PREDICTION OF INDIGENOUS CHICKEN GENOMIC BREEDING VALUES AND PERFORMANCE USING MACHINE LEARNING

EGERTON UNIV ITY LIBRARY
17 JAN 2021
P.O. BOX 536, EGERTON
ACQUISITION SECTION

CHESANG SUMUKWO

EULIB
072988

A Thesis Submitted to Graduate School in Partial Fulfillment of the Requirements for the Master of Science Degree in Animal Breeding and Genomics of Egerton University

EGERTON UNIVERSITY LIBRARY

EGERTON UNIVERSITY LIBRARY
30 MAR 2021
P.O. BOX 536, EGERTON
AFRICANA SECTION

EGERTON UNIVERSITY

JULY 2020

2021/113516

DECLARATION AND RECOMMENDATION

Declaration

This thesis is my own original work and has not been presented for an award for any degree in this University or any other University

Signature: .....

Date 13/7/2020

Chesang Sumukwo

KM111/12137/17

Recommendation

This thesis has been presented with our approval as the official university supervisors.

Signature: ..... Date: 13/07/2020

.....

Dr. Thomas, K. Muasya, (*Dr. rer. agr.*)

Department of Animal Sciences

Egerton University

Signature: ..... Date: 13/7/2020

.....

Dr. Kiplangat Ngeno, PhD

Department of Animal Sciences

Egerton University

COPYRIGHT

©2020, Chesang Sumukwo

All rights reserved. No part or whole of this thesis may be reproduced or transmitted in any form or means such as electronic, mechanical including photocopying, recording or any information storage and retrieval system without prior permission of the author or Egerton University on behalf of the author.

DEDICATION

This work is dedicated to my mother and more so my guardian Wesley Komen for his outstanding support throughout my academic journey.

ACKNOWLEDGEMENTS

I am very thankful to the Almighty God for taking me through this Thesis development and throughout my MSc program. My gratitude goes to Egerton University's Graduate School for admission and registration, Department of Animal Sciences for giving me support and an opportunity to pursue postgraduate studies. I am grateful to my supervisors Dr. T. K. Muasya and Dr. K. Ngeno for moral support. Thanks so much for your outstanding technical and professional guidance since we started this work. I thank you so much for accepting me to be your student and nurturing me throughout this academic journey. Am happy to report to you that I have learned a lot from you and I can now be an independent thinker and approach issues as a scientific professional.

I am further thankful to the center of excellence in agriculture and agribusiness management (CESAAM) for giving me a scholarship opportunity to pursue my MSc. Program at Egerton University. I really want to appreciate you for funding this research work and your unmatched support throughout my studies as your student. Also, I want to appreciate Yang Yuan for his generosity of making data available to be used in this study and Dr. Eva KF Chan for provisions of scripts used for coding the data used for this study.

I am also grateful to my fellow postgraduate students for their moral support and encouragement throughout this period. I want to acknowledge my family more so my mother whom she believed that I will make it, and constantly prayed and supported me both financially and morally. To my guardian I am really humbled for your accepting me as your son and your unwavering support and encouragement you gave me since I was in primary school until now, may God bless you abundantly.

