www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 7 July 2017, Page No. 21895-21915
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i7.04

Diagnosing Dengue: A Faster, Artificial Intelligence Based Hack

[1]Nitin Tyagi, [1,#]Apoorvaa Singh, [1]Himanshi Sharma, [1]Krishna Tripathi

[1] Department of Computer Science and Engineering, Inderprastha Engineering College, Ghaziabad

* Corresponding author email: apoorvaa.singh30@gmail.com,

Abstract:

Dengue fever is a seasonal, vector borne disease which is often deadly. At present, dengue is usually diagnosed through two stages of tests in India. A patient showing the physical symptoms of dengue is first subjected to a screening test, the CBC test. The second test, the dengue serology test, is the truly confirmatory test, but may take up to 10 days to return a correct reading. We propose a simple neural network based model which can detect whether the patient has dengue, with just the preliminary CBC test report's data. Patient data was collected from a single hospital located in Ghaziabad, India. We found that the system correctly classified the unseen test cases with a significant degree of accuracy. We further propose as future research directions the application and comparison of the modelling results of more pattern recognition techniques for this classification task, testing of the system in real-time hospital conditions, and the inclusion of locality specific factors to build as general and as widely-reproducible a model as possible.

Keywords**:** Artificial Neural Networks, Dengue, Pattern Recognition, Health Informatics.

1. Introduction

Dengue fever is an often-deadly disease spread by mosquitoes. It occurs almost every year during the monsoon in sub-tropical and tropical regions around the world. At present, there is no specific medicine prescribed for its treatment – the patient must be given complete rest and plenty of fluids to help the body flush out the toxins. Time is of essence in the treatment of dengue though. Once the physical symptoms of the patient are suspected to be of dengue, the general practice is to prescribe two stages of tests, the preliminary CBC test and the confirmatory serology test. But the same parameter variations as in the CBC test of a true dengue patient are also found in CBC reports of patients suffering from other diseases. The dengue serology test, though fast and convenient, often takes around 10 days from the day of the infection to detect dengue in the patient's blood, and the patient's health can steadily deteriorate within the span of these 10 days.

We propose that the pattern learning capability of artificial neural networks can be applied to facilitate accurate and fast diagnosis of dengue from the CBC report itself. A neural network model is applied to learn

the true model behind the variations that occur in the normal ranges of leucocytes, haemoglobin, platelets and packed cell volume, when a person gets infected with the dengue virus. We first discuss the diagnostics related to dengue and the issues that warrant resolution. Then we discuss the application of neural networks to medical science and some celebrated success stories. Next, we describe the data and the models built, and wrap up with an analysis of the test sets results and the future directions in which this research can proceed.

1.1. Introduction To Diagnostics Related To The Dengue Disease And Difficulties That Arise Therein

The diagnosis of dengue is through physical symptoms and clinical tests. There are several laboratory-based tests available for dengue:

- virus isolation from cell cultures,
- PCR based detection of nucleic acids,
- detection of viral antigens (usually the NS1 antigen)
- detection of specific antibodies (serology test).

In India, the usual practice among medical practitioners is to prescribe two levels of tests if the patient shows the physical symptoms of dengue – a preliminary blood test and a confirmatory serology test. The other tests mentioned above are not used as confirmatory tests either due to higher cost or due to reduced sensitivity in case of secondary infections. [11]

The blood test, or the CBC test, tests the level of leucocytes (TLC), platelets (Pt), haemoglobin (Hb) and packed cell volume (PCV) in the blood of the patient. If the TLC is between 1800 to 4000, the Hb is between 9 to 15 thousand and the platelet count is less than 70,000, the patient is suspected to be suffering from dengue. For two CBC tests conducted on the same patient on two consecutive days, the second test's report shows a marked decline in the haemoglobin, platelet and leucocyte count of the patient as the infection progresses. For severe cases, the decline can be significant on an hourly basis too. But these variations in the given parameters are also exhibited in the case of simple viral infections, thrombocytopenia, etc.

The serology test checks for dengue virus specific antibodies – types IgG and IgM. Both are produced 5 to 7 days after the infection. IgG antibodies, once produced during the first or the primary infection, remain in the blood for years. The IgM antibodies, though, become undetectable 30 to 90 days after both primary and secondary infections. IgG concentrations reach peak levels in the blood 14 to 21 days after the primary infection and earlier in case of secondary infections. Thus, the result of the serology test can take a long time to come out as positive even if the patient truly has been infected with dengue. Also, the test can cross-react with other flavivirus infections – it may result in a false positive if the patient has suffered from infections of, or taken vaccinations for, diseases like yellow fever and Japanese encephalitis in the recent past. [11]

DOI: 10.18535/ijecs/v6i7.04

2. Artificial Neural Networks

Artificial Neural Network (ANN) can be visualized as a computing system made up of several small, highly interconnected processing elements called nodes. These nodes process information responding dynamically to external inputs. [Dr. Robert Hecht-Nielsen] An ANN is an information processing paradigm which is inspired by the way biological neural systems, like the brain, process information. It is also commonly referred to as a "neural network". [1]

2.1. ANN As A Graphical Model

Several models of artificial neural networks, each with their own advantages, have been proposed over the years, but the basic structure remains the same across all of them. The network has many small processing units, arranged in three types of layers – one input layer, one or more than one hidden layers, and one output layer. The number of nodes to be included in the input layer is taken from the number of features in the input data set – the number of nodes is equal to the number of features. On the other hand, the number of nodes in the hidden layer is adjustable – it should be such that the training data is neither over-fitted nor under-fitted during the training process. The number of nodes in the output layer corresponds to the number of features we need as output – there is just one node in case of simple classification problems, the output being 1 if the classification is the desired one and 0 if it is not. [Haykin, 13] [4]

All the nodes in all three layers are connected to each other. The connections have weights assigned to them, which are a way of representing how much significance the connection holds. Each node has an activation function that processes the incoming input from each connection, multiplied by the weight assigned to that connection, to calculate that node's output. This output then serves as the input to the next node to which this output node is connected. [C.M. Bishop, 14]

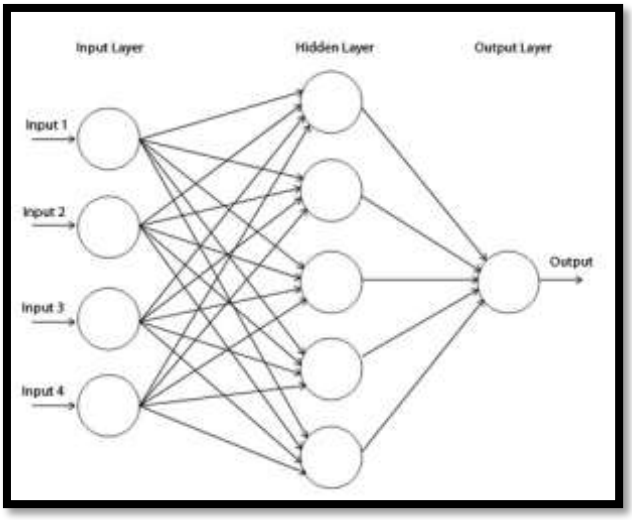

Figure 1. Pictorial representation of a neural network with 1 hidden layer containing 5 neurons.

The neural network for solving a given problem gets created in three stages – the training, testing and validation phases. Training can be done in two main ways – supervised learning in which each training data set has the target output attached to it and the neural network has to adjust itself to bring its outputs close to the target outputs, or the learning process can be unsupervised with the input data not having target outputs associated with them. [Hagan, 15] [6]

Data corresponding to the given problem's domain is fed to the neural network and its weights are adjusted till the output for each input data set has the least possible amount of error. The network developer also has the choice of deciding the training algorithm, the activation function, and the number of hidden layers and the number of nodes in them. [5]

2.2. Applications of Ann in Medical Sciences

The scientific community has been trying to address this question for a long time – how to use computers and software to make the work of doctors easier. Such systems, depending on the specific functionalities they provide, can be known by various names like decision support systems, expert systems, health informatics systems, etc. Usually, decision support systems refer to software that help in the management, organization and planning of various decision making activities. In the medical context, they also refer to software that help confirm diagnosis decisions. Expert systems are computer systems that emulate the decision-making ability of a human expert. They are designed to solve complex problems by reasoning about knowledge, which is usually represented in the form of if–then rules, rather than conventional procedural code. Health informatics systems are used for dealing with the resources, devices, and methods that are required to optimize the procurement, storage, access and use of information in health and biomedicine.

One of the first attempts to build a decision support system was MYCIN. [16] It was a simple question-answer, text based system that identified what dosage of a given medicine for bacterial infections the patient should be given, based on the observations inputted by the medical observer. The system was a brilliant idea but unfortunately, it was not robust enough to use it practically. Significant progress has been made since the days of MYCIN though.

