Chuka University
info@chuka.ac.ke | 020 231 0512 / 020 231 0518

APPLICATION OF ARTIFICIAL NEURAL NETWORK TO EVALUATE EXTEND OF NON-LINEARITY AMONG EXPLANATORY VARIABLES WITHIN AND BETWEEN GENOTYPES AND PHENOTYPES

Sumukwo Chesang, Thomas Kainga Muasya and Kiplangat Ngeno *Department of Animal Science, Chuka University, P. O. Box 109-60400, Chuka [2]Department of Animal Science, Egerton University, P. O. Box 536-20115, Egerton*
Email: chesangsumukwo@gmail.com, muasyakt@yahoo.com, aarapngeno@gmail.com; csumukwo@chuka.ac.ke

How to cite:
Chesang S, Muasya, T. K and Ngeno K. (2022). Application of artificial neural network to evaluate extend of non-linearity among explanatory variables within and between genotypes and phenotypes. In: Isutsa, D.K. (Ed) *Proceedings of the 8th International Research Conference held in Chuka University from 7th to 8th October2021, Chuka, Kenya, p.124-134***.**

ABSTRACT
Artificial neural networks (ANN) have been described as one of the models used for marker-based genomic predictions of complex traits in the field of animal breeding. It accommodates noisy, non-linearity in data set and makes decisions based on prior knowledge. This study evaluated the extent of non-linearity among explanatory variables within and between genotypes and phenotypes using ANN. A feedforward ANN was adopted with different number of neurons where Levenberg-Marquardt back-propagation algorithm was used to train the network. The construction and training of the network were done with matrix laboratory (MATLAB). Mean absolute error (MAE) and Pearson's correlation coefficients (R) were used to measure the ANN predictive performance as a measure of extent of non-linearity among explanatory variables within and between genotypes and phenotypes. Results showed that the ANN models differed in predictive performance depending on the number of neurons in the hidden layer, for instance the neural network with one hidden layer containing 10 neurons in the hidden layer yielded high R-value of 0.86 and MAE of 2.98E-3. When the network dimension was increased to 16 neurons the performance decreased to 0.67 for R and MAE increased to 7.73E-2. After a further increase of neurons to 32 the model yielded R value of 0.27 and MAE of 7.60E-2. The benchmark model for this study had an R of 0.77 and MAE of 5.72. Thus a model with 10 neurons is enough to handle non-linearity in this kind of data set thus chosen as the best non-linear model. This is because dimension reduction of neurons in the hidden layer led to higher, more accurate and more consistent predictions for growth rate. In comparison to linear model, the best non-linear model performed better though the more complex non-linear architectures with 16 and 32 neurons could not outperform the linear model. Thus, linear models can as well produce reliable results for making genomic predictions.
Keywords: Artificial neural network, Backpropagation, Mean absolute error

INTRODUCTION
Artificial neural networks (ANN) is a mathematical model which mimic the way the biological neural network of human brain works. It has the linear ability to learn from experience to improve its performance and to adapt themselves to changes in the environment (Njubi *et al*., 2010). The ANN architecture comprises of the input, hidden and output layers. These layers are connected by synapses denoted as weights. The weights are multiplied by input to give the desired output. ANNs have been successfully applied in fields of engineering, medical diagnosis, economic predictions and image recognition (Gorgulu, 2012). Consequently, it has proven to be a powerful modelling tool in comparison with other conventional models because it has an ability to predict outcome of non- linear and noisy data (Ehret *et al*., 2015).

Although ANN model has shown a lot of inspiring success in prediction of outcome, its application in animal breeding is still scarce (Gorgulu, 2012), considering huge dataset available to be analyzed in this field. The first application of ANNs in the field of animal breeding was by Macrossam *et al*. (1999), where they used ANN as a method for optimizing mating allocation to maximize production traits in Australian dairy industry. Since then the development and applications of ANNs in animal breeding are increasing steadily, owing to their flexibility in classification

recognition, prediction and forecasting, process control, optimization and decision support (Hanrahan, 2011). Artificial neural networks have been described as an additional model in performance prediction in the field of animal breeding (Gianola *et al*., 2011).

It has been pointed out to be a promising tool for molecular marker-based genomic predictions of complex traits in animals (Ehret, *et al*., 2015). Early prediction of performance is important in reducing generation intervals, making management decisions and increasing intensity of selection thus greater genetic progress (Gorgulu, 2012). Therefore, phenotypic prediction from genotypes plays an important role in identifying and selecting superior candidates to be used in a breeding program, this facilitates early utilization of best genotypes for economic gain. Genomic prediction is always applied early in a breeding program as a way to increase the overall selection pressurethus increasing the rate of genetic gain (Mcdowell, 2016). Selection and mating strategies are important breeding

improvement tools. This genetic improvement tools plays a major role in designing a breeding program and have been adopted successfully in developed countries unlike in developing countries where application of these practicesis still inadequate (Njubi *et al*., 2010).