ABSTRACT

Genomic selection (GS) is a new breeding strategy that is rapidly becoming the method of choice of selection. It predicts the phenotypes of quantitative traits based on genome-wide markers of genotypes using conventional predictive models. However, these conventional predictive models are faced with a statistical challenge related to high dimensionality of marker data and typically make strong assumptions related to linear regression analysis. They also do not capture the complex, non-linear relationships within genotypes and between genotypes and phenotypes. Machine learning has been reported to be a promising tool, which accounts for these shortcomings. This study aimed at determining the adoption of machine learning in genomic prediction and performance estimation of indigenous chicken (IC). To compare predictive ability of genomic models, DeepGS, ridge-regression best linear unbiased prediction (RR-BLUP), Ensemble and artificial neural network (ANN) models were adopted in predicting body weight (BW) of IC based on genome-wide markers. The Pearson correlation coefficient (PCC) results for the four models were 0.891, 0.889, 0.892 and 0.812 respectively. All models were not significant difference ($p>0.05$) from each other, therefore, all the four models can be used in complement. For performance prediction, mean normalized discounted cumulative gain value (MNDCGV) were adopted. The MNDCGV results showed that the accuracy of genomic estimated breeding values (GEBVs) estimated using DeepGS RR-BLUP and Ensemble were 0.75~0.78, 0.66~0.76 and 0.76~0.79 respectively. Thus, the Ensemble and DeepGS model outperformed the RR-BLUP model by a significant margin ($P<0.05$), therefore they can be used as a supplement to RR-BLUP. To evaluate the extent of non-linearity among explanatory variables within genotypes and between genotypes and phenotypes, a multilayer perceptron ANN was adopted. Mean absolute error (MAE) and PCC were used to measure the ANN performance. The results showed that the neural network with one hidden layer containing 10 neurons yielded high PCC value of 0.86 and MSE of 2.98E-3. Further increase of network dimension to 16 and 32 neurons the performance decreased to 0.67 and 0.27 for PCC and MAE increased to 7.73E-2 and 7.60E-2 respectively. Thus, model with 10 neurons is enough to handle non-linearity of data set thus chosen as the best non-linear model. This study concludes that DeepGS, Ensemble, RR-BLUP and ANN models can be used interchangeably in making phenotypic predictions, also the Ensemble and DeepGS model are appropriate in phenotypic ranking of individuals.

TABLE OF CONTENTS

DECLARATION AND RECOMMENDATION ii
COPYRIGHT iii
DEDICATION iv
ACKNOWLEDGEMENTS v
ABSTRACT vi
TABLE OF CONTENTS vii
LIST OF TABLES xi
LIST OF FIGURES xi
LIST OF ABBREVIATIONS AND ACRONYMS xii
CHAPTER ONE 1
INTRODUCTION 1
1.1 Background information 1
1.2 Statement of the problem 2
1.3 Objectives of the study 2
1.3.1 Broad objectives 2
1.3.2 Specific Objectives 2
1.4 Research Questions 3
1.5 Justification 3
1.6 Limitation of the study 3
CHAPTER TWO 4
LITERATURE REVIEW 4
2.1 Indigenous chicken (IC) production in Kenya 4
2.2 Production performance of indigenous chicken 4
2.3 Performance improvement strategies 4
2.3.1 General management 5
2.4 Genetic improvement 6
2.4.1 Crossbreeding 6
2.4.2 Breed substitution 7
2.4.3 Within breed selection 7
2.5 Breeding programs 7
2.6 Genomic selection 8
2.7 Prediction accuracy of genomic estimated breeding values 9

2.8 Strategies to improve the accuracy of genomic estimated breeding values 10
2.9 The transition from traditional breeding schemes to genomic selection era 10
2.10 Conventional genomic selection models, ability and limitations 10
2.11 Genomic selection: from linear model to complex models 11
2.12 Machine learning in genomic enabled prediction 12
2.12.1 Artificial Neural Networks 12
2.12.2 Deep learning 13
CHAPTER THREE 14
PREDICTION OF INDIGENOUS CHICKEN PHENOTYPES FROM GENOTYPES: COMPARISON BETWEEN MACHINE LEARNING AND CONVENTIONAL LINEAR MODELS 14
Abstract 14
3.1 Introduction 15
3.2 Materials and methods 16
3.2.1 Data source 16
3.2.2 Phenotyping 16
3.2.3 Sample collection, DNA extraction and genotyping 16
3.2.4 Genotype preprocessing and coding 16
3.2.5 Data analysis 17
3.2.6 Statistical analysis 19
3.3 Results 19
3.4 Discussion 21
3.5 Conclusion 22
CHAPTER FOUR 24
RANKING INDIVIDUALS BASED ON PREDICTED PERFORMANCE USING MEAN NORMALIZED DISCOUNTED CUMULATIVE GAIN VALUE 24
Abstract 24
4.1 Introduction 24
4.2 Materials and methods 26
4.2.1 Phenotypes and genotypes source 26
4.2.2 Missing markers and coding 26
4.2.3 Data analysis 26
4.2.4 Statistical analysis 27