In the past two decades, artificial neural networks are being used for creating decision support systems, due to the much higher degrees of accuracy and learning abilities that such systems possess. Several universities and hospitals around the world have ongoing projects on various diseases and disorders that use artificial neural networks to predict the possibility of their occurrence. [18] Other researchers and institutes around the globe have suggested the use of demographic data, various tests' data, etc. to draw inferences regarding the possibility of occurrence and reoccurrence of diseases/disorders like narcolepsy, lung cancer, obstructive sleep apnea, etc. [7]

2.3. Autoassociative Ann Model

Autoassociative memories or attractor neural networks emulate the temporal visual memory processing in the human brain. They store the patterns of the inputs they receive and can recall that pattern when they receive as input a fragment of that pattern. In effect, they work as short term, content addressable memories.

This network is implemented using recurrent collateral synapses. Using the Hebb rule for learning and in a completely connected network, this ensures that the synaptic weight matrix is symmetric. Each neuron i is made to fire in such a way that its output value r_i is determined by the external input e_i. This happens during the learning phase. During the recall phase, given an external input e_i, the result of the firing of the neuron is fed back through recurrent collateral axons. This feedback, called internal recall, is combined with the external input e_i to produce the pattern r_i. Here, the activation function on the feedback loop must be non-linear.

The autoassociative memory learning algorithm is often called "one-shot", because the network can learn a pattern when presented with a single instance of that pattern. In recall, given a fragment of a pattern, the input circulates through the network and in each loop, the entire pattern is recalled with more clarity. The network is also able to generalize a new pattern based on similarities to previously learned patterns.

Thus, overall, autoassociative memories represent a powerful new model of artificial neural networks.

2.4. ANN Based Classifier For Dengue

The task at hand is pattern recognition using numerical data from the CBC report of various patients, where for each data set the target values are known. The target value has the domain [0,1], where 0 represents a dengue negative case and 1 represents a dengue positive case. A simple feedforward neural network with backpropagation learning is used for the job.

3. Methodology

3.1. Data Collection Procedure

Building the best possible neural network, for any given situation, is a data intensive task. The need for accuracy is especially more pronounced when the problem domain is medical science, because millions of lives are at stake. Hence, our first task was collection of reliable medical data.

The data was collected from Sarvodaya Hospital, Ghaziabad, Uttar Pradesh, India. The CBC report data is usually recorded in registers by the hospital pathology lab staff. The cases ranged from January 2005 to January 2015.

Overall, 1209 cases of dengue suspected patients were recorded. The TLC, PCV, Hb and Pt count from the CBC reports of these patients were taken. Of these 1209 cases, 643 cases of suspected dengue were ultimately proven to be negative by the confirmatory serology test. The remaining 566 cases of suspected dengue were proven to be positive by the serology test. From this data, 1000 cases were used for training and the remaining 209 cases were used for testing.

The following four parameters were chosen for use in the neural network. These are the same parameters that the doctors check from the CBC report during the first phase of dengue diagnosis.

- Haemoglobin count
- Total leucocyte count
- Packed cell volume
- Platelet count

3.2. System Design

We have constructed three different neural networks, corresponding to original data, {PCA1, PCA2, PCA3}, and {NLPCA1, NLPCA2} as input vectors, respectively. Data representation in reduced dimensions with a high degree of information retention and efficiency in the neural network constructed from the reduced dimensions, is always a preferable occurrence – although this exercise is not strictly required in this specific study since there are only 4 variables in the input feature vector set. PCA reduces the given n-dimensional data into a straight line. NLPCA reduces the given n-dimensional data to a curve.

Positive cases of dengue are represented by 1, and the negative cases are represented by 0. The number of hidden layers and the number of neurons in them depends on the application. Since there are atmost 4 features in the input data vector and only one output, the number of hidden layers is kept as one. The number of neurons in the hidden layer are kept at either (number_of_input_features + 1) or (number_of_input_features + 2).

3.2.1. Control Checks

Control checks were applied in the form of a group of 50 cases of non-dengue infected, healthy human beings. The TLC, PCV, Hb and Pt counts vary as below in healthy persons:

Parameter	**Male**	**Female**
Haemoglobin (g/dl)	13.0-17.5	11.5-15.3
Total Leucocyte	4-11	4-11

Count ($x10^9$/L)		
Platelet ($x10^9$/L)	150-400	150-400
Packed Cell Volume (%)	42-50	36-45

Table 4.1.

If the TLC is between 1800 and 4000, the Hb is between 9,000 and 15,000, and the platelet count is less than 70,000 for a given patient, then the patient is suspected to be suffering from dengue. But these variations in the given parameters can also be exhibited in the case of other diseases like viral infections, thrombocytopenia, etc.

3.2.2. Features

As described in the preceding section, four primary features – haemoglobin, total leucocyte count, platelet count and packed cell volume - were selected to build the first ANN model. Dimensionality reduced forms of the primary feature vector have been used in the other two neural networks.

3.3. Preprocessing

The raw data procured from various sources is often not clean, formatted and organized enough to begin the main data processing tasks. Also, before beginning with the building of the neural network, we would like to know more about the statistical organization of the data to better guide decisions and techniques to be used in the latter part of the project. Hence, the following processing steps have been performed on the raw medical data procured from Sarvodaya Hospital, Ghaziabad.

3.3.1. Normalization

Since the data for the four parameters had a widely varying range, we normalized all four parameters to lie within the range of 0.2 to 0.8.

If a parameter A has a higher range than a parameter B, and both parameters are a part of the input vector for a given neural network, then it is highly likely that the network will be more biased towards parameter A than towards parameter B even if in the true model of the system B has more importance than A. We also observed this phenomenon during the training phase of the neural network model, whenever the non-normalized data set was used. Hence all four parameters were normalized to lie in the same range.

The formula used for normalization is the simple z-score measure :

DOI: 10.18535/ijecs/v6i7.04

$$xnorm = \left(\frac{x - xmin}{xmax - xmin}\right) * (newmax - newmin) + newmin$$

or,

$$xnorm = \left(\frac{x - xmin}{xmax - xmin}\right) * 0.6 + 0.2$$

3.3.2. Correlation analysis

The value of the correlation matrix of the 4 parameters is shown below :

	Hb	TLC	PCV	Pt
Hb	1	-0.1269	0.9406	-0.1041
TLC	-0.1269	1	-0.602	0.5458
PCV	0.9406	-0.602	1	0.0408
Pt	-0.1041	0.5458	0.0408	1

These values indicate that there is not a heavy interdependence among the individual values of each of the four parameters, except between PCV and Hb. But PCV and Hb show low correlation to the other two parameters. That is, each of the four parameters can contribute non-overlapping, unique information to the neural network.

3.3.3. Principal Component Analysis

The original data set was subjected to principal component analysis to see whether a compressed form of the data could contribute the same or nearly the same amount of information as the original data.

Three PCAs give information retention of 99.1492% cumulatively.

DOI: 10.18535/ijecs/v6i7.04

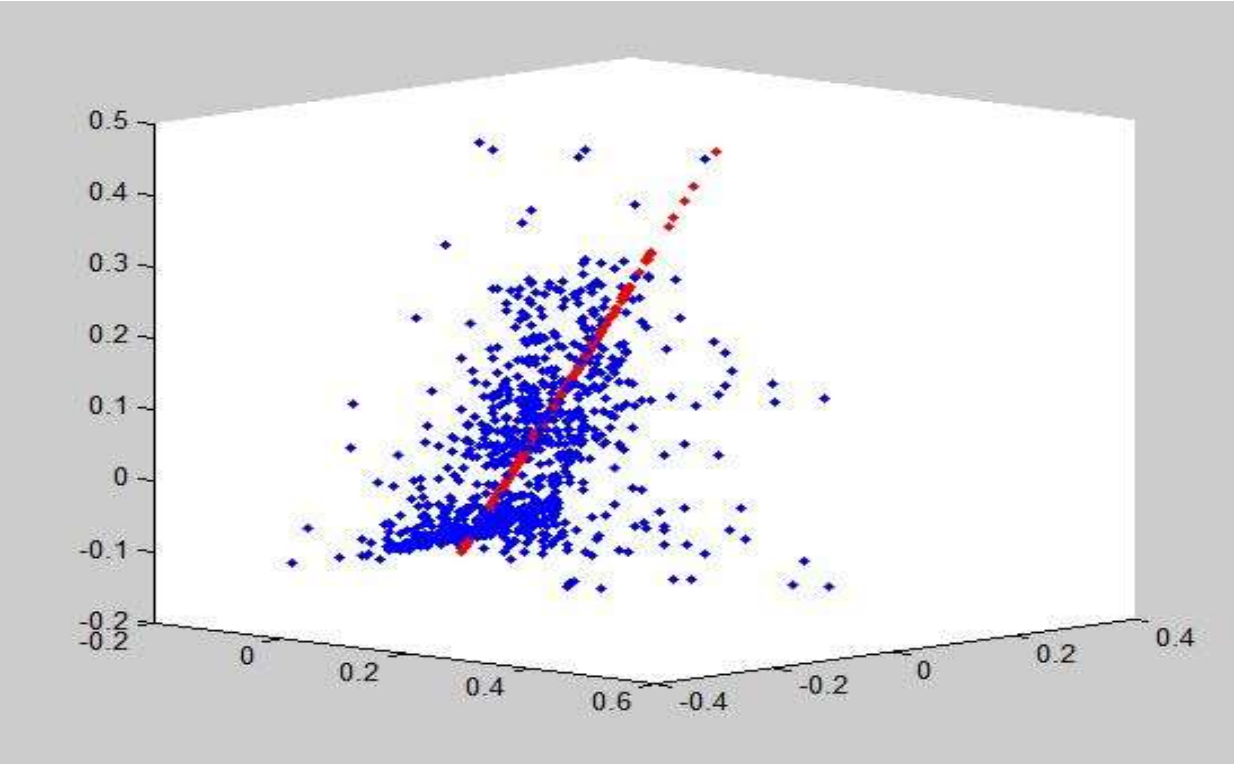

Figure 4.1.2 : Plot showing original data in blue and PCA1, PCA2, PCA3 in red.

