Optimizing Input variables for the Artificial Neural Network Model Using Genetic Algorithm

Chidubem Damian Dibie

Damiandibie@gmail.com

Departmentof Civil Engineering

University of Johannesburg, South Africa.

ABSTRACT

ANN models are known to give satisfactory results when it comes to simulation and ultimate prediction of events irrespective of the process as far as there are sufficient historical data which the machine learning technique can simulate. However, the process can be modified such that the ANN model can even give better results than it is giving presently; Genetic Algorithm Technique in WinGAmma software was utilized in analysing the dataset in order to select the most relevant combination of input variables from the dataset which will produce better performance of the ANN model and the output was compared with the normal ANN model which utilized all the input variables in the dataset and the GA-ANN model outperformed the ANN model with all input variables, which goes on the show the GA technique is very reliable and a welcome technique in selecting input variables for the optimal performance of the ANN model in simulation and prediction of events. The event simulated in this work is the Daily suspended sediment load with several input variables ranging from present and antecedent parameters of Discharge, Sediment Discharge and Suspended Sediment load. Although the result of the ANN model with all 7 input variables was satisfactory, but the GA-ANN model gave a much better result with lesser but more relevant input variables as selected by the GA technique.

Keywords: Artificial Neural Network (ANN), Genetic Algorithm (GA), nodes, hidden layer, input variable.

INTRODUCTION

The cardinal objective of creating the Artificial Intelligence models majorly the Artificial Neural Network (ANN) is sourced from envisioning a technique that can imitate the Human brain in making a well-informed decision based on knowledge of the process in question (Oyebode and Stretch, 2019). Over the past decade the adoption of Artificial Intelligence technique has increased exponentially and more readily used in all aspect of life which is typified in hydrological event description (Dibie, 2019). Water resources management has received adequate attention from the artificial intelligence techniques; prediction of

hydrological events such sediment load, sediment discharge, water quality etc) have been marred by the nonlinearity and interaction by different parameters due to spatial and temporal variations. ANN is known to be able to study pattern between inputs and outputs without any clarity of relationship (Hence the black box model) and give a very impressive result (Mustafa *et al.,* 2012).

More so, one of the strengths of Artificial Intelligence techniques is the ability to combine several techniques thereby making up for each other's deficiencies and making a more improved estimation that their individual usage (Nourani, 2014). More so, AI models have been very resourceful in analysing events whose mechanisms are very complex and not easily describable which is very conspicuous in hydrological processes (Chen *et al.,* 2008).

In this work, the hydrological process to be described is the prediction of suspended sediment load in a dam using hybrid of ANN ad GA; whereby GA will optimize the relevant input data set required to for the ANN model to make reliable predictions; this result will be compared with using all input variables to train and also for prediction with the ANN model to know which will give a better result.

1.1 ARTIFICIAL NEURAL NETWORK

ANN are models that imitate the operation of the human brain through the simulation of the biological network of neurons in the brain to transmit information to and from the brain and other parts of the body. ANN are data driven models that seek patterns and not assumptions in the events or processes such as forecasting (Besaw and Rizzo, 2007; Jain *et al.,* 1996).

The typical architecture of the ANN layout comprises three layers namely: Input where data is fed into the network, Output where the desired results required is gotten and the Hidden layer where the complex processes such learning, pattern detection and others are done in the network for the purpose of simulation and forecast of the ANN model. Each respective neuron in each layer is interconnected to all neurons in the next layer with a link with weights of different magnitudes as determined during the training/learning process which implies their level of influence and interrelationship but no neuron in the same layer are linked together (Yesilnacar *et al.,* 2008)

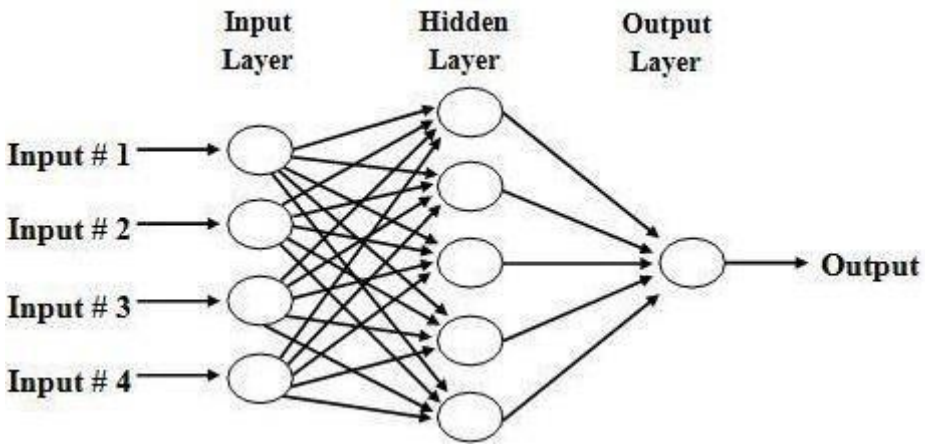


Figure 1: ANN Typical Structure (Chen *et al.,* 2008)

The most crucial part of ascertaining the type of ANN to be adopted is dependent on the ANN architecture (Palani *et al.,* 2008; Starrett *et al.,* 1998). Sequel to the architecture being the most crucial aspect in the efficacy of the ANN model in terms of weight appropriation in the neurons, information transmission through the network, it is of great importance to give much attention to the architecture while developing the model; it has also proven to be the most difficult aspect to be accomplished while constructing the ANN model (Chen and Chang, 2009; Singh and Datta, 2007).