4.3 Results 28
4.4 Discussion 30
4.5 Conclusion 31
CHAPTER FIVE 33
APPLICATION OF ARTIFICIAL NEURAL NETWORK (ANN) TO EVALUATE EXTEND OF NON-LINEARITY AMONG EXPLANATORY VARIABLES WITHIN AND BETWEEN GENOTYPES AND PHENOTYPES 33
Abstract 33
5.1 Introduction 34
5.2 Materials and methods 35
5.2.1 Data source 35
5.2.2 Imputation of missing genotypes and coding 35
5.2.3 Statistical analysis 35
5.2.4 Feedforward artificial neural network 36
5.2.5 Network design and training 37
5.3 Results 39
5.4 Discussion 44
5.5 Conclusion 46
CHAPTER SIX 48
GENERAL DISCUSSION, COCLUSIONS AND RECOMMENDATIONS 48
6.1 Aim of the study 48
6.2 Study Methodology 49
6.3 Comparison between machine learning and conventional linear models 49
6.4 Evaluation of prediction performance of genomic models 50
6.5 Extend of non-linearity among explanatory variables within and between genotypes and phenotypes 51
6.6 Conclusions 52
6.7 Recommendations 52
REFERENCES 53

APPENDICES **60**

Appendix I. Data preparation and parameters setting syntaxes for DeepGS 60

Appendix II. SAS output for objective one 64

Appendix III. Publications abstract 69

Appendix IV. Research permit 72

LIST OF TABLES

Table 1: mean separation using the least significant difference (LSD) of the GS models 20
Table 2: PCC between observed and predicted BW based on DeepGS and RRBLUP model 21
Table 3: Mean separation using least significant difference (LSD) for GS models 29
Table 4: Prediction performance based on MNDCGV for Ensemble, DeepGS and RR-BLUP with top-ranked alpha increasing from 10% to 70% 29
Table 5: The Pearson's correlation coefficient (R) and mean absolute error (MAE) for training, validation and training data set for different neural network architecture 40

LIST OF FIGURES

Figure 1: Pearson's correlation coefficient (PCC) analysis results between observed and predicted phenotypic values for the four GS models. The correlation results from different GS models are shown in different colours. 19

Figure 2: Box plot of the absolute errors between the observed and predicted phenotypic values. 20

Figure 3: Mean normalized discounted cumulative gain values (MNDCGV) curves for DeepGS, RR-BLUP, and Ensemble model with top-ranked alpha increasing from 10% to 70% 28

Figure 4: The absolute increase in MNDCGV of DeepGS and the ensemble GS models over RR-BLUP evaluated using fivefold cross-validation with five replicates 30

Figure 5: Single feed-forward network 37

Figure 6: Fully connected feed-forward network with one hidden layer 37

Figure 7: Pearson's correlation coefficients for the training, validation and testing data set related to 10 neuron architecture 41

Figure 8: Correlation coefficients for the training, validation and testing data set related to 16 neuron architecture 42

Figure 9: Correlation coefficients for the training, validation and testing data set related to 32 neuron architecture 43

Figure 10: Representation of linear model correlation coefficients for the training, validation and testing data set related to one neuron architecture model 44

LIST OF ABBREVIATIONS AND ACRONYMS

ADG	Average daily growth
AI	Artificial intelligence
ANN	Artificial neural network
BLUP	Best Linear Unbiased Prediction
BW	Body weight
CPU	Central-processing-unit
CV	Cross Validation
DCNN	Deep convolutional neural network
DL	Deep learning
EBV	Estimated breeding values
GEBV	Genomic estimated breeding values
GS	Genomic Selection
GWS	Genomic wide selection
IC	Indigenous chicken
LSD	Least significant difference
MAE	Mean absolute error
MAF	Minor Allele Frequency
MAS	Marker assisted selection
MATLAB	MATrix LABoratory
ML	Machine learning
MLP	Multilayer Perceptron
MNDCGV	Mean normalized discounted cumulative gain value
PCC	Pearson's correlation coefficient
PSO	Particle swarm optimization
QTL	Quantitative trait loci
ReLU	Rectified linear unit
RR-BLUP	Ridge Regression– Best Linear Unbiased Prediction
SNP	Single nucleotides polymorphism