Breeding programs in the tropics are based primarily on performance based on production traits, because they are readily available and often used to measure economic profitability of an enterprise (Njubi *et al*., 2010). Production traits such as milk production are affected by linear and non-linear interactions between environmental conditions, management, and seasons. The commonly used linear algebraic methods (Mrode, 2014) possess inherent restrictions which makes them not able to consider all these interactions. ANN have been reported to possess an ability to accommodate noisy data, linear and non-linear relationships between variables and interactions between explanatory variables and finally ambiguity of data from environmental influence (Ehret *et al*., 2015; Gonzalez-Camacho *et al*., 2012). They are also known for making accurate selection decisions based on prior knowledge of the outcome (Hosseinia *et al*., 2007). The focus of this study was, therefore, to evaluate the extent of non-linearity among explanatory variables within and between genotypes and phenotypes using ANN.

MATERIALS AND METHODS

Data Source

The phenotypic and genotypic data were obtained from a wide-genome study in four Chinese indigenous chicken breeds (Yuan *et al*., 2018) as described in Section 3.1 of this study. Live body weight (BW) was measured at hatch and every week until 12 weeks of age. For genotypic data, blood sample were collected from the sampled birds followed DNA extraction using the phenol-chloroform method. Illumina 60K Chicken SNP BeadChip described by Groenen *et al*. (2011). Quality control was conducted on all genotypes, after imposing the quality control checks, a total of 46211 SNPs was retained for analysis. A.mat function of the rrBLUP package installed in R software were used to impute missing genotypes (SNPs) where markers with 50% missing genotypes were not imputed. Genotypes were coded as {0 1 2} based on R code script by Eva KF Chan.

The data were split into training and test data set. Training data was splinted into 70:30, with 70% of the data being used for training, and 30% being used for validation. The sampling was random in order to avoid any selection bias in the dataset. Testing was done with the remaining samples that were not used during the training phase. Data preprocessing and analysis was performed using the R software. To investigate the performance of ANNs in evaluating the extent of non-linearity among explanatory variables within and between genotypes and phenotypes, the most frequently used feedforward ANN was adopted (Gianola *et al*., 2011). This study made a direct comparisonof the different neural network types using identical datasets to determine the most suitable architecture forpredicting a phenotypical trait based on genotypes.

This study applied two types of ANN, the single-layer feedforward network which are frequently used for regression problems and forecasting (Besic *et al*., 2017). They are the simplest form of a layered network consisting of an input layer of source nodes that project directly onto an output layer of neurons. The multilayer feedforward network which is distinguished by the presence of one or more hidden layers, whose computational nodes are correspondingly called hidden neurons were adopted. According to Badnjevic *et al*. (2015) linear feedforward neural network is often sufficient to properly perform classification tasks and is also applicable to regression tasks. The information generated was relayed in one direction without any loops or cycles between the input and output. Random weights were assigned to the neurons first. The linear combination (sum of the product) of the weights and inputs were calculated at each neuron.

The network was trained depending on the output obtained, when the obtained values were greater than a given threshold value, then the neuron “fires” assumes the activated value. In situations when the threshold is not reached, it assumes a deactivated value. The back-propagation algorithm was used to minimize the error term between the output of the neural network and the actual desired output value. The error term was calculated by comparing the net output to the desired output which was then feedback through the network, causing the synaptic weights to be configured in an effort to minimize error by activating training algorithm, which modifies the ANN parameters sequentially. This process was repeated until a sufficiently low level of error was reached, or until a predefined cutoff point was reached (Larose, 2014). A representation of a single and multilayer feedforward neural network adopted are shown in Figure 1 and 2, respectively