More recently has seen ANN models being combined with other techniques to give better results, even at that using ANN model as a stand alone has yielded reliable and impressive results across different areas of usage such as prediction of Uniaxial Compressive Strength (UCS) of rocks using the back propagation neural network and even elucidating the interrelationship between different parameters relevant to the study (Ferentinou and Fakir, 2017). However, there have been a growing usage of ANN being combined with several other models or techniques to bring about better results.

1.2 GENETIC ALGORITHM

Algorithm is a series of steps undertaken to solve a problem (Saini, 2017). Genetic algorithm was first created by Holland (Holland, 1992) and then subsequently Goldberg (1989) introduced it as an evolutionary algorithm. GAs are very adaptive and resourceful in a variety of applications for problems of optimization. The philosophy of the GA is based on genetic mechanisms of living organisms. Across several generations, evolution of populations are based on the ideology of selection and survival of the fittest concept. Initially this theory was first stated by Charles Darwin (1859). Genetic Algorithm has its application very useful in optimization problems. Goren *et al.* (2010) give steps to adhere to when applying GA as:

1. Choosing a scheme of representation with the intent of deriving a solution
2. How to constitute the initial population
3. Well defined fitness function
4. Concise elucidation of the genetic operators to be adopted such as: mutation, crossover, reproduction, elitism.
5. Ascertaining the relevant parameters like size of population
6. Termination rule

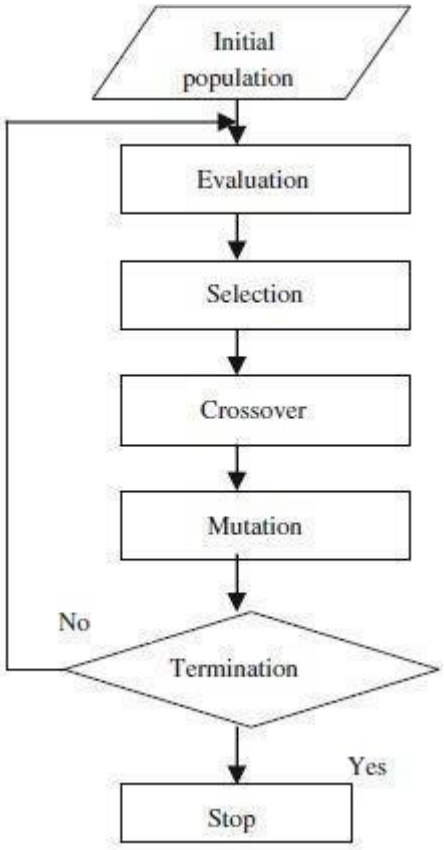


Figure 2: Flow Chart of GA

2.0 PROCESS OF MODEL DEVELOPMENT

2.1 PREPROCESSING AND HANDLING OF RAW DATA

It is of great importance to pre-process raw data as there may be variability embedded in the data set and by so doing in an efficient manner helps to reduce noise in data to the barest minimum. It was hence postulated that all variables in a data set need be

standardized to attain equality with respect to magnitude and ultimately experience expected attention in training phase (Maier and Dandy, 2000). Standardization is the rescaling of data to be enclosed within the activation function boundaries applicable in the output layer. Minns and Hall (1996) suggested that data scaling should be defined with the range being proportional to the bounds of activation function embedded in the output layer; if not then input variables documented at various magnitude orders can negatively impact training process (Dawson and Wilby, 2001).

Another credible technique in pre-processing of data is Normalization, this implies rescaling data to a Gaussian function. Maier and Dandy (2000) admonished against scaling of data to ranges at the extreme of the activation function as this can bring about reduction in the probable weight updates, consequently leading to flat spots in training. More so, Abrahart *et al.* (2012) suggested a an inquiry in the application of hydrological data, since an array of mixed results imply inputs may either give better or poor results than those gotten from the original without pre-processing.

2.2 DETERMINATION OF INPUT VARIABLES FOR THE MODEL

The fundamental step in making us of a model is not just to collect data but to ensure the variables that are relevant in having a positive impact in the model's description of the hydrological system. The data sets are always large and hence needs to be screened. As this will help in the selection of only relevant input variables that will account for proper representation of the process being described (Prasad *et al.,* 2017). More so, the usage of unrelated input variables somewhat increases network extent, model complexity, impedes learning and hence resulting in bad generalization (Bowden *et al.,* 2005). In this work, Genetic Algorithm will be used to determine which relevant input variables are relevant for the ANN model to describe Suspended sediment load.

