical tools that can be utilized to make predictions for medical data. Neural networks provide a very general way of approaching problems. When the output of the network is continuous, it is performing prediction and when the output has discrete values, and then it is doing classification. A simple rearrangement of the neurons and the network becomes adept at detecting clusters [8]. With the computerization in hospitals, a huge amount of data can be collected. Employing the technology especially artificial neural network (ANN) techniques in medical applications could result in reducing the mortality rate, cost, time, medical error and need of human expertise [9].

2. LITERATURE REVIEW

Computer Assisted Decision Support in medicine has at least the role of enhancing the consistency of care. It has the potential to cover rare conditions, since no clinical expert can be expected to possess encyclopaedic knowledge of all of the exceptional manifestations of diseases, even within a specialist domain. Saritas *et.al*. [10] examined Fuzzy Expert System Design for Diagnosis of Prostate Cancer. They developed a rule-based fuzzy expert system (FES) that used the laboratory data and simulated an expert-doctor's behaviour. The system was developed with the aid of MATLAB 6.5. It was shown that the FES gave close results with the data of 4641 patients from Brawer *et.al*. [11] and Seker *et.al*. [12]. Adam *et.al*. [13] proposed Computerized Breast Cancer Diagnosis with Genetic Algorithms and Neural Network. Optimum initial weight for each layer in the NN architecture was calculated with GA approach, following the method suggested by Khairuddin Omar [14]. While for mining process, multilayer perceptrons with back propagation was used. Available datasets for breast cancer were studied and pre-process for cleaning data was carried out to prepare it for mining process. They found out that medical data were best kept in its original value as it gave high accuracy percentage as compared to the altered data.

Guru *et.al.* [15] examined Decision Support System for Heart Diseases Diagnosis Using Neural Network. They suggested supervised network for diagnosis of heart diseases and trained it using back propagation algorithm. The system was trained for 78 patients' record. They concluded that the success rate for imprecise inputs to retrieve the desired output or the data that was closest to the desired output was 100%. Thakur *et.al.* [16] improved on the work done by Guru *et.al.* [15] by proposing Early Diagnosis of Ischemia Stroke using Neural Network. They provided 280 samples of patients with 10 combining risk factors and symptoms for cerebral ischemia stroke. The limit for training error was kept at 1e-8. The network was validated and tested with datasets. Two third data are used for validation and one third data are used for testing. Regarding this, MATLAB program was developed to model and train NN achieving accuracy of 99.99%.

Kavitha *et.al.* [17] improved on the work done by Guru *et.al.* [15] by combining Genetic Algorithm with Neural Network to detect heart diseases. They proposed Modelling and Design of Evolutionary Neural Network for Heart Disease Detection. The data set contained 270 patient's record and each patient condition was defined by 13 parameters, 150 patient's record taken for training data set and remaining 120 for test data set. They envisaged developing an expert system for heart disease detection in future. With the proposed system they hope that, design of diagnosis system for heart disease detection would become easy, cost effective, reliable and efficient. Karegowda *et.al*. [8] did a recent research on diabetes prediction by proposing Application of Genetic Algorithm Optimized Neural Network Connection Weights for Medical Diagnosis of Pima Indians Diabetes. A total of 768 cases were examined. The data set was divided into Training and Test using 60-40 ratio. They concluded that the hybrid GA-BPN showed substantial improvement in classification accuracy of BPN. Significant features selected by DT (Decision Tree) and GA-CFS (Correlation-based Feature Selection) further enhanced classification accuracy of GA-BPN.

3. MATHEMATICAL MODELS AND ARCHITECTURE

3.1 Modelling Equations

The binary sigmoidal transfer function considered as the activation function for each node in the network is defined thus:

$$y_k = F_N(Z_k) = \frac{1}{1+\exp^{(-Z_k)}} \quad \text{.........................} \quad 3.1$$

where y_k is the Sigmoidal Transfer function and limits the output of all nodes in the network to be between 0 and 1. Z_k which is the sum of the inputs x_i multiplied by their respective weights is defined thus:

$$Z_k = \sum_j w_{kj} x_j \quad \text{.........................} \quad 3.2$$

w_{kj} is weight of the jth input to the kth neuron of the output layer and x_j is the jth input to the neuron.
The error function calculated thus:

$$error = \frac{1}{2}\sum_{j=1}^{m}[d_j(n) - y_j(n)] \qquad .... \quad 3.3$$

where d_j = practical data of jth output neuron

y_j = computed data of jth output neuron

m = neuron number

n = training step

The Back Propagation Algorithm for training the artificial neural network and updating the weights is calculated thus:

$$w_{ij}^{k+1} = w_{ij}^{k} + \eta\delta_j^k I_i f'(s) \qquad .... \quad 3.4$$

where

w_{ij}^k = weights of the connection from unit i in layer k to unit j in layer k+1.

η = learning rate (constant)

δ_j^k = signal error.

I_i = input vector to the networks.

f '(s) = derivative of the networks sigmoidal transfer function.

s = sum of all the weights.

3.2 Proposed Patients Data

A hypothetical data of 50 patients who have symptoms of stroke disease was formulated based on the information extracted from medical textbooks. The data were standardized so as to be error free in nature. Table 3.1 below shows the various input parameters for the prediction of stroke disease.

Code	Input Column Name
X1	Hypertensive
X2	Diabetes Mellitus
X3	Myocardial infarction
X4	Cardiac failure
X5	Atrial fibrillation
X6	Smoking
X7	Blood cholesterol
X8	Left arm and leg paralysis
X9	Right arm and leg paralysis
X10	Slurring of Speech
X11	Giddiness
X12	Constant headache
X13	Vomiting
X14	Memory loss
X15	Swallowing difficulty
X16	Vision impairment
X17	Transient Double Vision
X18	Vertigo
X19	Numbness
X20	Dizziness

Table 3.1: **Input Parameters for Prediction of Stroke**

The various categories of Stroke diseases and their classification which are the results generated from the input given to the neural network are presented in Table 3.2 below.

Code	Output
D1	Transient Ischemia Attack
D2	Left Hemiplegia
D3	Right Hemiplegia
D4	Dysphasia
D5	Monoplegia
D6	Left Hemianopia
D7	Aphasia
D8	Right Hemianesthesia
D9	Dysphagia
D10	Quadruplegia

Table 3.2: **Categories of Stroke Diseases and their Classification**

3.2 Neural Network Architecture and Genetic Algorithm Framework.

The neural network architecture is multi-layered feed-forward network architecture with 20 input nodes, 10 hidden nodes, and 10 output nodes. The number of input nodes is determined by the finalized data; the numbers of hidden nodes are determined through trial and error; and the numbers of output nodes are represented as a range showing the disease classification. Training a neural network involves modifying the weights and biases of the network in order to minimize a cost function. The cost function is an error term which measures the closeness of the network's predictions to the class labels for the examples in the training set. The Neural Network architecture is shown in figure 3.1 while the Genetic Algorithm framework is shown in figure 3.2.