The information gain through PCA1 was 51.9130%, 33.9543% through PCA2 and 13.2819% through PCA3. Thus, to build an effective classifier using PCA components as input, at least PCA1, PCA2 and PCA3 together would have to be used. This does not offer much in the way of dimensionality reduction, since the original data itself is of 4 dimensions.

The scatter plots of PCA 1 and PCA 2, and of PCA 1, PCA2, PCA3 are shown below, colour coded black for positive dengue cases and red for negative dengue cases.

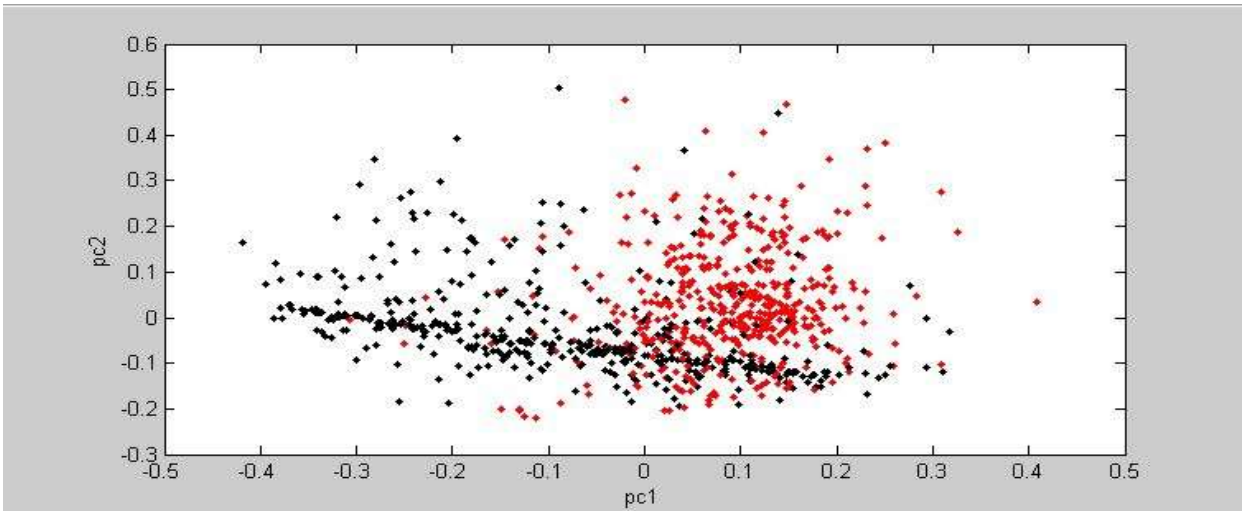

Figure 4.2.1 : PCA1 vs. PCA2. Positive cases are shown in black, negative cases in red.

DOI: 10.18535/ijecs/v6i7.04

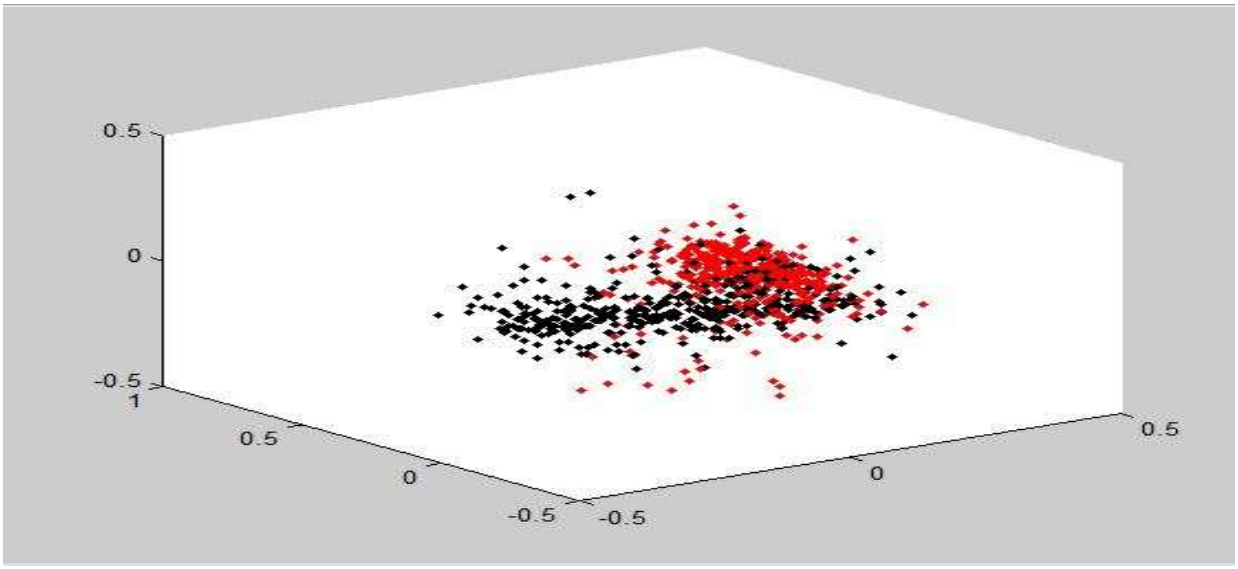

Figure 4.2.2 : PCA1 vs. PCA2 vs PCA3. Positive cases are shown in black, negative cases in red.

3.3.4. Non-Linear Principal Component Analysis

Many times, the data distribution is such that a straight line is not adequate to represent the data efficiently. Non-linear principal component analysis has the same concept has PCA, except that it tries to reduce data dimensionality to a curve instead of a straight line.

The dengue CBC data showed better response to dimensionality reduction through non-linear PCA, as is evident in the graphs below.

The first two non-linear principal components were evaluated, with an information retention of 74% and 24% respectively. Thus, the 4-dimensional dengue data can be represented through the first two non-linear principal components without a significant loss of information.