3.0 STUDY AREA

The Shiroro dam is mainly used for hydropower and it is situated in the Shiroro Gorge estimated between $9^{\circ}46^{I}35.29^{II}$ and $10^{\circ}08^{I}36.65^{II}$ N latitude and $6^{\circ}50^{I}51.23^{II}$ and $6^{\circ}53^{I}14.53^{II}$ N longitude. The dam is located about 90km southwest of Kaduna, across River Dinya. Its power generating capacity is 600MW (Kolo, 1999). The impounded reservoir of the dam has a surface area of 320km^2 with a corresponding storage capacity of about 7 billion m^3 (Suleiman and Ifabiyi, 2015).

River Kaduna contributes about 70% inflow to the reservoir with other lateral inputs from Rivers Sarkin-Pawa, Muyi and Dinya (Adie *et al.,* 2012; Eze, 2005). The impounded reservoir is within the River Kaduna catchment in the Guinea Savanna zone of Nigeria. The climate of the region aligns with that of the whole country (Adegun *et al.,* 2018)

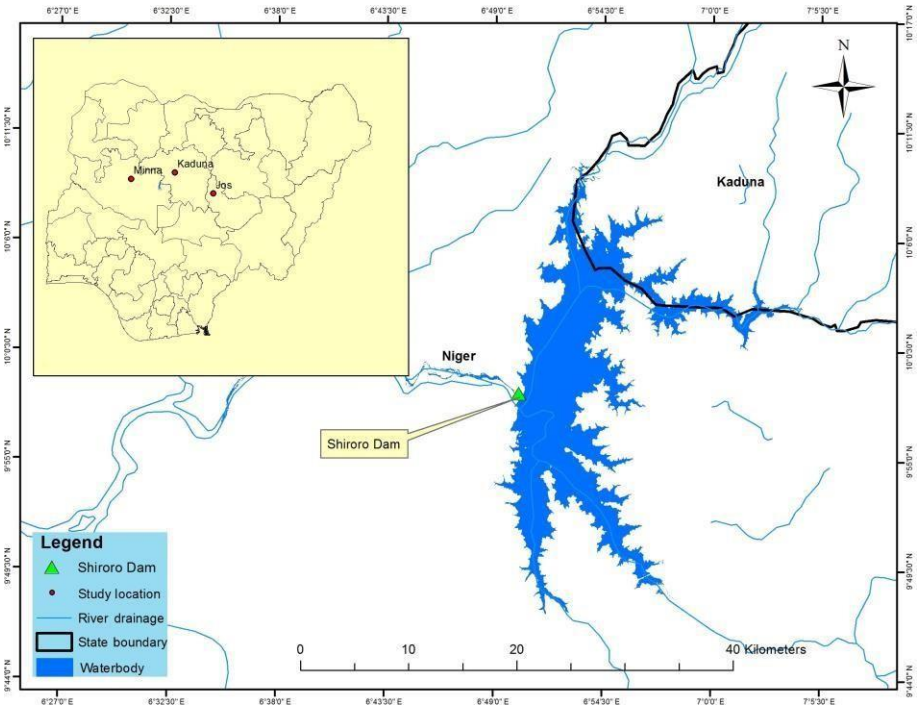


Figure 3: Location of Shiroro dam

MATERIALS AND METHOD

The data used for this article was sourced from the Nigerian Meteorological Agency (NIMET) which are daily records of sediment discharge, discharge and suspended sediment load for Shiroro dam.

The data used are the current and antecedence of discharge, sediment discharge and Suspended sediment load which was done for the ANN model and the GA-ANN model with details of their respective architecture as shown in the table of results below. The number of neurons used where selected by trial and error method as the variability in the data set impeded the use of any empirical formula or rule of thumb.

The data was normalized using excel within a range of 0 – 1. The data is then saved as a CSV format to be exported to WinGamma software for further analysis. Genetic Algorithm technique in WinGamma software was utilized to select the most relevant input variables for the ANN model. The data set was subjected a graph of Standard Error against Near Neighbour was used to select the suitable near neighbour as seen in the graph below:

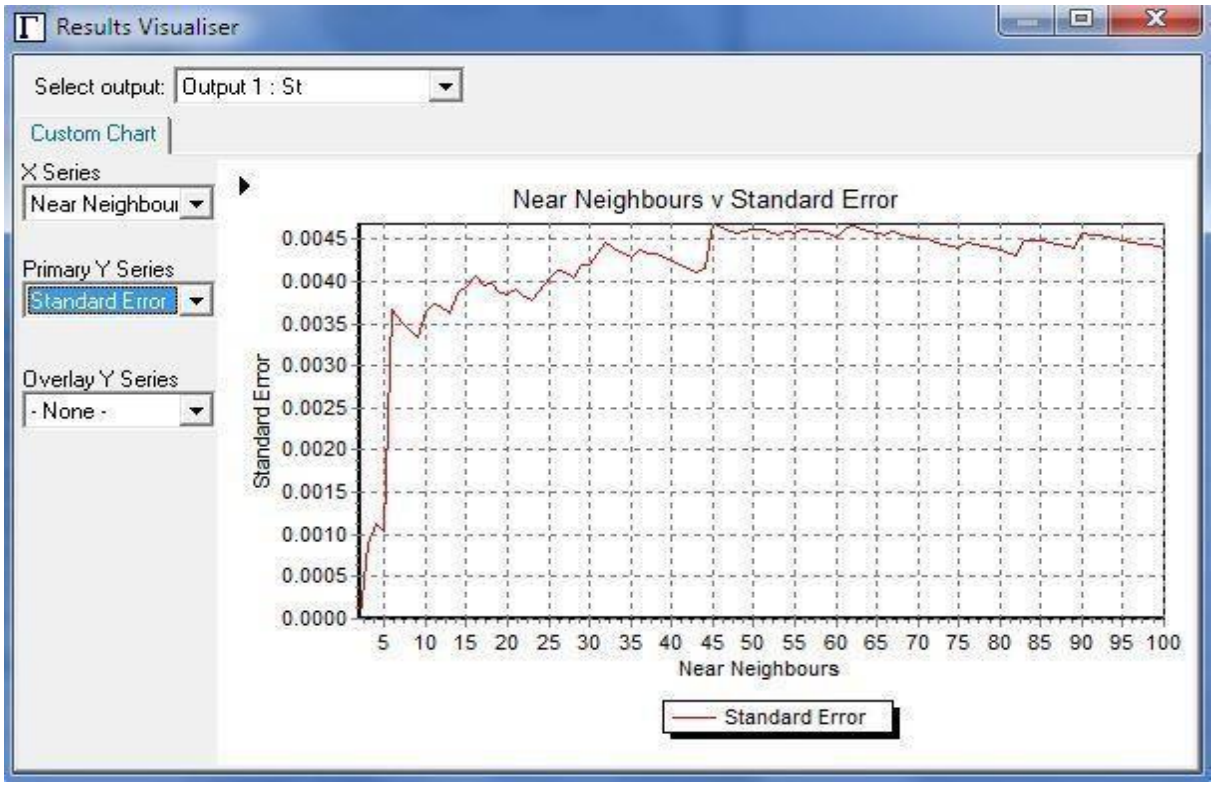


Figure 4: Plot of Standard Error against Near Neighbours in WInGamma

The value of the suitable of near neighbour from the plot above was used in configuring the genetic algorithm to ascertain which data set is relevant as input parameters. The output of the genetic algorithm for the Mask result is shown below:

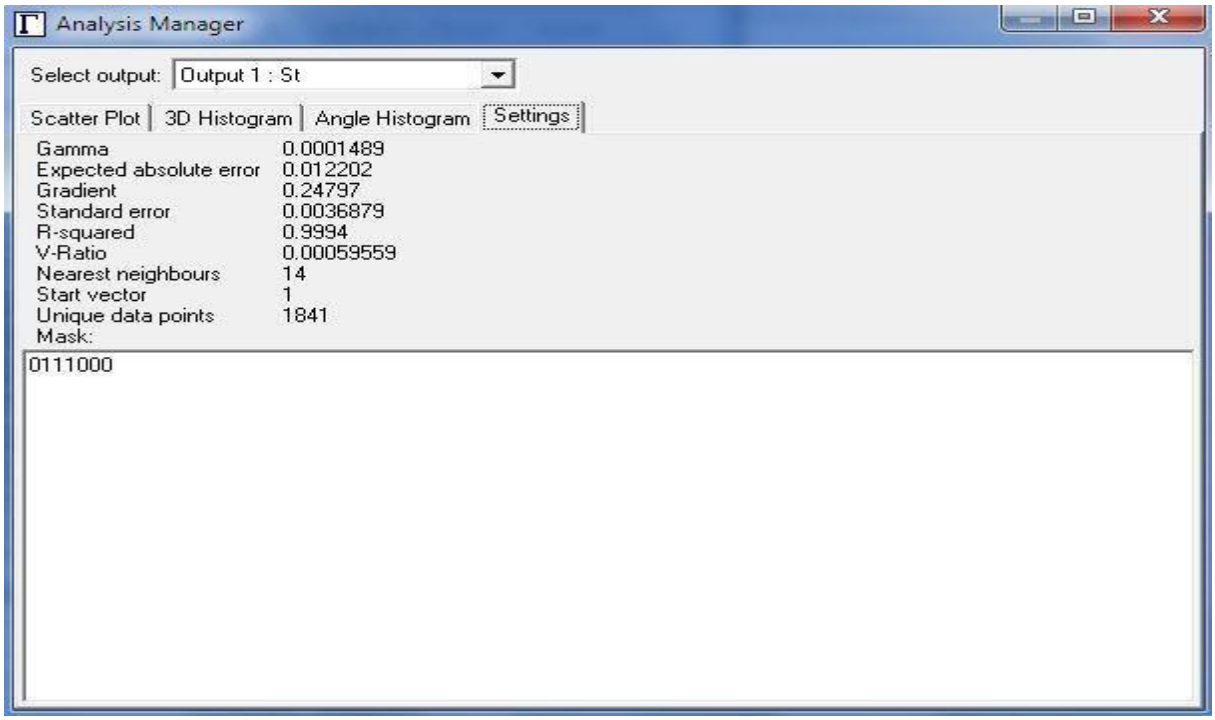


Figure 5: Mask Result of Input variables from Genetic Algorithm

The binary numbers of 0 and 1 are assigned to each input variable as from the imported dataset from excel which is shown in the table below:

Table 1: Interpreting the Mask result of the Genetic Algorithm from WinGamma

Q_t (Discharge)	Q_{t-1} (Discharge)	Q_{t-2} (Discharge)	Q_s (Sediment Discharge)	Q_{s-1} (Sediment Discharge)	Q_{s-2} (Sediment Discharge)	S_{t-1} (Suspended Sediment load)
0	1	1	1	0	0	0

The input variables with 0 values show much noise in the data and so may not give the ANN model good results, on the other hand the variables with 1 value assigned shows good parameters for optimum performance of the ANN model. Thus, the ANN model will be developed for the 7 input variables and used to predict suspended sediment load and the result will be compared with the ANN model output of the GA selected 3 input variables and compared to see which gives a better result. Q_{t-1}

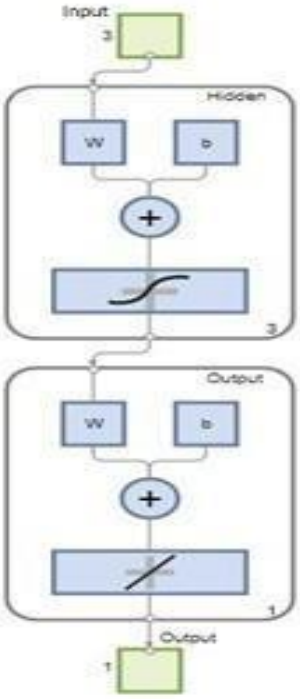
The normalized data set for the different data set was imported to MATLAB and used to build the ANN architecture. The data set was split in 3 parts, 70 per cent of the data was used for training, 15 per cent was used for testing while the remaining 15 per cent was used for validation

More so, the pruning of the ANN models was done by rule of thumb which is illustrated by adopting 7 nodes in the hidden layer for the 7 input variables while for the model with 3 input variables the nodes in the hidden layer is 3; further analysis was done by also using the 7 hidden nodes for the model with 3 input variables while the model with 7 input variables utilized 3 hidden nodes to evaluate their performances under switched hidden nodes magnitude for the sake of uniformity across board and all the models were trained and used for testing as well as validation of the data set.

RESULTS

The outputs for the ANN model with 7 input variables as well as the GA-ANN model with 3 input variables with different nodes in the hidden layer are shown with performance evaluation of the four models done with coefficient of determination (R^2) and the Root Mean Square Error (RMSE) to ascertain which model has performed best based on their different techniques are detailed below:

The evaluation of the GA-ANN model with 3 Hidden Nodes is shown below



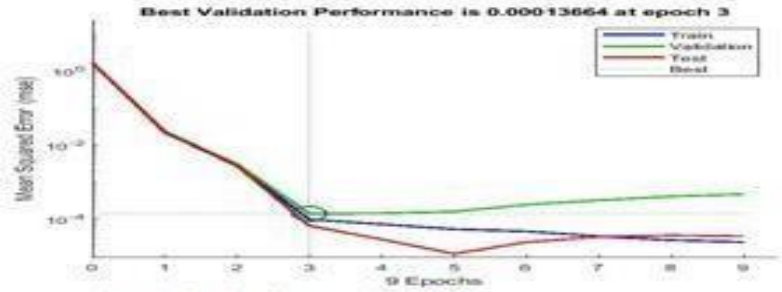


Figure 6b: the plot of MSE against Epoch

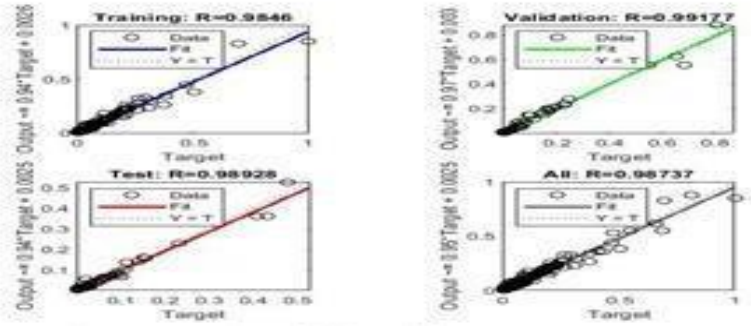


Figure 6c: Regression analysis the Model

From the figures above, the figure 6a shows the model architecture while figure 6b illustrates the at which iteration each part of the data set reach minimum error, the test data set reach its minimum error at the fifth iteration which is epoch 5 while validation and training each have their minimum errors at the third iteration which epoch 3; more so, from figure 6c the coefficient of correlation showed that the model has reliable performance with the model performing best at validation and a very good result in testing.