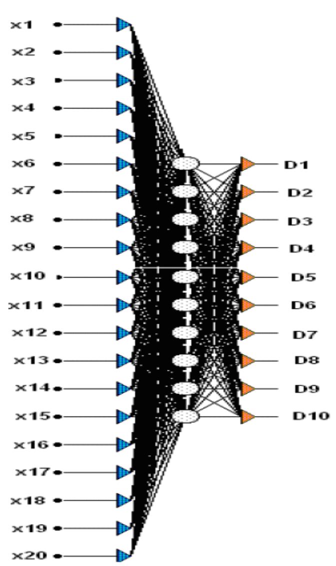


Fig 3.1: Best Neural Network Architecture for Stroke Prediction

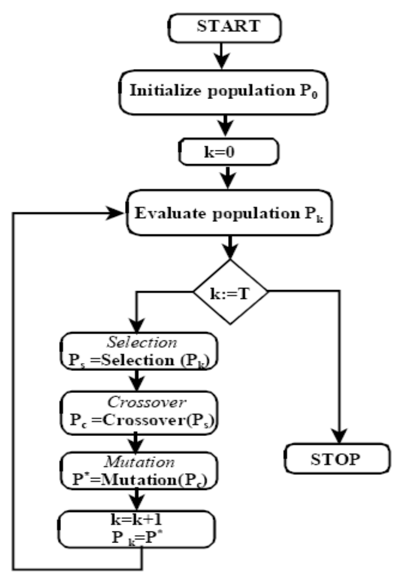
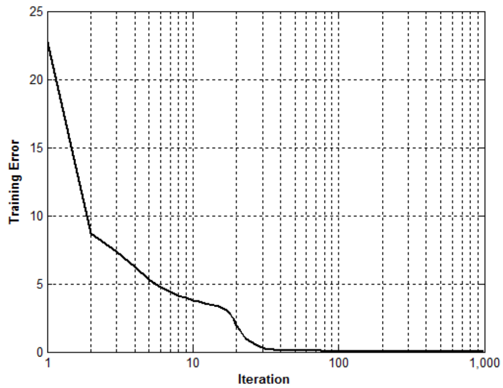


Fig 3.2: Standard Genetic Algorithm

4. RESULTS AND DISCUSSION

4.1 Results

In this research work, fifty (50) samples of patients with different symptoms of stroke were provided to train and test the network. Neural network program written in MATLAB language was used to determine the best possible test results. Different neural network architectures were also tested. The results obtained after running the program are depicted below.



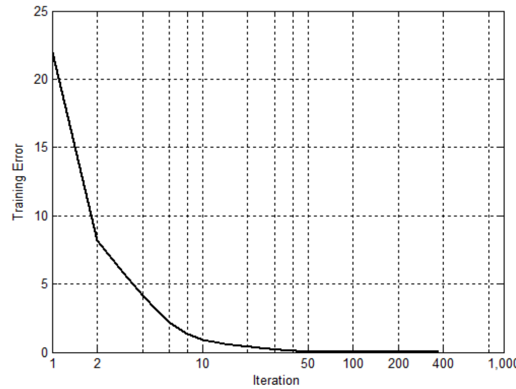
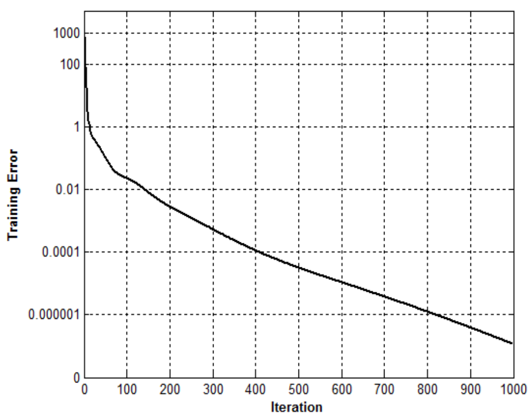


Fig 4.1: Training Error Vs. Iteration using 5 Hidden Neurons

Fig 4.2: Training Error Vs. Iteration using 10 Hidden Neurons



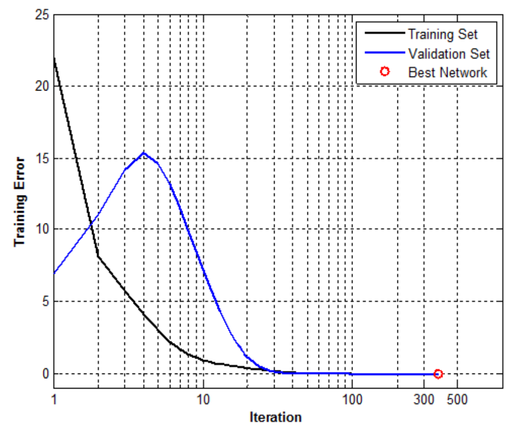


Fig 4.3: Training Error Vs. Iteration using 20 Hidden Neurons

Figure 4.4: Data Set Errors of the Best Architecture

4.2 Discussion

Training the neural network allows setting the best weights on the inputs of each of the units. The goal is to use the training set to produce weights whose output of the network is as close to the desired output for the prediction of stroke. The training set is a part of the input data set used for the adjustment of network weights. The test set is a part of the input data set used to test how well the neural network will perform on new data. The test set is used after the network has been trained to test what errors will occur during future network application. This set is not used during training and thus can be considered as consisting of new data entered by the user for the neural network application.