DOI: 10.18535/ijecs/v6i7.04

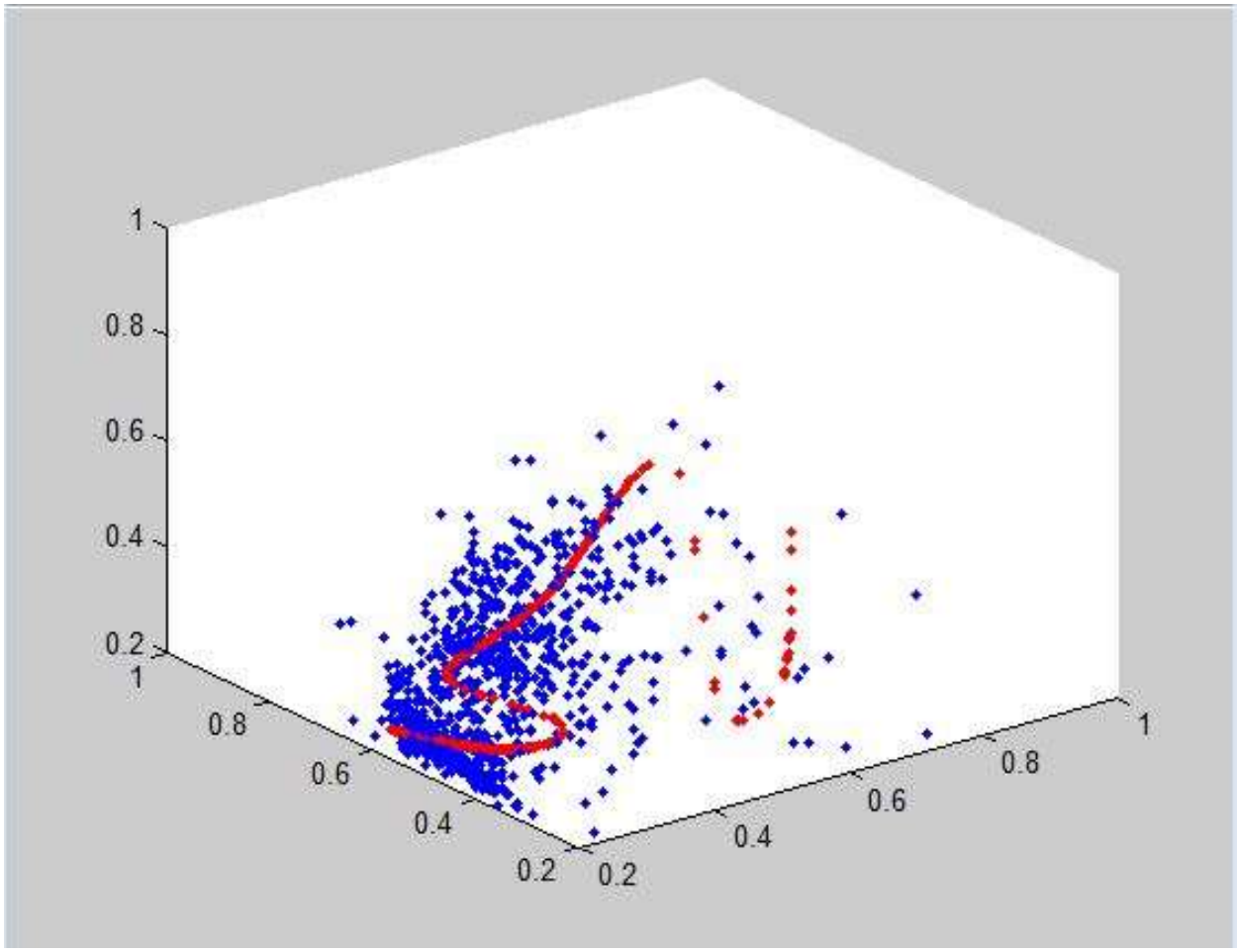

Figure 4.3.2 : Plot showing original data in blue and NLPCA1 and NLPCA2 in red.

3.3.5. NLPCA vs. PCA

What PCA could achieve in three dimensions, NLPCA achieved in just two dimensions. The information retention for (PCA1, PCA2, PCA3) and for (NLPCA, NLPCA2) is nearly the same at approximately 99% and 98%, respectively.

The following two graphs show how both NLPCA and PCA fit to the original data. As is evident, NLPCA provides better information retention during dimensionality reduction.

DOI: 10.18535/ijecs/v6i7.04

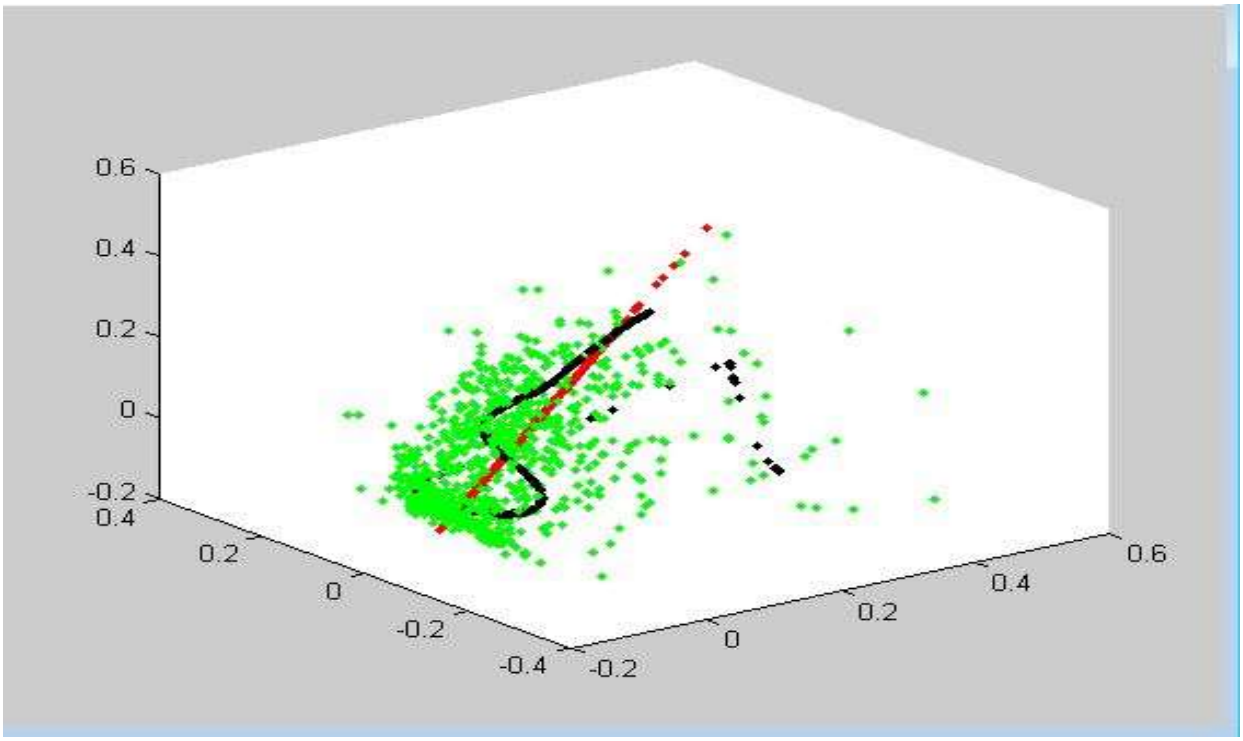

Figure 4.4.1 : Original data in green, NLPCA1 & NLPCA2 in black, PCA1 & PCA2 & PCA3 in red.

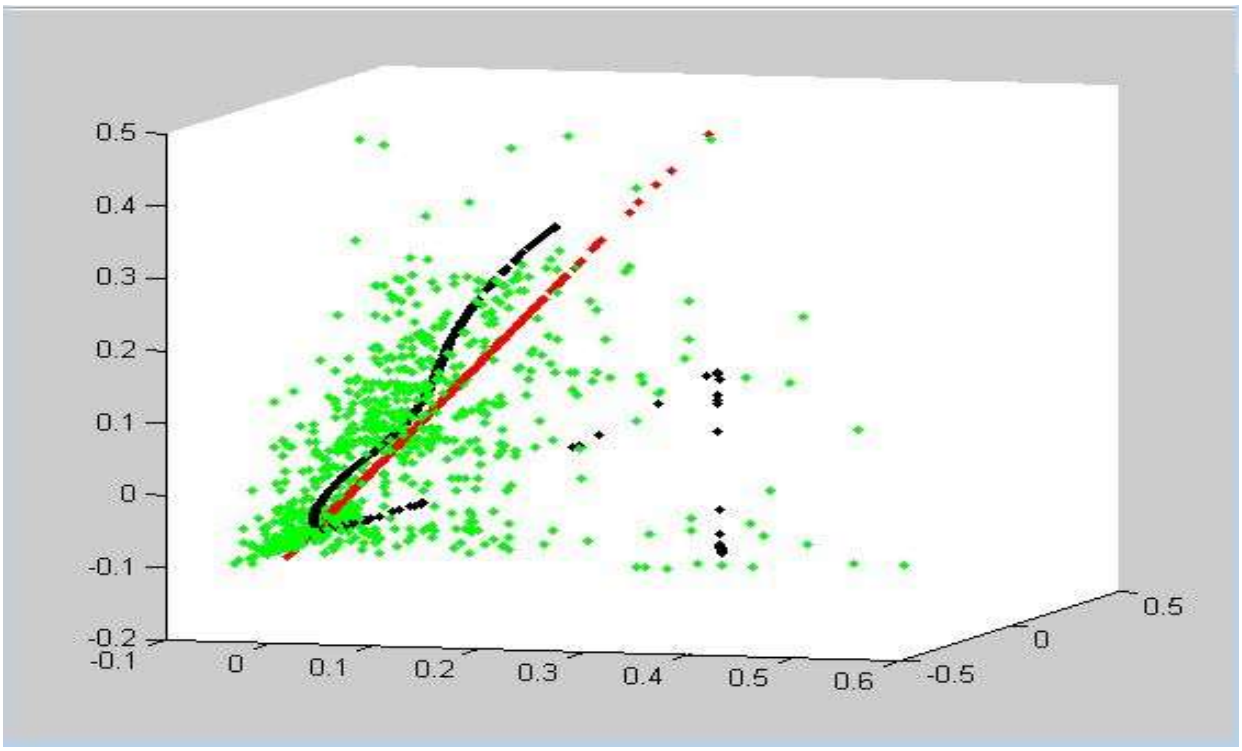

Figure 4.4.2 : Original data in green, NLPCA1 & NLPCA2 in black, PCA1 & PCA2 & PCA3 in red.

3.4. Training, Testing and Validation

The first neural network takes as input feature vector the four parameters TLC, PCV, Hb and Pt. For the second neural network, input sets with reduced dimensionality were used – more specifically, the first three

principal components of the original data set were used. The third neural network takes as input the first two no-linear principal components of the original data set.

3.4.1. ANN with original data

The first neural network constructed was the ANN with the original data set as input. Its structure and performance charts are as below.