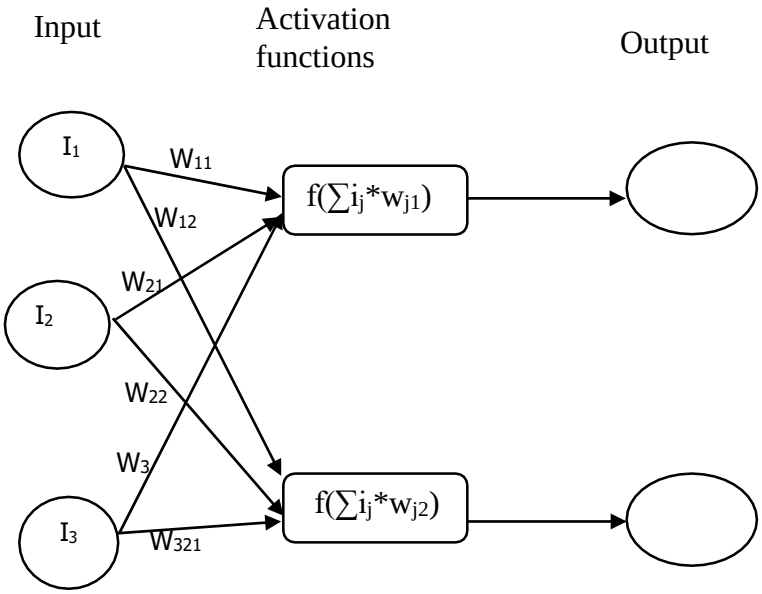

Figure 1: Single feed-forward network

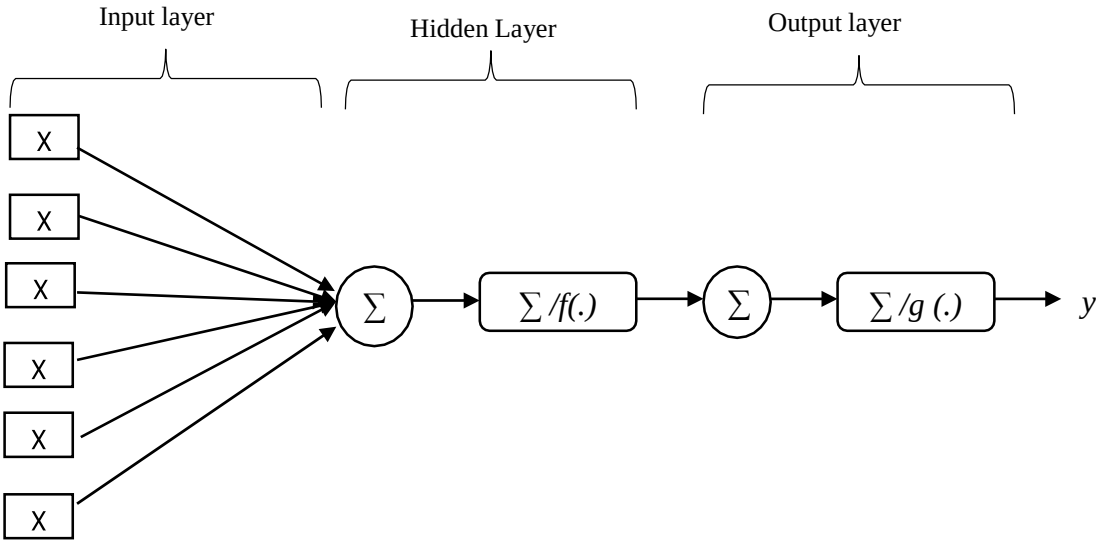

Fig. 2: Fully connected feed-forward network with one hidden layer

Network design and training

For this study different architecture of the feed-forward network was used, with each unit connected to all units in the next layer. The input to each hidden neuron is a linear combination of a vector of weights, input SNP variants and a "bias" weight for the feedforward networks. The input to each neuron was obtained as:

$$q_i^{[t]} = f_t\left(a_t + \sum_{i=1}^{m} w_{1j}^{[t]} X_{1j}\right)$$

where

$q_i^{[t]}$ = the hidden neuron
$w_{1j}^{[t]}$ = vector of weights
f_t =the sigmoid activation

x_{ij} = input value of the SNP variant

a_t = the bias weight
j= input SNP variant (range of 1 to m)
m = total number of input SNP variants

The sigmoid activation function was then used to transform the results obtained using the following equation to produce the hidden neurons output value:

$$f_t = \frac{1}{1+e^{-t}}$$

The inputs to neurons in the output layer is a linear combination of outputs of neurons in the hidden layer, weights of the output layer, and an output layer bias neuron. The value obtained is transformed by the linear transformation function $p_t(.)$ to generate the value of the predicted phenotypes (body weight) of an individual as:

$$y_i = p_t\left(b + \sum_{t=1}^{s} w_{2t} q_i^t\right)$$

where
y_i =predicted body weight
p_t= the linear transformation function
b = the bias neuron: t = the hidden neuron (range of 1 to s)
s = total number of hidden neurons
w_{2t} = vector of weights
q_i^t = hidden neuron

The training was done using the backpropagation procedure with training sets presented in a random order. The optimal weights were established using Levenberg-Marquardt back-propagation algorithm (*trainlm)* with a maximum number of iteration (epochs) equal to 2000, which is commonly used for training ANNs (Fojnica *et al.*, 2016), this minimizes the error between the predicted and the actual weight. The process of optimization was performed until an optimal mean error squared level is reached or stopping criteria was fulfilled. The sigmoid function was used as the activation function in the hidden layer, with the linear transformation function as shown in equation 3 being used as the activation function in the output layer.

Once the network was trained the testing dataset was finally presented to test the model, and the output for each subject was recorded, based on their marker genotype. The performance of neural networks built using feedforward architecture with the Levenberg-Marquardt training method was examined where the number of neurons in the hidden layer was repeatedly increased from one neuron, 10 neurons, 16 neurons and 32 neurons. Mean absolute error (MAE) and Pearson's correlation coefficients were used to measure the ANN predictive performance as a measure of extent of non-linearity among explanatory variables within and between genotypes and phenotypes. The Mean absolute error was calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | f_i - y_i | = \frac{1}{n}\sum_{i=1}^{n} | e_i |$$

where
f_i = the predicted value
y_i = the actual value
e_i = the error value
n = the number of sample
Pearson's correlation coefficient of the predicted and actual values *(r)* was calculated as a measure of linearity between input and output values as follows, where *n* is the total number of samples:

$$r = \frac{n\sum f_i y_i - \left(\sum f_i\right)\left(\sum y_i\right)}{\sqrt{[n\sum f_i^2 - \left(\sum f_i\right)^2][n\sum y_i^2 - \left(\sum y_i\right)^2]}}$$

RESULTS

The architectures adopted for this study was a single hidden layer with varying number of neurons i.e. from 1 to 32 neurons to examine the extent of non-linearity among explanatory variables within and between genotypes and phenotypes. Performance of different neural models was used to measure the extent of non-linearity. After subjecting the data sets to different neural network architecture, the MAE and R was used to measure the performance. The results for the training, validation and testing data sets are in Table 1.