In order to obtain the best neural network architecture and avoid overfitting problem, three different architectures with 5, 10 and 20 hidden neurons in the network were tested. Overfitting occurs when the network simply memories all training examples, thus prevent it from generalizing and producing correct output when presented with test data. This happens if the number of hidden neurons used for the network is too big. Figure 4.1 depicts the graph of training error against iteration using 5 hidden neurons. The training error decreases steadily from 23 until it is almost 0 (i.e. constant) after 100 iterations. Figure 4.2 represents the graph of training error versus iteration using 10 hidden neurons. The training error converges to 0 after 50 iterations. Figure 4.3 represents the graph of training error against iteration using 20 hidden neurons. The training error reduces with increasing iteration number but could not achieve convergence after 1000 iteration steps.

This occurs as a result of overfitting which is the result of using too many neurons in the network architecture. The reverse is the case if the numbers of hidden neurons are few. In this case, the neural network will not give accurate results. This occurs when 5 hidden neurons were used in the architecture. When 10 hidden neurons were used, quick convergence was achieved after 50 iterations with accurate result unlike other two cases where convergence was delayed with inaccurate results. Hence, neural network architecture with 10 hidden neurons is the best architecture for the stroke prediction data set.

The trained network was tested with a test set, in which the outcomes are known but not provided to the network, to see how well the training has performed. The data set was divided into training set and test set while the validation data set were chosen randomly from the testing data. Forty- two (42) data set were used for the training making 84% of the data set while eight (8) data set were used for the test making 16% of the hypothetical data set. Four (4) data were chosen from the test set and used for the validation. Figure 4.4 shows the various data set errors with respect to training set, validation set and the best network. After training through repeated iterations of over 400 it reaches the level of best network. The test data set demonstrates that the ANN based prediction of stroke disease improves the diagnosis accuracy with higher consistency. This ANN exhibits good performance in the prediction of stroke disease in general.

Figure 4.5 is the chart showing the various stroke diseases classification and their percentage based on the number of hypothetical data used for the stroke prediction. The analysis shows clearly that 2% each of the respondents have the symptoms of Dysphagia and Quadruplegia respectively, 4% have the symptoms of Right Hemianesthesia, 6% are suffering from Monoplegia, 8% each have the symptoms of Left Hemianopia and Aphasia respectively, 10% of the patients have the symptoms of Dysphasia, 14% each have the symptoms of TIA and Right Hemiplegia respectively and lastly, 32% of the respondents have the symptoms of Left Hemiplegia.