The analysis of the ANN model with 7 input variables are shown below

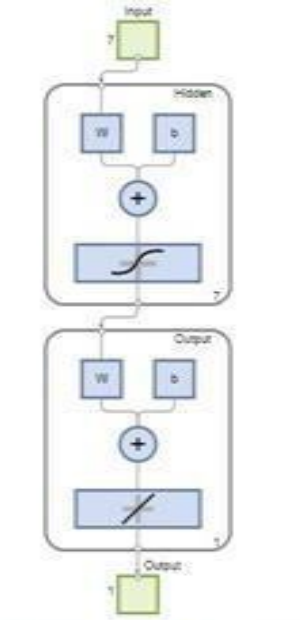


Figure 7a: ANN Architecture

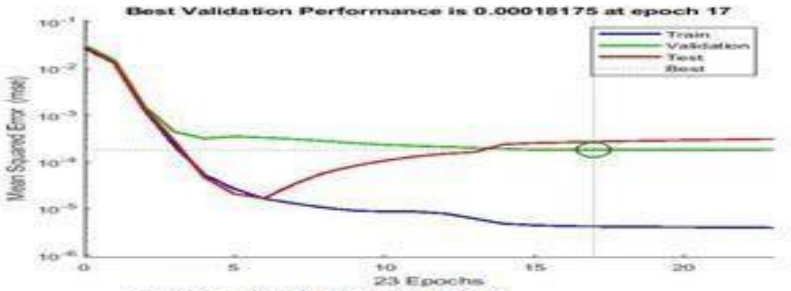


Figure 7b: the plot of MSE against Epoch

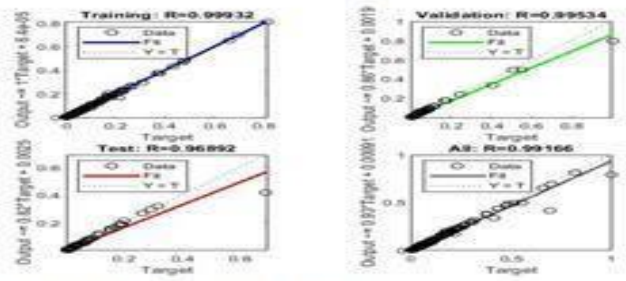


Figure 7c: Regression analysis the Model

As illustrated above, in figure 7b the validation data set has much error compared to the data set for training and testing. However, the coefficient of correlation and MSE is quite good and also has a reliable result as well with the training data set performing the best.

For the sake of uniformity across board, the GA-ANN model was also pruned with the magnitude of nodes in the hidden layer as determined for the ANN model of 7 input variables hence the GA-ANN model with 3 input variables was pruned with 7 nodes in the hidden layer as shown below with the results and analysis. Shown below is the GA-ANN model with 3 input variables and 7 hidden nodes and the details of the result.

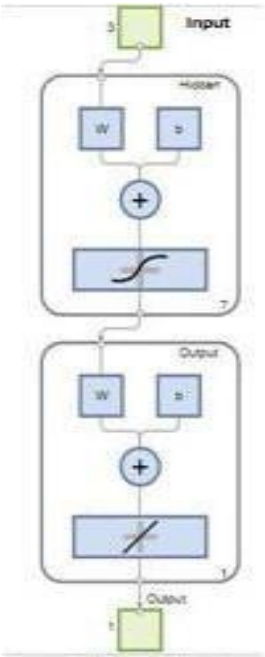


Figure 8a: ANN Architecture

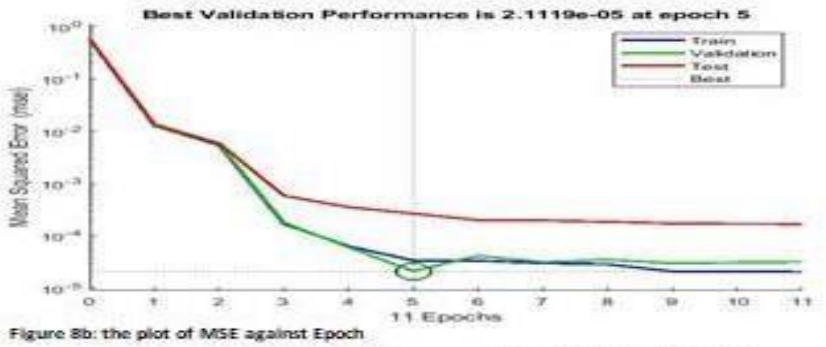


Figure 8b: the plot of MSE against Epoch

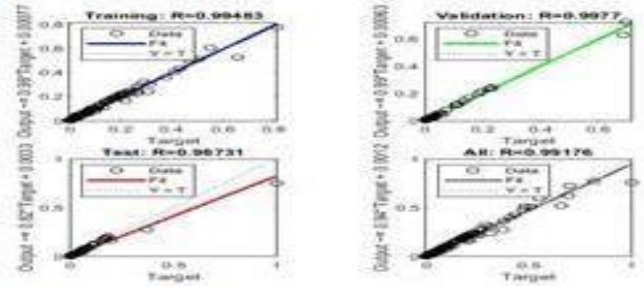


Figure 9c: Regression analysis the Model

As seen above, the result is very reliable as well and the ealiest iteration was achived at epoch 5 for Validation data setbut the test has the most errors after analysis; even with the most errors experieiced with the data set of the test data, the coefficient of determination and coefficient of correlation is produced a very reliable result.