As a benchmark, a linear model with one neuron in the hidden layer and linear activation functions in the hidden layer as well as in the output layer was adopted. The network is similar to genomic best linear unbiased prediction (GBLUP), in that network performs a multiple linear regression, in which weights of hidden layer can be interpretedas regression coefficients (Ehret *et al*., 2015).

Table 1: The Pearson's correlation coefficient (R) and mean absolute error (MAE) for training, validation and training data set for different neural network architecture

Architecture			R	MAE
Single hidden layer with 1 neuron	Linear	Train set	0.92	3.25E-3
		Validation set	0.62	7.75E-3
		Test set	0.77	5.72E-3
Single hidden layer with 10 neurons	Non-linear			
		Train set	0.97	8.31E-4
		Validation set	0.79	4.67E-3
		Test set	0.86	2.98E-3
Single hidden layer with 16 neurons	Non-linear			
		Train set	0.96	1.10E-3
		Validation set	0.70	8.07E-3
		Test set	0.68	7.73E-2
Single hidden layer with 32 neurons	Non-linear			
		Train set	0.97	9.31E-4
		Validation set	0.37	6.88E-2
		Test set	0.27	7.60E-2

The MAE is the average squared difference between the output and the target values, where, lower values are better and zero value means no error. The R-value measures the correlation between the output and the target/ actual values where R close to 1 means perfect relationship. The ANN models differed in predictive performance depending on the number of neurons in the hidden layer, for instance the neural network with one hidden layer containing 10 neurons in the hidden layer yield high correlation of 0.86 and performance (MAE) of 2.98E-3 as shown in Figure 3.

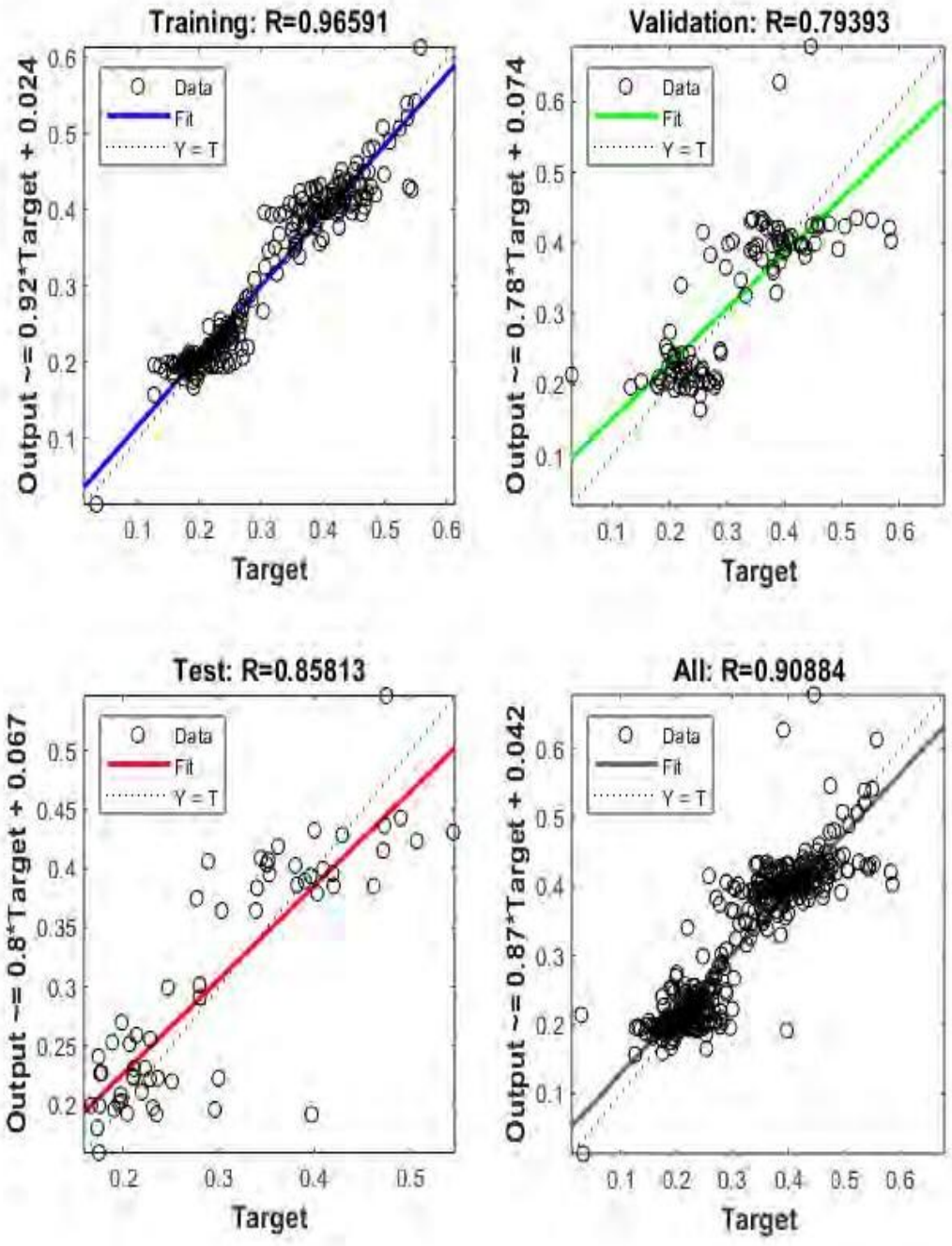

Figure 3: Pearson's correlation coefficients for the training, validation and testing data set related to 10 neuron architecture

When the network dimension was increased to 16 neurons, the performance of the neural network for test and target model decreased to 0.67 and 0.84 respectively for R and MAE increases to 7.73E-2 as shown in Figure 4.