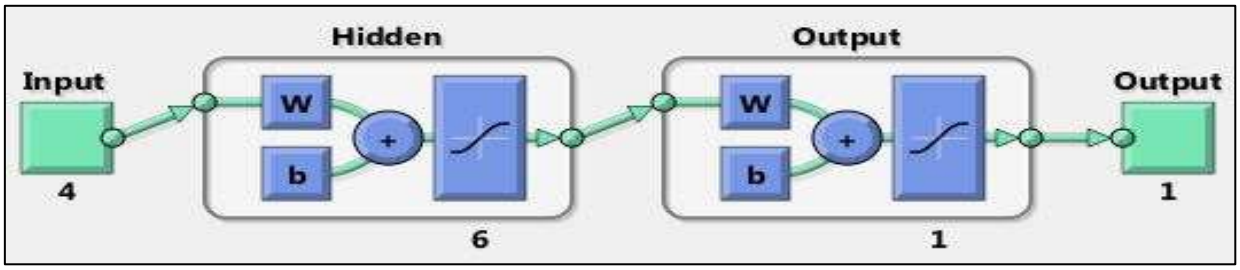

Figure 4.5.1 : Neural Network 1 Structure

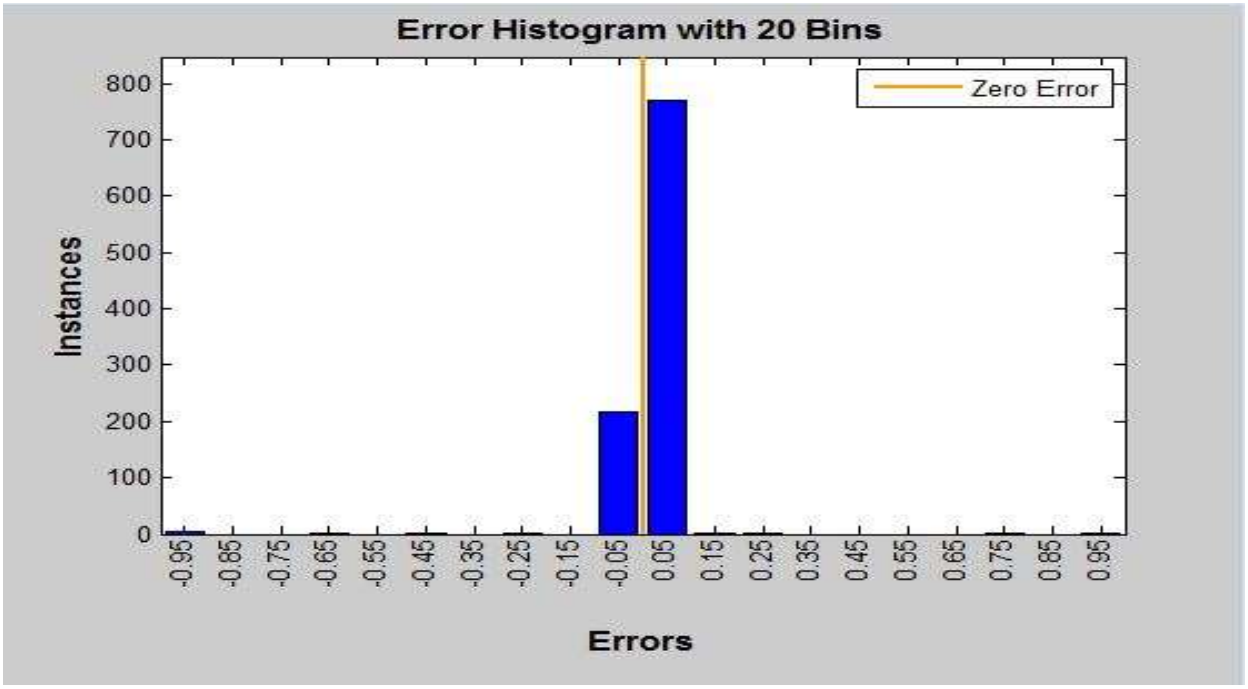

Figure 4.5.2 : Network 1 Error Histrogram

DOI: 10.18535/ijecs/v6i7.04

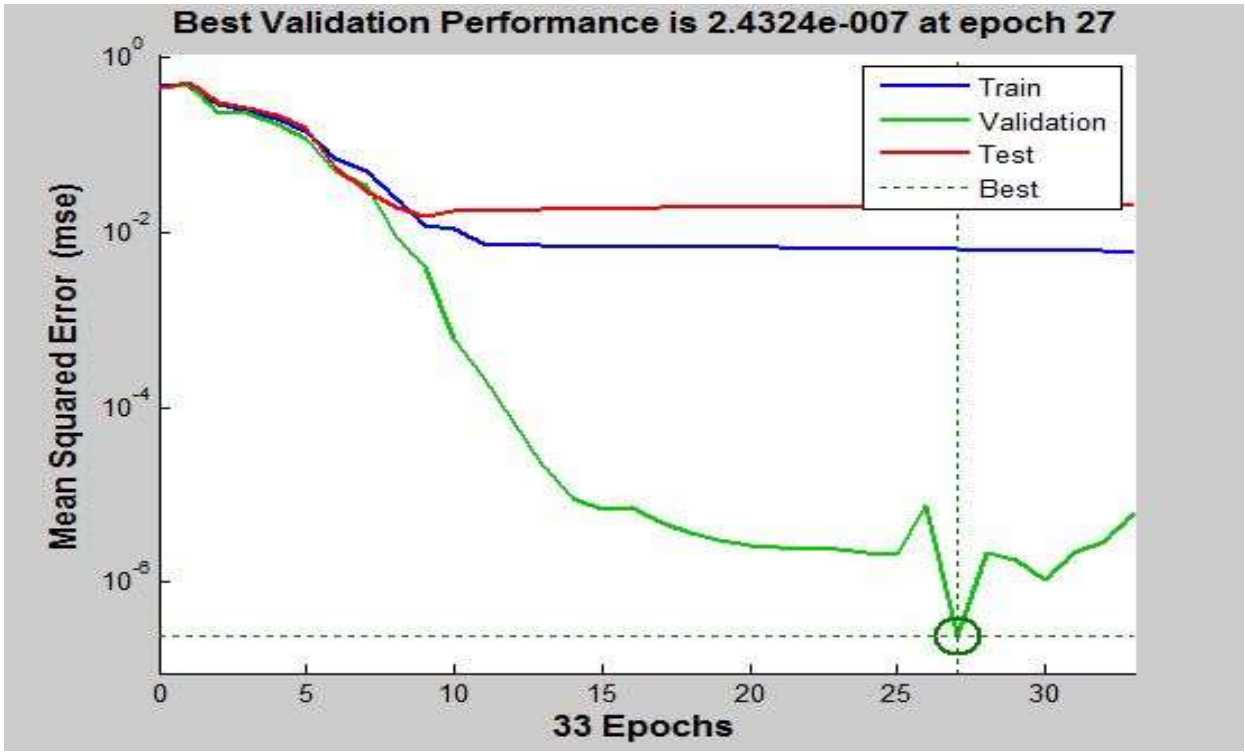

Figure 4.5.3 : Network 1 Convergence

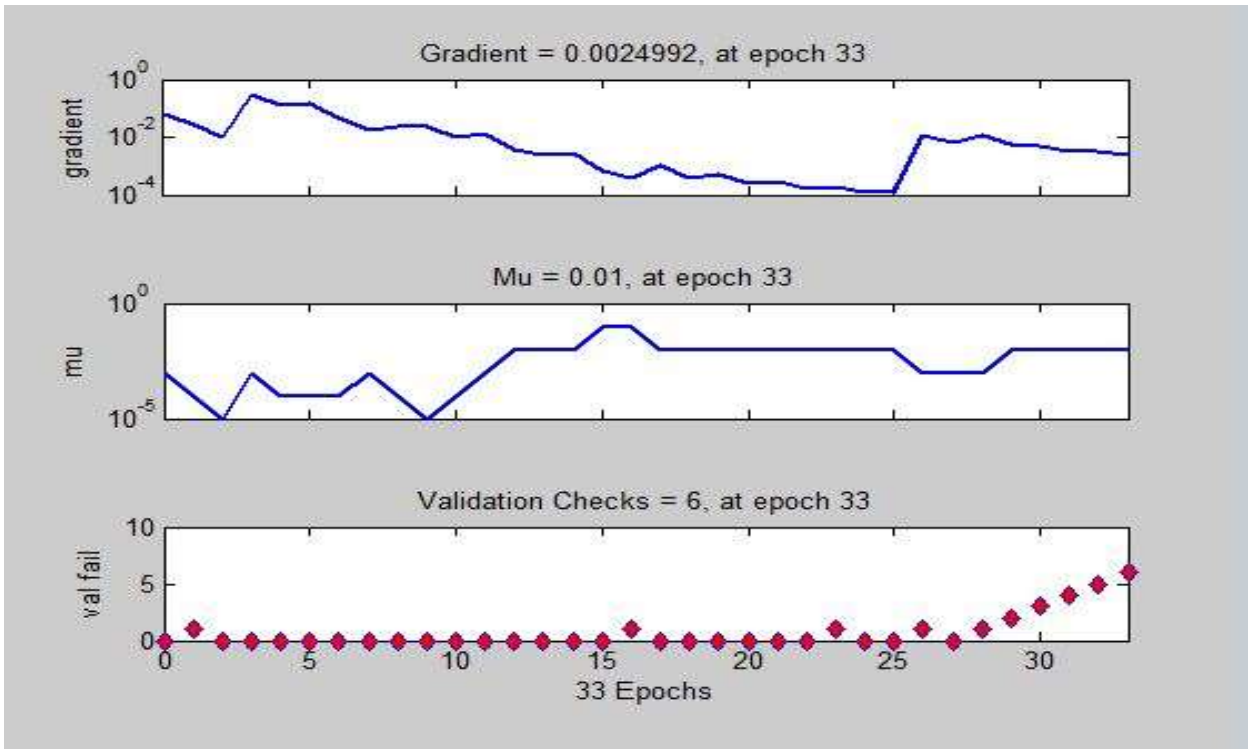

Figure 4.5.4: Network 1 Epochs and Gradients

3.4.2. ANN with PCA data

The second neural network constructed was the ANN with the first three principal components as input vector. Its structure and performance charts are as below.