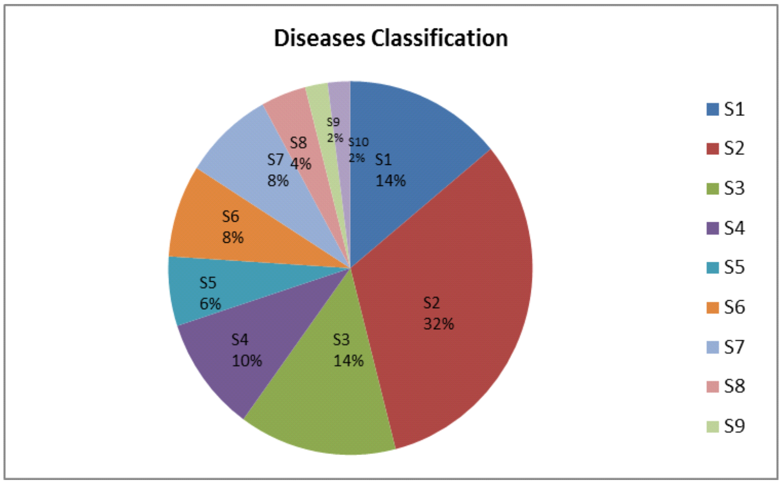


Fig 4.5: Various Stroke Diseases and their Respective Number of Cases

5. CONCLUSIONS

The validation of Neural Network algorithm had been achieved using hypothetical data of patients with symptoms of stroke diseases. A MATLAB computer programing language had been written to simulate the hypothetical data. Also it has been proved that Genetic Algorithm and Back-Propagation neural network hybrids in selecting the input features for the neural network reveals the performance of ANN can be improved by selecting good combination of input variables. The Neural Network architecture with 10 hidden neurons had been chosen to be the best architecture of the data sets in order to avoid overfitting and inaccuracy. Neural networks are useful medical tools for early detection of diseases and hence, can be used in medical diagnosis and prediction of thrombo-embolic stroke. ANN based model has been used to develop a system in which patients would be able to self-diagnose themselves before undergoing a more thorough examination. Neural networks will never replace human experts but they can help in screening and can be used by experts to double-check their diagnosis. In conclusion, when the ANN was trained and tested after optimizing the input parameters using Genetic Algorithm, the overall predictive accuracy obtained was 90%.

REFERENCES

[1]. Fiez, J.A. "Lesion Segmentation and Manual Warping to a Reference Brain: Intra- and Inter observer Reliability", Human Brain Mapping 2000, Vol. 9, pp. 192–211.
[2]. Sudlow, C.L.M. "Comparable Studies of the Incidence of Stroke and Its Pathological Types-Results from an International Collaboration", Stroke 1997, Vol. 28(3), pp. 491–499.
[3]. Evers, M. "Economic Evaluation in Stroke Research", J Cerebrovascular Dis 2000, Vol.11, pp. 82-91.
[4]. Stein, J. "Stroke and the family: a new guide", Harvard University Press, 2004.
[5]. Shanthi, D., Sahoo, G. and Saravanan, N. (2008) "Input Feature Selection using Hybrid Neuro-Genetic Approach in the Diagnosis of Stroke", International Journal of Computer Science and Network Security, ISSN: 1738-7906, Vol.8, No.12, pp.99- 107.

[6]. Mohr, J. P. (2001), "Stroke Analysis", 4th Edition, Oxford Press, New York.
[7]. Haykin, S. (1999), "Neural Networks - A Comprehensive Foundation", 2nd Ed. Prentice Hall.
[8]. Karegowda, A.G., Manjunath, A.S. and Jayaram, M.A. "Application of Genetic Algorithm Optimized Neural Network Connection Weights for Medical Diagnosis of Pima Indians Diabetes", International Journal on Soft Computing, Vol.2, No.2, May 2011.
[9]. Ahmed, F.E. "Artificial Neural Networks for Diagnosis and Survival Prediction in Colon Cancer", Mol Cancer 2005, Vol. 4, pp. 24-29.
[10]. Saritas, I., Allahverdi, N. and Sert, I.U. (2003) "A Fuzzy Expert System Design for Diagnosis of Prostate Cancer", International Conference on Computer Systems and Technologies.
[11]. Brawer, M. K. and Kirby, R."Prostate Specific Antigen", Second Edition, Health Press, pp. 92, 1999.
[12]. Seker, H., Odetayo, M., Petrovic, D. and Naguib, R.N.G., "A Fuzzy Logic Based Method for Prognostic Decision Making in Breast and Prostate Cancers", IEEE Trans. On Information Technology in Biomedicine, 2003.
[13]. Adam, A. and Omar, K. (2006), "Computerized Breast Cancer Diagnosis with Genetic Algorithms and Neural Network", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 14, No 6, pg. 686-691.
[14]. Khairuddin Omar., Pengawalan Pemberat unutk Perambat Balik Menggunakan Algoritma Genetik: Hasil Ujikaji, Laporan Teknikal Perkembangan Penyelidikan PhD Thesis, Universiti Putra Malaysia, 2000.
[15]. Guru, N., Dahiya, A. and Rajpal, N. (2007), "Decision Support System for Heart Diseases Diagnosis Using Neural Network", Delhi Business Review, Vol.8, No.1.
[16]. Thakur, A., Bhanot, S. and Mishra, S.N. (2009), "Early Diagnosis of Ischemia Stroke using Neural Network", Proceedings of the International Conference on Man-Machine Systems, Batu Ferringhi, Penang, Malaysia.
[17]. Kavitha, K.S., Ramakrishnan, K.V. and Singh, M.K. "Modelling and Design of Evolutionary Neural Network for Heart Disease Detection", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010.