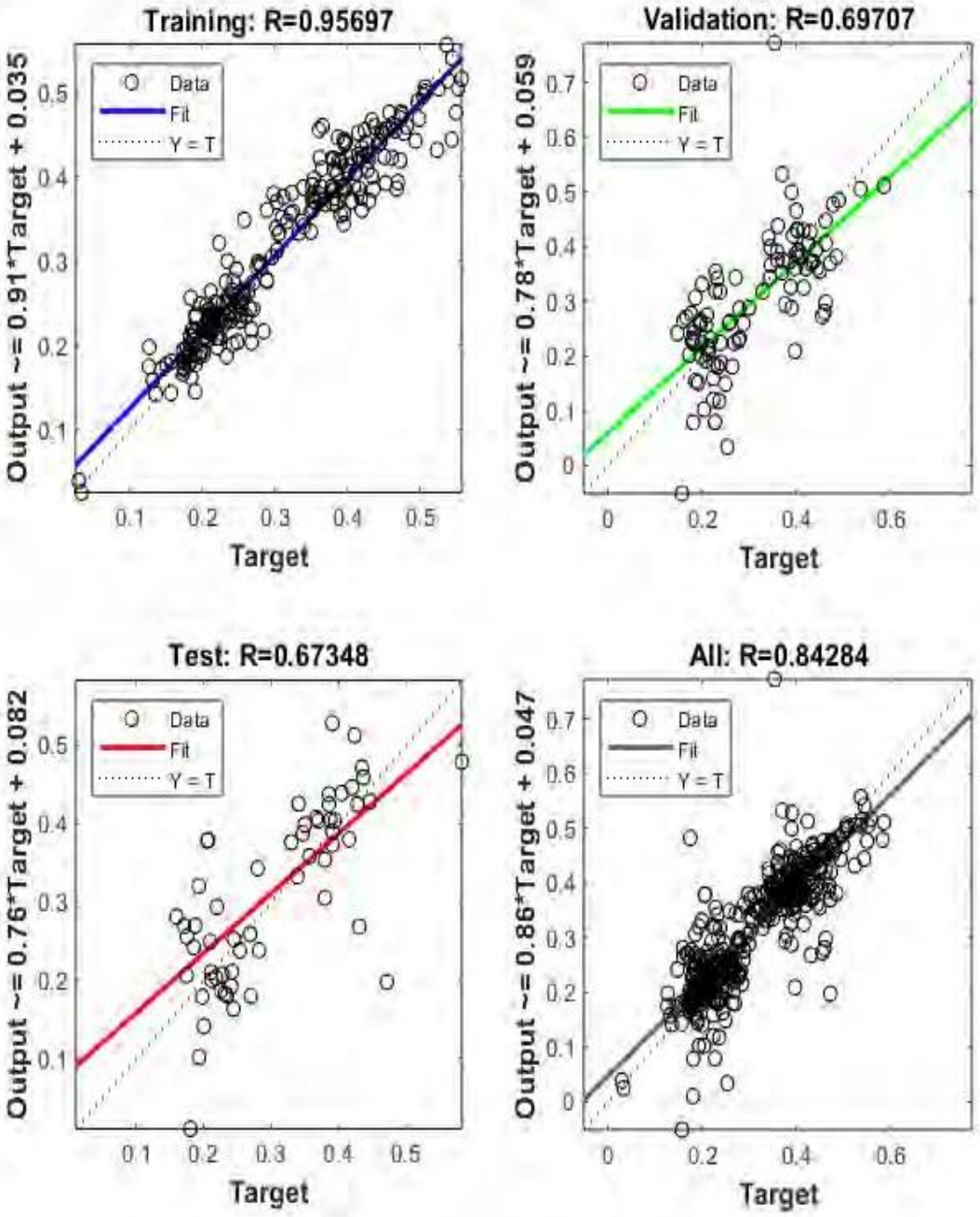

Figure 4: Correlation coefficients for the training, validation and testing data set related to 16 neuron architecture

Further increase of the network neurons to 32 the correlation coefficient for test and target decreases to 0.27 and 0.52 respectively for R and 7.60E-2 for MSE as shown in Figure 5.