DOI: 10.18535/ijecs/v6i7.04

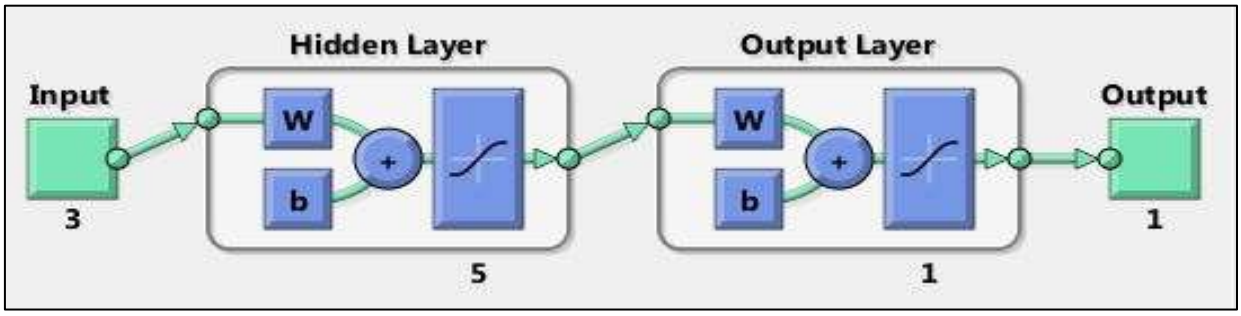

Figure 4.6.1 : Neural Network 2 Structure

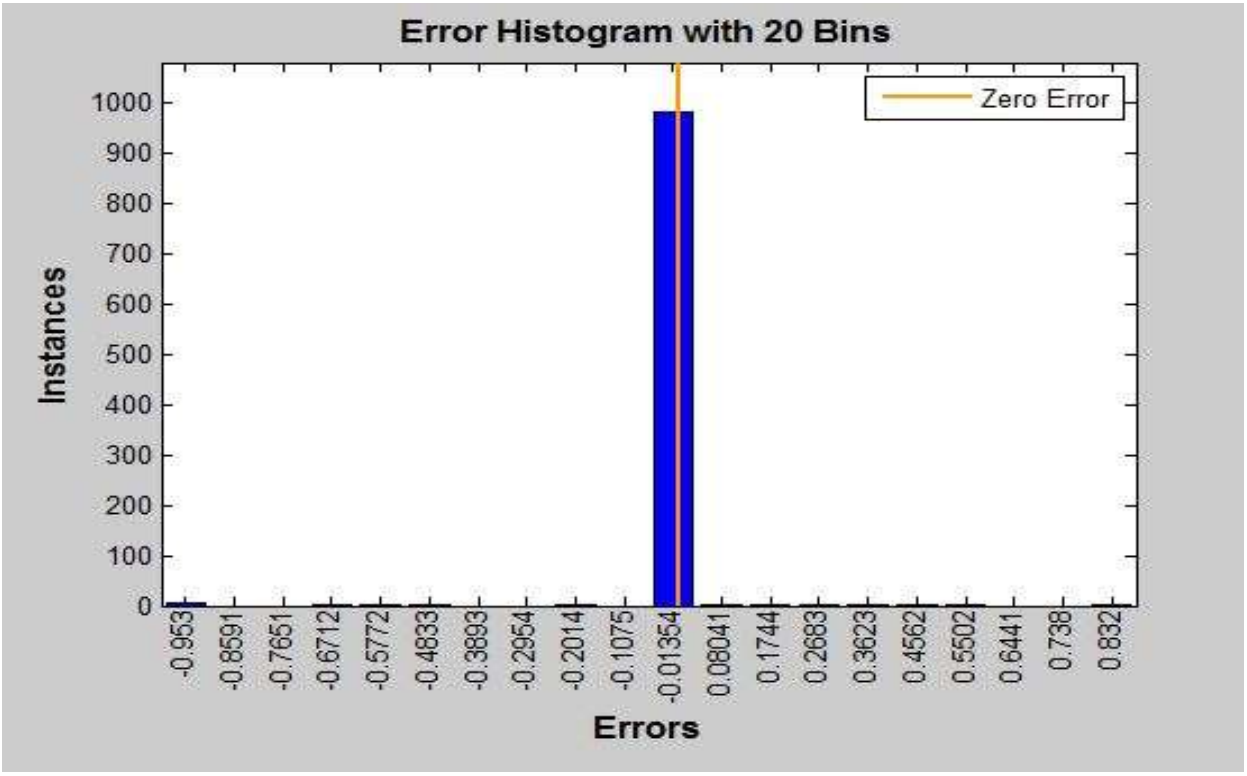

Figure 4.6.2 : Network 2 Error Histogram

DOI: 10.18535/ijecs/v6i7.04

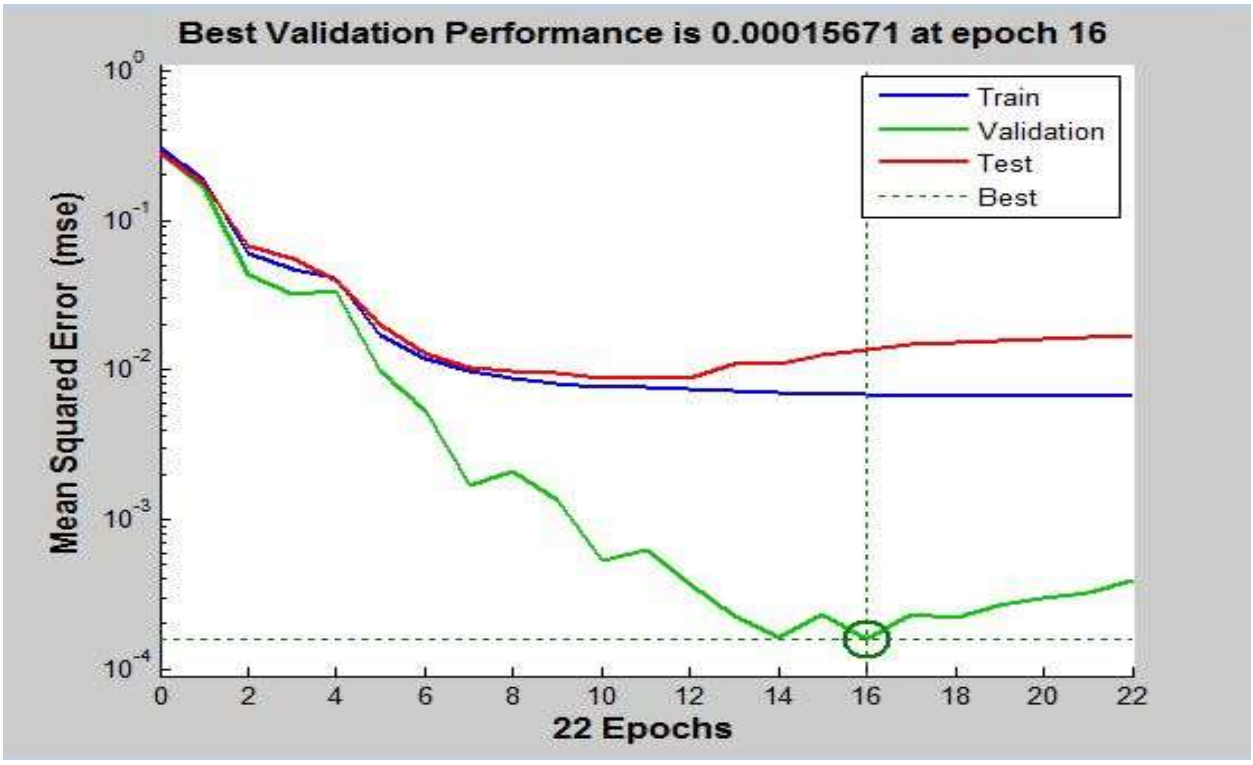

Figure 4.6.3 : Network 2 Convergence

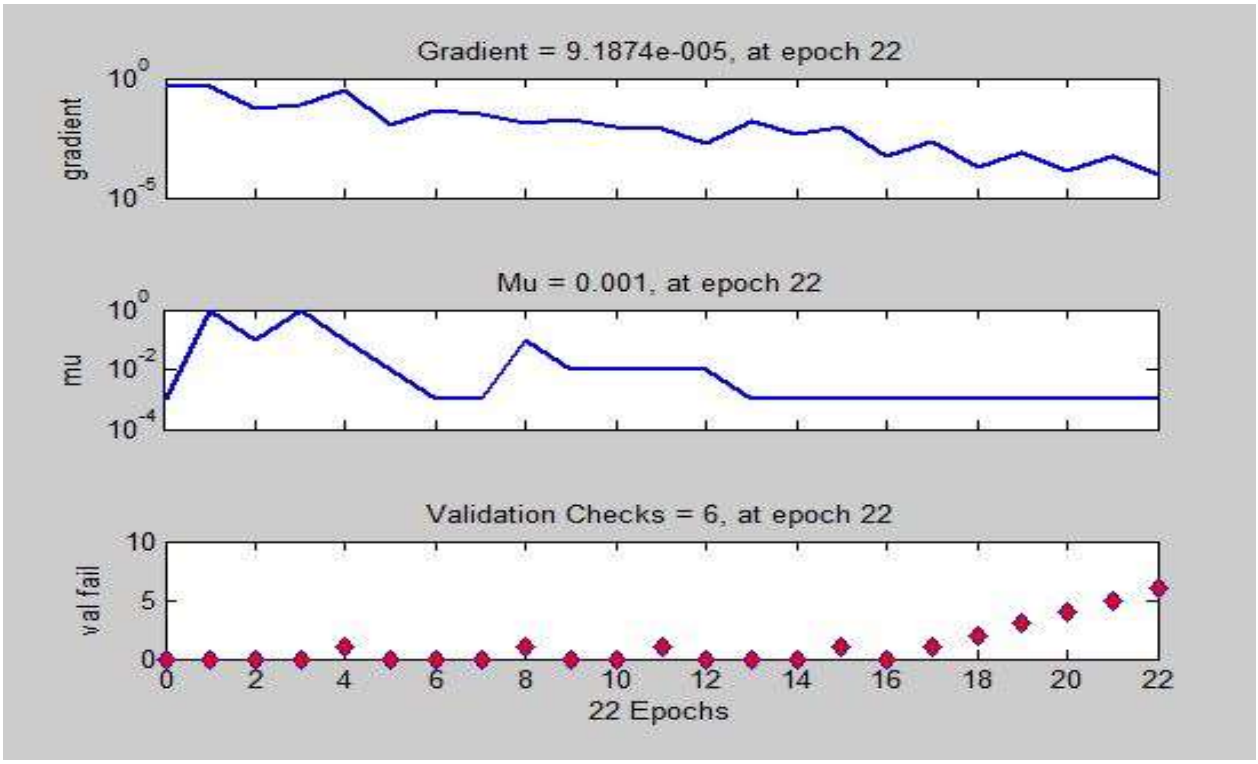

Figure 4.6.4 : Network 2 Epochs and Gradients

3.4.3. ANN with NLPCA data

DOI: 10.18535/ijecs/v6i7.04

The third neural network constructed was the ANN with the first two non linear principal components as input vector. Its structure and performance charts are as below.