CHAPTER ONE

INTRODUCTION

1.1 Background information

Genomic selection (GS) has received interest from animal breeders and is rapidly becoming the method of choice of selection in dairy cattle breeding programs around the world (Hayes *et al.*, 2009). The main benefit of GS is its ability to accurately identify genetically superior animals at a young age as a result of the estimation of genomic breeding values derived from genomic information. Genomic estimated breeding values (GEBVs) are considerably more accurate than the traditionally used pedigree indexes for young animals derived solely from information on the genetic merit of their sires and dams (Calus *et al.*, 2013). The increased accuracy and persistence of prediction based on conventional statistical models at a young age make GS more efficient than progeny testing. The latter is hindered by the long generation interval and low reproductive rate thus lowering genetic progress.

The success of the genomic wide selection is related to the choice of the prediction method of the GEBVs. Several statistical models have been adopted for predicting GEBVs, but not limited to the Best Linear Unbiased Prediction (BLUP) (Henderson, 1975), Stepwise Regression, Ridge Regression–BLUP (RR-BLUP) and Bayesian estimation (Meuwissen *et al.*, 2001). These conventional predictive models, however, are faced with statistical challenge related to high dimensionality of marker data. They make strong assumptions and perform linear regression analysis only, thus limiting their accuracies since they do not capture non-linear relationships within genotypes and between genotypes and phenotypes (Ma *et al.*, 2018).

Machine learning, through use of semi- and non-parametric methods for genome-enabled prediction of quantitative traits is able to account for non-additive gene effects and higher non-linearities as well as genotype by environment interactions (Gianola *et al.*, 2006). Several variants of machine learning have been proposed for analyzing large DNA marker datasets to aid in phenotype prediction and genomic selection. Deep convolutional neural network (DCNN) also referred as Deep GS and Artificial neuron network (ANN) (Ehret *et al.*, 2015), are among many machine learning methods that provide good predictive capability with many advanced features (Ma *et al.*, 2018). Deep multi-layered neural network architecture is one of these features with an ability to accommodate noisy data, linear and non-linear relationships between variables and interactions between explanatory variables (Gonzalez-Camacho *et al.*, 2012). Therefore, unlike conventional statistical models, DCNN has an ability to account for non-additive effect in data and can automatically "learn"

complex relationships from the training dataset, without predefined rules (e.g., normal distribution and non-zero variance) (Ma *et al.*, 2018). Deep learning also has the ability to reduce the complexity of high-dimensional marker data through dimensionality reduction using convolution, sampling and dropout strategies. This aims at avoiding overfitting or underfitting of DCNN model thus providing an added advantage.

1.2 Statement of the problem

Conventional genomic prediction methods lead to higher accuracy of GEBVs and identification of superior individual and reduction of generation interval due to utilization of genomic information at an early age, leading to faster genetic gains and reduction of cost of breeding programs. However, conventional statistical models do not handle non-linearity among explanatory variables and association between predictors and target variables which may involve non-additive effects, leading to biased results and low estimated breeding values (EBVs). Conventional linear models also utilize performance data and do not have predictive ability. As a result, generation interval may be increased and the expected gains of a selection program may be negated.

1.3 Objectives of the study

1.3.1 Broad objectives

The overall objective of this study was to contribute to the genetic improvement of indigenous chicken (IC) through selection by adopting machine learning techniques as predictive models.

1.3.2 Specific Objectives

i. To compare the predictive ability of the conventional neural network, conventional statistical models and deep convolutional neural network for body weight of indigenous chicken at 12 weeks based on genome-wide markers of genotypes
ii. To evaluate the prediction performance of the conventional neural network, conventional genomic models and deep convolutional neural network
iii. To evaluate the extent of non-linearity among explanatory variables within and between genotypes and phenotypes using artificial neural network

1.4 Research Questions

i. What is the predictive ability of the conventional neural network, conventional statistical models and deep convolutional neural network for body weight of indigenous chicken at 12 weeks based on genome-wide markers of genotypes?

ii. What are the prediction performances of the conventional neural network, conventional genomic models and deep convolutional neural network?

iii. What is the extent of non-linearity among explanatory variable within and between genotypes and phenotypes using the artificial neural network?