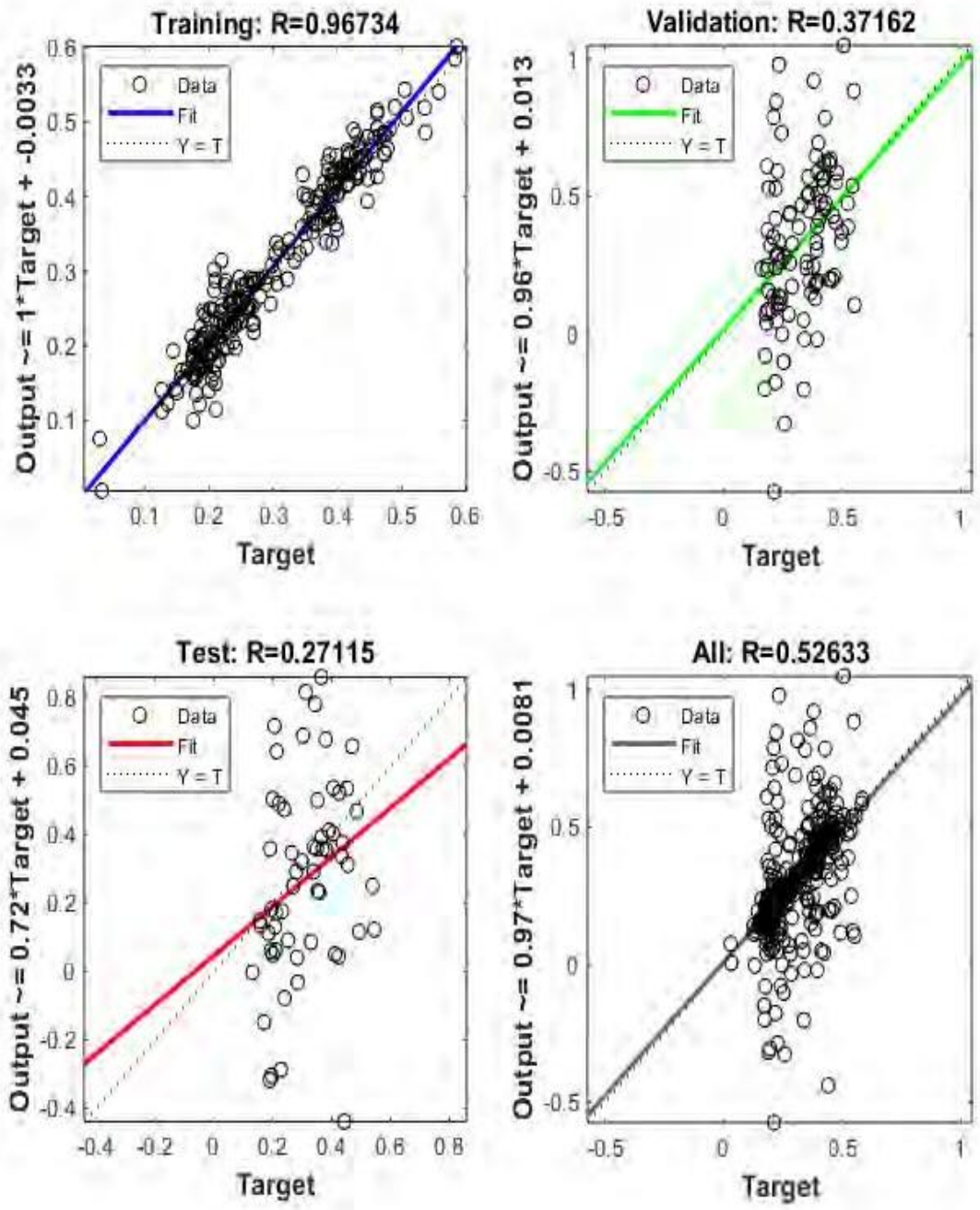

Figure 5: Correlation coefficients for the training, validation and testing data set related to 32 neuron architecture

As a benchmark for non-linear models, the linear model with one neuron in the hidden layer was adopted. The results from the testing data set of the model were 0.77 for R and 5.72E-3 for the MAE as shown in Figure 6. Computational time was long during the training process this is as a result of all input information passing through a single neuron.