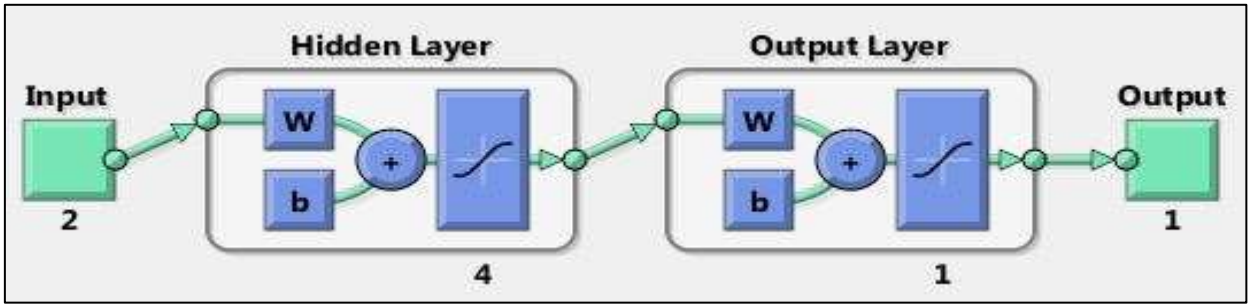

Figure 4.7.1 : Neural Network 3 Structure

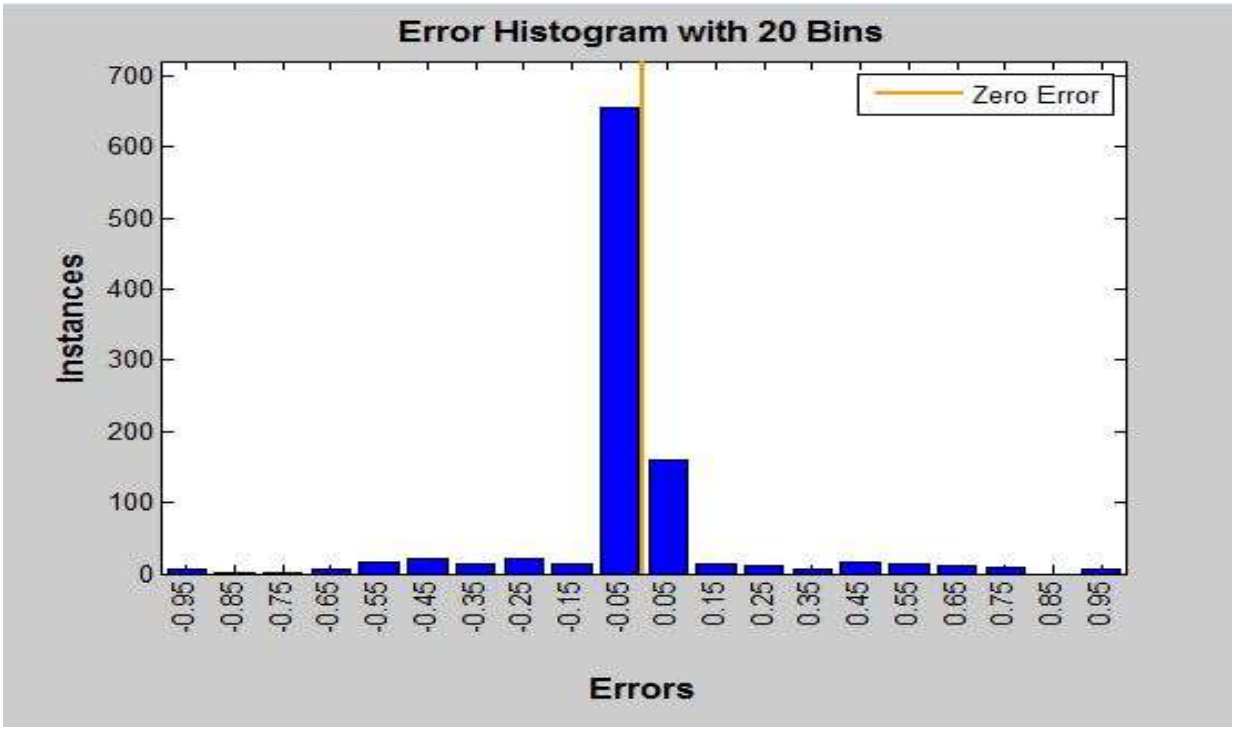

Figure 4.7.2 : Network 3 Error Histogram

DOI: 10.18535/ijecs/v6i7.04

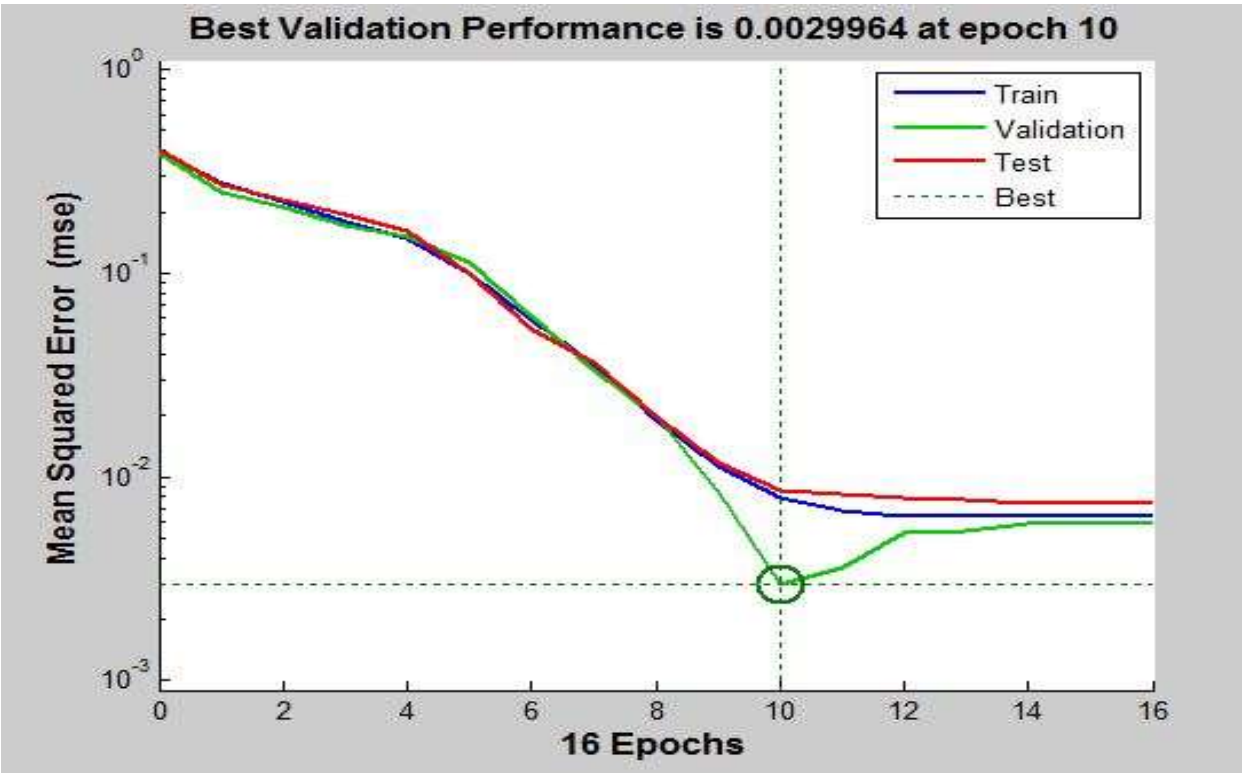

Figure 4.7.3 : Network 3 Convergence

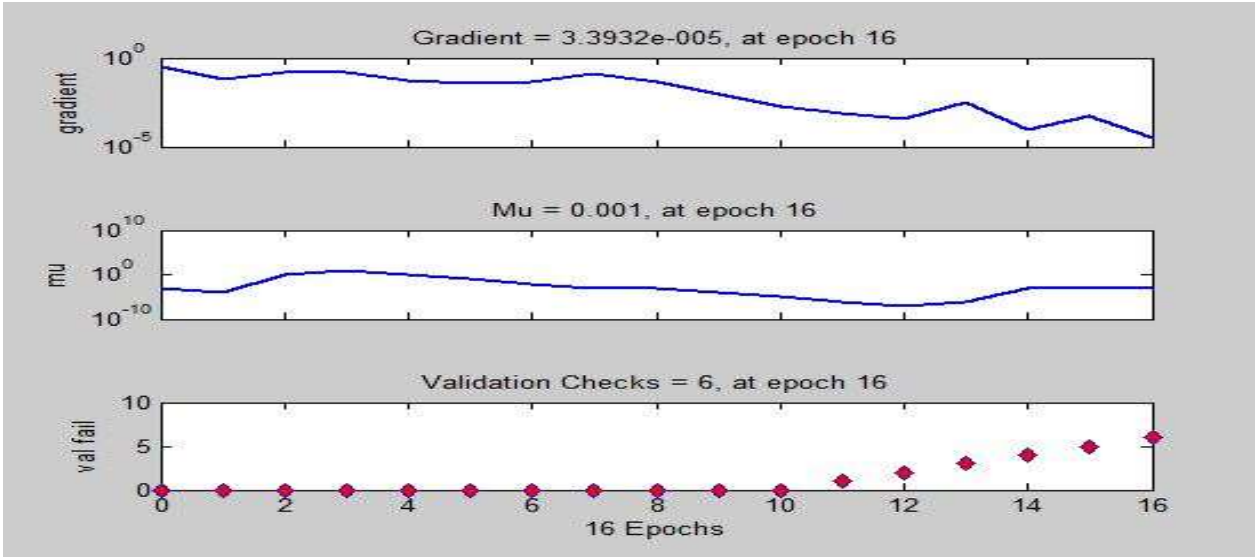

Figure 4.7.4 : Network 3 Epochs and Gradients

4. Results and Discussion

Overall, we have found that all three neural networks have been very successful in the diagnosis of whether a given patient has dengue.

4.1. ANN Efficiency

DOI: 10.18535/ijecs/v6i7.04

The neural network with original data set has an efficiency of 99.6%. The neural network with the PCA as input vector has an efficiency of 99.2% too. The neural network with NLPCA as input vector has an efficiency of 92.9%.

Overall, the best useable neural network is the third network – because it gives a good performance with just half the number of features present in the original data set.

We have taken the first neural network as the practically used network in the decision support software implementation though, because the number of features is not very large and it has the highest efficiency.

4.2. Hit And Miss Analysis

The following was the result of the hit and miss analysis of the outputs of the first neural network on a test data set of 209 cases,

	Positive Diagnosis By NN	Negative Diagnosis By NN
Patient has dengue	58	9
Patient does not have dengue	0	142

Table 5.2.

- Hit Ratio = 0.9569
- Miss Ratio = 0.0431
- False Positive Ratio = 0
- True Positive Ratio = 0.8656
- False Negative Ratio = 0.1343

There is a need to bring down the false negatives even further – in this scenario, falsely rejecting patients with dengue can prove to be a fatal mistake. Overall though, the performance of the neural network is deemed to be quite good.

5. Conclusion and Future Scope

Thus, the two-step diagnostic process of dengue is effectively reduced to a one-step process with the help of this classifier. The model proposed has a test set accuracy of nearly 95%. It eliminates the need for the more time consuming, but truly confirmatory serology test. Time being a crucial factor in treatment of dengue, this decision support software system thus has the potential to help doctors save many more lives.

Further refinements to the dengue classifier include the trial and use of more parameters apart from the four already used here. The classifier should also be made diverse enough to classify diseases that are similar to dengue based on physical symptoms and lab test report readings.