The final comparison is done with ANN model of 7 input variables and 3 nodes in the hidden layer to further check if the result is relaible and juxtaposed with the previous results for viability and to scrutinize which model perfroms best. Hence, the figures below show the analysis done

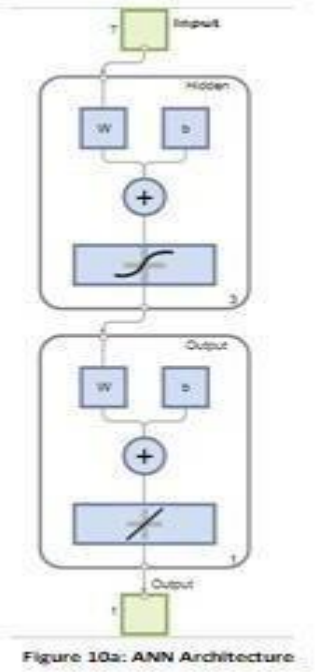


Figure 10a: ANN Architecture

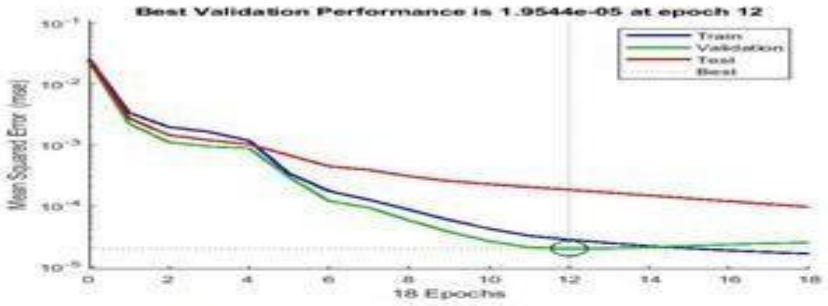


Figure 10b: the plot of MSE against Epoch

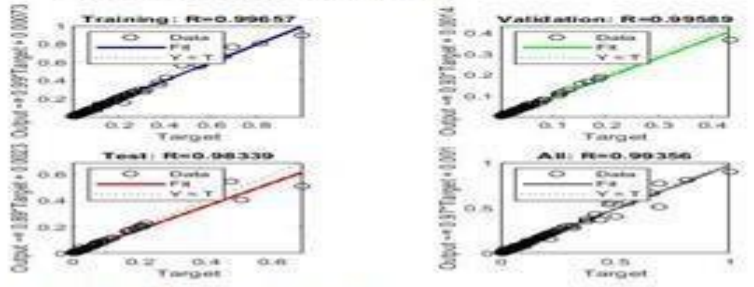


Figure 10c: Regression analysis the Model

The analysis show a very reliable result as well from figure 10c with good coefficient of correlation with a very low error during training data set and validation data set.

Table 2: Results of ANN models and GA-ANN models with different hidden nodes

Model	**Input Variables**	**Number of nodes in hidden layer**	**Output Variable**	**R**	**R^2**	**MSE**	**RMSE**
ANN	Q_t, Q_{t-1}, Q_{t-2}, Q_s, Q_{s-1},Q_{s-2}, S_{t-1}	7	S_t	0.9689	0.9388	0.0002694	0.00000007258
ANN	Q_t, Q_{t-1}, Q_{t-2}, Q_s, Q_{s-1},Q_{s-2}, S_{t-1}	3	S_t	0.9834	0.9671	0.0001813	0.00000003287
GA - ANN	Q_{t-1}, Q_{t-2}, Q_s	3	S_t	0.9893	0.9787	0.0000653	0.00000000426
GA - ANN	Q_{t-1}, Q_{t-2}, Q_s	7	S_t	0.9873	0.9748	0.0002655	0.00000007049

From the table above, it shows the ANN model with 7 input variables and 3 hidden nodes had a satisfactory result with $R^2 = 0.9671$ however the ANN model with 7 input variables and 7 hidden nodes as stated by the rule of thumb performed better with $R^2 = 0.9388$ but the GA-ANN model with 3 input variables with 7 hidden nodes performed better with $R^2 = 0.9748$ while the GA-ANN model with 3 input variables and 5 hidden nodes as computed by

the rule of thumb performed best with a $R^2 = 0.9787$ and has the lowest RMSE magnitude of 0.00000000426 as detailed above with the corresponding coefficient of correlation as well as Mean Square Error results.