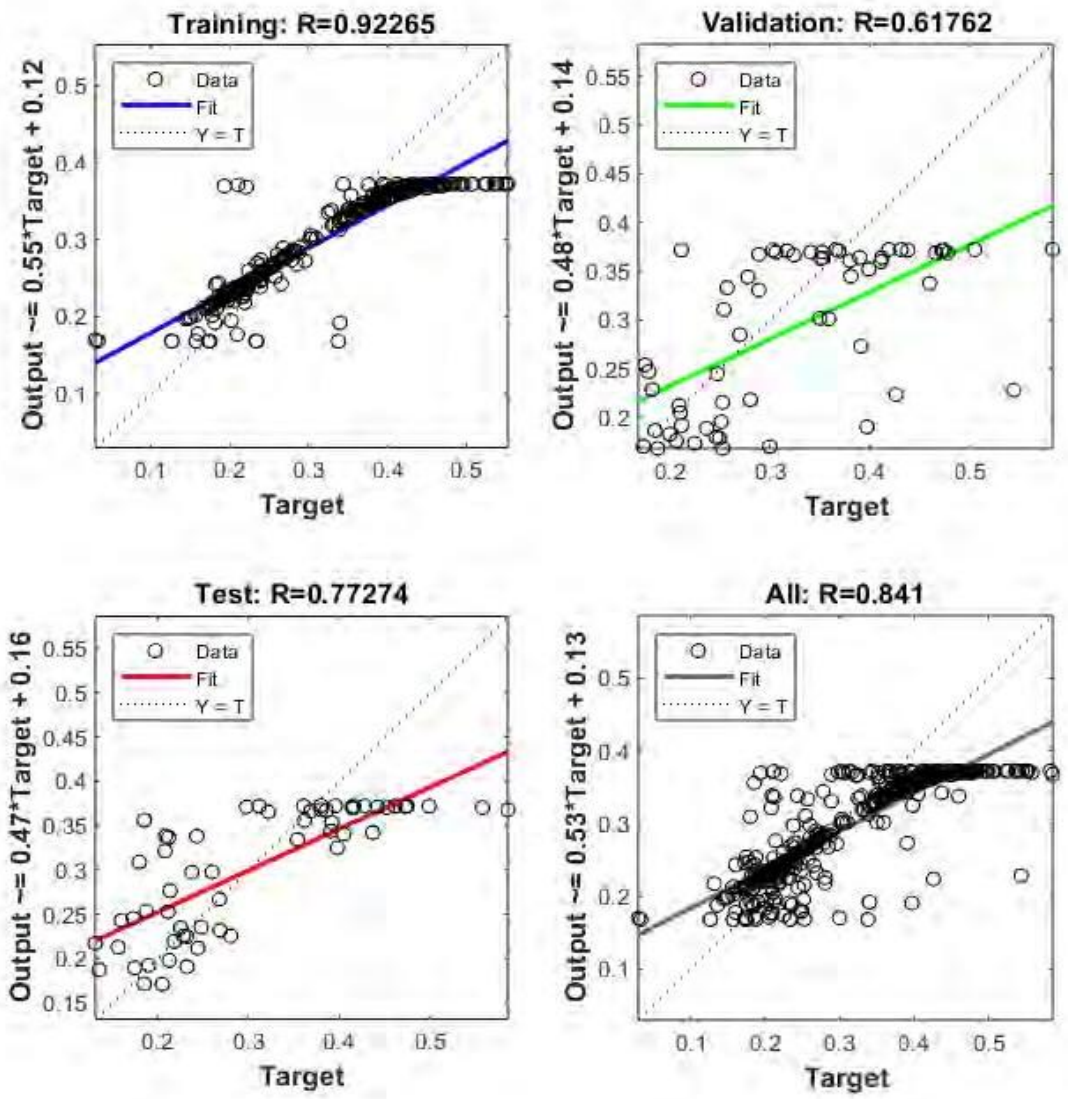

Figure 6: Representation of linear model correlation coefficients for the training, validation and testing data set related to one neuron architecture model

DISCUSSION

Artificial Neural Networks are machine learning models based on operating principles of the human brain. It has capability of modelling non-linear systems, with the information learned through experience and also they have ability to handle noisy data (Ehret *et al*., 2015). This model has been reported in literature to act as an alternative predictive model, because of their ability to act as universal approximators of complex functions and can capture non-linear relationships between predictors and responses (Gianola *et al*., 2011). The performance of ANN is determined by the network architecture i.e. the number of hidden layers and neurons, type of training, linear or non- linear transformation process and the nature of input data set. Therefore, for better performance of ANN, all these factors have to be put into consideration. For instance, the network dimensionality has immense effect through network overfitting or under fitting leading to reduction in performance predictions.

This study compared the extent of ANN model with different architecture in handling non-linearity of data for the prediction of a phenotypic trait based on SNP data. The effects of network architecture and number of neurons in

hidden layer on performance prediction were used as a measure of the best performing non-linear neural network model for the prediction of yet to be observed growth rate in indigenous chicken. The performance of different non-linear ANN models adopted for this study differed in predictions performance depending on the number of neurons in the hidden layer. The model with the highest correlation (R) and lower MAE is considered to be the best model, this is because MAE is the average squared difference between the output and the target, where the lower the values are better the model likewise R close to 1 means perfect relationship between output and target.

According to the results from test data set obtained from this study, the model with a single hidden layer and 10 neurons was considered to be the best non-linear model with R corresponding to 0.86 and lowest MAE of the value of 2.98E-3. Further increase in the number of neurons in the hidden layer to 16 and 32 the correlation for test data decreases to 0.68 and 0.27 respectively, subsequently the MAE was high corresponding to 7.73E-2 and 7.760E-2 respectively. These results, therefore, indicate that dimension reduction of neurons in the hidden layer resulted in higher, more accurate and more consistent predictions for growth rate. These results are consistent with those of Hamidi *et al*. (2017) and Ehret *et al*. (2015) who concluded that ANN is useful for functional traits with potential of mapping non-linear relationships between genotype and phenotype.

The more the number of neurons in the hidden layer, the worse the ANN model becomes. This is attributed to the fact that increasing the dimensionality of network architecture makes ANN learn irrelevant details of the data set andthe number of features increases making the model more complex. The more the number of features, the more the chances of a model being under or over-fitted thus making the model worse for prediction of future target values (Ehret *et al*., 2015). According to Mcdowell *et al*. (2015), network structure with many hidden layers is notoriously difficult to train, thus affecting the prediction performance. Therefore, increase in dimensions of the neural architecture through number of neurons in the hidden layer is only necessary in the case of noisy datasets, where it served to handle the noisy data (Besic *et al*., 2017). As a recommendation, for better predictions of phenotypes it's important to consider the nature of neural network, the choice of input values and its distribution in the input data set. From this study, therefore, a single hidden layer with 10 neurons has an ability to account for non-linearity in thedata set thus resulting in better performance predictions.