We further propose as future research directions the application and comparison of the modelling results of more pattern recognition techniques for this classification task, testing of the system in real-time hospital conditions, and the inclusion of locality specific factors to build as general and as widely-reproducible a model as possible

Acknowlegements

We acknowledge the help and guidance in terms of medical expertise provided by the esteemed doctors at Sarvodaya Hospital, Ghaziabad and by Mr. Kapil Tyagi. The pathology lab data was provided by Sarvodaya Hospital, Ghaziabad, on the strict condition of total patient-identity anonymity.

References

[1] N Ganesan, K Venkatesh, M A Rama, Malathi A Palani, "Application of Neural Networks in Diagnosing Cancer Disease using Demographic Data".

[2] Miss. Manjusha B. Wadhonkar 1, Prof. P. A. Tijare 2 and Prof. S. N. Sawalkar 3, "Classification of Heart Disease Dataset using Multilayer Feed forward backpropogation Algorithm".

[3] Shraddha Srivastava 1, K.C. Tripathi 1, "Artificial Neural Network and Non-Linear Regression: A Comparative Study", (1 Inderprastha Engineering College, Ghaziabad)

[4] Jayanta Kumar Basu 1, Debnath Bhattacharyya 2, Tai-hoon Kim 2, "Use of Artificial Neural Network in Pattern Recognition", (1 Computer Science and Engineering Department, Heritage Institute of Technology, Kolkata, India; 2 Multimedia Engineering Department, Hannam University, Daejeon, Korea)

[5] Dr. Rama Kishore, Taranjit Kaur, "Backpropagation Algorithm: An Artificial Neural Network Approach for Pattern Recognition".

[6] Ms. Sonali. B. Maind 1, Ms. Priyanka Wankar 1,"Research Paper on Basic of Artificial Neural Network"*(1 Datta Meghe Institute of Engineering, Technology & Research, Sawangi (M), Wardha).*

[7] http://www.ijcaonline.org/archives/number26/476-783

[8] http://www.ncbi.nlm.nih.gov/pubmed/20639791

[9] http://informahealthcare.com/doi/abs/10.1586/eri.10.53

[10] Simon Haykin, (2001) "Neural Networks – A Comprehensive Foundation", Pearson Education.

[11] Hagan, M.T., Demuth H.B., Beale M.H. 1997. "Neural Network Design, PWSpublishing", Boston, M

[12] Bishop C.M. 1997. "Neural Networks for pattern Recognition", Oxford University Press, New York.

RESOURCES

[1] http://www.cs.cf.ac.uk/Dave/AI1/mycin.html

[2] http://in.mathworks.com/products/matlab/

[3] http://www.hopkinsmedicine.org/Research/johns_hopkins_research_topics.html