CONCLUSION

The results illustrated for the ANN model and the GA-ANN model showed that the GA-ANN model outperformed the ANN model irrespective of the number of hidden nodes with less effort as against the common method of using GA to train the ANN model which requires rigorous iteration that are effort and time consuming in order to attain a desirable output. This method is less time consuming and the GA technique in WinGamma does all the iteration and computations in several minutes and arrives at a conclusion on which input variables will be desirable for the ANN model in order to attain an optimized result. Hence, the GA technique in WinGamma can be used reliably to select relevant input variables from the dataset for the ANN model to produce a better output for the ANN model

References

Abrahart, R. J., Anctil, F., Coulibaly, P., Dawson, C. W., Mount, N. J., See, L. M., Shamseldin, A. Y., Solomatine, D. P., Toth, E., & Wilby, R. L. (2012). Two decades of anarchy? emerging themes and outstanding challenges for neural network river forecasting. *Progress in Physical Geography, 36*(4), 480-513.

Adegun, O., Ajayi, O., Badru, G., & Odunuga, S. (2018). Water, energy and agricultural landuse trends at shiroro hydropower station and environs. *Proceedings of the International Association of Hydrological Sciences, 376*, 35.

Adie, D. B., Ismail, A., Muhammad, M. M., & Aliyu, U. B. (2012). Analysis of the water resources potential and useful life of the shiroro dam, nigeria. *Nigerian Journal of Basic and Applied Sciences, 20*(4), 341-348.

Besaw, L. E., & Rizzo, D. M. (2007). Stochastic simulation and spatial estimation with multiple data types using artificial neural networks. *Water Resources Research, 43*(11)

Bowden, G. J., Dandy, G. C., & Maier, H. R. (2005). Input determination for neural network models in water resources applications. part 1—background and methodology. *Journal of Hydrology, 301*(1-4), 75-92.

Chen, S. H., Jakeman, A. J., & Norton, J. P. (2008). Artificial intelligence techniques: An introduction to their use for modelling environmental systems. *Mathematics and Computers in Simulation, 78*(2-3), 379-400.

Chen, Y., & Chang, F. (2009). Evolutionary artificial neural networks for hydrological systems forecasting. *Journal of Hydrology, 367*(1-2), 125-137.

Dibie, C. D. (2019). A critical review on artificial intelligence models in hydrological forecasting how reliable are artificial intelligence models.

Darwin, C. (1859). The origin of species by means of natural selection.

Dawson, C. W., & Wilby, R. L. (2001). Hydrological modelling using artificial neural networks. *Progress in Physical Geography, 25*(1), 80-108.

Eze, J. N. (2005). No title. *Vulnerability and Adaptation to Climate Variability and Extremes: A Case Study of Flooding in Niger State, Nigeria,*

Ferentinou, M., & Fakir, M. (2017). An ANN approach for the prediction of uniaxial compressive strength, of some sedimentary and igneous rocks in eastern KwaZulu- natal. *Procedia Engineering, 191*, 1117-1125.

Goldberg, D. E. (1989). Genetic algorithms in search. *Optimization, and MachineLearning,*

Goren, H. G., Tunali, S., & Jans, R. (2010). A review of applications of genetic algorithms in lot sizing. *Journal of Intelligent Manufacturing, 21*(4), 575-590.

Holland, J. H. (1992). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence* MIT press.

Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. *Computer, 29*(3), 31-44.

Kolo, R. J. (1999). The assessment of physico-chemical parameters of shiroro lake and its major tributaries.

Maier, H. R., & Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. *Environmental Modelling & Software, 15*(1), 101-124.

Minns, A. W., & Hall, M. J. (1996). Artificial neural networks as rainfall-runoff models. *Hydrological Sciences Journal, 41*(3), 399-417.

Palani, S., Liong, S., & Tkalich, P. (2008). An ANN application for water quality forecasting. *Marine Pollution Bulletin, 56*(9), 1586-1597.

Prasad, R., Deo, R. C., Li, Y., & Maraseni, T. (2017). Input selection and performance optimization of ANN-based streamflow forecasts in the drought-prone murray darling basin region using IIS and MODWT algorithm. *Atmospheric Research, 197*, 42-63.

Saini, N. (2017). Review of selection methods in genetic algorithms. *International Journal of Engineering and Computer Science, 6*(12), 22261-22263.

Singh, R. M., & Datta, B. (2007). Artificial neural network modeling for identification of unknown pollution sources in groundwater with partially missing concentration observation data. *Water Resources Management, 21*(3), 557-572.

Starrett, S. K., Starrett, S. K., Najjar, Y., Adams, G., & Hill, J. (1998). Modeling pesticide leaching from golf courses using artificial neural networks. *Communications in Soil Science and Plant Analysis, 29*(19-20), 3093-3106.

Suleiman, Y. M., & Ifabiyi, L. P. (2015). The role of rainfall variability in reservoir storage management at shiroro hydropower dam, nigeria. *Momona Ethiopian Journal of Science, 7*(1), 55-63.

Yesilnacar, M. I., Sahinkaya, E., Naz, M., & Ozkaya, B. (2008). Neural network prediction of nitrate in groundwater of harran plain, turkey. *Environmental Geology, 56*(1), 19-25.