The current study also investigated the performance of the linear model as a benchmark in prediction of growth rate in comparison with the non-linear models. The results for the test data showed that the linear model had acorrelation corresponding to 0.77 and MAE of 5.72 E-3 as compared to those of non-linear models shown in Table 1. The model chosen as the best non-linear model performed better than the linear model even though the more complex non-linear architectures with 16 and 32 neurons could not outperform the linear ANN. These results were inconsistent with those of Ehret *et al*. (2015) which reported that the linear ANN and the best non-linear model performance were almost similar. This indicates that linear models can as well produce reliable results for making genomic predictions. Non-linear ANNs for predicting performance can more accurately evaluate effects of each genotype without much interference of environmental effect, therefore selection of the superior progenies in a breeding program will be more efficient using the superior genotypic value, and not the superior phenotypic value (Peixoto *et al*., 2015). The ANN has been reported to greatly approximate unknown relationships, and work much better in the absence of noise in the dataset (Besic *et al*., 2017), thus the best ANNs have ability to handle non- linearity in dataset.

CONCLUSION

The results of this study showed the extent of non-linearity between and within genotypes and phenotypes as determined by the ANN architecture. Thus, for efficient consideration of non-linearity in data set, this study recommends adoption of a single hidden layer with 10 neurons because dimension reduction leads to higher, more accurate and more consistent predictions performance for growth rate. This study reveals the fact that increase of dimensionality of network architecture leads to model under fitting or overfitting as proven by the results obtained from the complex non-linear models (with 16 and 32 neurons)

REFERENCES

Badnjevic A, Cifrek M, Koruga D, Osmankovic D. Neuro-fuzzy classification of asthma and chronic obstructive pulmonary disease. *BMC medical informatics and decision making*. 2015; 15(3): S1.

Besic L, Muhovic I, Asic A, Catic A, Gurbeta L, Badnjevic A. Application of neural networks to the prediction of a phenotypic trait of pacific lampreys based on single nucleotide polymorphism (SNP) genetic markers, 2017.

Ehret A, Hochstuhl D, Gianola D, Thaller G. Application of neural networks with back-propagation to genome- enabled prediction of complex traits in Holstein-Friesian and German Fleckvieh cattle. *Genetics, Selection, Evolution***.** 2015.

Fojnica A, Osmanović A, & Badnjević A. Dynamical model of tuberculosis-multiple strain prediction based on artificial neural network. In 2016 *5th Mediterranean Conference on Embedded Computing (MECO) 2016*; 290-293. IEEE.

Gianola D, Okut H, Weigel KA, Rosa GJ. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC genetics*. 2011; 12(1):87.

Gonzalez-Camacho JM, de Los Campos G, Perez P, Gianola D, Cairns JE, Mahuku G, et al. Genome-enabled prediction of genetic values using radial basis function neural networks. *Theoretical and Applied Genetics*. 2012; 125(4):759-771.

Gorgulu O. Prediction of 305-day milk yield in Brown Swiss cattle using artificial neural networks. *South African Journal of Animal Science*, 2012, 42(3).

Groenen MA, Megens HJ, Zare Y, Warren WC, Hillier LW, Crooijmans RP, et al. The development and characterization of a 60K SNP chip for chicken. *BMC Genomics*. 2011; 12(1):274-283.

Hanrahan G. Artificial Neural Networks in Biological and Environmental Analysis. CRC Press, 2011.

Hosseinia P, Edrisi M, Edriss MA, and Nilforooshan MA. Prediction of second parity milk yield and fat percentage of dairy cows based on first parity information using neural network system. *Journal of Applied Science.* 2007; 7:3274-3279.

Larose DT, Larose CD. 2014. *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.

Macrossan P, D Hand, J Kok, M Berthold, H Abbass, K Mengersen, et al. Bayesian neural network learning for prediction in the Australian dairy industry. *Advances in Intelligent Data Analysis*. Springer Berlin Heidelberg. 1999; 1642:395-406

McDowell RM. *Genomic selection with deep neural networks*. MSc Thesis, Iowa State University, Ames, Iowa, 2016.

Njubi DM, Wakhungu JW, Badamana MS. Use of test-day records to predict first lactation 305-day milk yield usingartificial neural network in Kenyan Holstein–Friesian dairy cows. Trop. *Animal Health Production*. 2010; 42:639-644.

Peixoto LA, Bhering LL, Cruz CD. Artificial neural networks reveal efficiency in genetic value prediction. *Genetic Molecular Resource*. 2015; 14(2):6796-807.

Pour Hamidi S, Mohammadabadi MR, Asadi Foozi M, Nezamabadi-pour H. Prediction of breeding values for the milk production trait in Iranian Holstein cows applying artificial neural networks. *Journal of Livestock Science and Technologies*. 2017; 5(2):53-61.

Team RC. R: A language and environment for statistical computing, 2017.

Yuan Y, Peng D, Gu X, Gong Y, Sheng Z, Hu X. Polygenic Basis and Variable Genetic Architectures Contribute to the Complex Nature of Body Weight—A Genome-Wide Study in Four Chinese Indigenous ChickenBreeds. *Frontiers in Genetics*. 2018; 9:229.