1.5 Justification

Implementation of genomic selection based on a machine leaning model as a predictive mechanism by researchers and commercial institutions facilitates know how of the production capabilities of different species of animals in advance because of increased prediction accuracy and effective predictions, which in turn benefits to farmer. Genomic selection information derived from genotypes also plays an important role in management decisions of selecting superior animals at early age and culling of low producing birds leading to improved profitability of the poultry enterprises. Early identification of superior birds based on performance also facilitate utilization of best genes at early age leading to higher response to selection because of reduced generation intervals and increase accuracy of selection thus cutting on cost which would have been otherwise used for progeny test.

1.6 Limitation of the study

This study was limited to one trait (i.e. body weight) in making predictions based on genotypes. Therefore, results of phenotypic prediction may not be generalized for other traits in the indigenous chicken population.

CHAPTER TWO

LITERATURE REVIEW

2.1 Indigenous chicken (IC) production in Kenya

Domestication of chicken in Kenya started in the western part of Kenya as a result of human migration from Egypt around 100 B.C (Magothe *et al.,* 2012). According to Mwacharo *et al.* (2013), Greco-Roman East Coast trade led to introduction of more chicken to Kenyan Coast. Since then, little effort was put in place to control their mating behaviour, healthcare and feeding. With the current rapid population growth and increasing demand for protein source, poultry farming has received more attention, and more improvement efforts have been put in place. The improvement efforts include crossing with exotic breeds to improve on meat and egg production (Magothe and Kahi, 2011). Population in Kenya is estimated at 32 million birds, of these 75% are IC (MOLFD, 2017). It is estimated that IC contribute 710% of total poultry meat and eggs produced in Kenya (Bett *et al.*, 2012). IC have been considered as a strategy for wealth creation, poverty alleviation, fighting malnutrition and hunger among the poor rural household in developing countries (Okeno *et al.*, 2012), because of their ability to be raised on free range and diseases resistance.

2.2 Production performance of indigenous chicken

In Kenya, several studies have been done on productive and reproductive performance of IC under different production systems (Magothe *et al.*, 2012; Ngeno *et al.*, 2015). Males grow faster and are heavier than females, averaging 2.2 kg and 1.6 kg respectively (Magothe *et al.*, 2012). The age at growth inflection of IC is observed at week 12, the maximum growth rate at week 16 and asymptote body weight at week 24 (Magothe *et al.*, 2010). The dressed weight of IC is 72% of their live weight compared to 75% for commercial culled layers and 80% for broilers (King'ori *et al.*, 2010). For meat-producing poultry, assessing average daily gain (ADG) and residual feed intake (RFI) during the linear growth phase is considered most effective given that high level of growth and development occurs during this period (Case *et al.*, 2012).

2.3 Performance improvement strategies

Achieving good chicken health, housing and feeding requires attention to detail and proper day to day operations as a result of chicken being sensitive to any kind of environmental and nutritional changes. A combination of quality nutrition, veterinary

guidance, housing, genetic improvement and general management ensures chicken have the best possible chance to perform at their maximum potentials.

2.3.1 General management

Indigenous chicken production is an important agricultural activity in that, they provide scarce animal protein in the form of meat and eggs as well as being a reliable source of petty cash for almost all rural communities (Mwalusanya *et al.,* 2002). However, they are not rated highly in the mainstream of national economies because of the lack of measurable indicators (Awuni, 2002). Therefore, management strategies aiming at increasing the IC productivity will play a major role in poverty alleviation, improve household food-security and protein intake of the rural communities. The following strategies has been put in place;

Housing

The Indigenous chicken production system can be classified depending on the production goal attached to the enterprise i.e. commercial and subsistence depending on households' land availability (Magothe *et al.*, 2012). According to Menge *et al.* (2005) and Okeno *et al.* (2012) IC production systems in Kenya are classified into free-range, semi-intensive and intensive systems. In free-range system the birds are mostly left to scavenge for feeds during the day and confined at night. The housing of poultry depends on economic class of household (King'ori, 2010). In semi-intensive system, chickens are kept in small flocks of between five and 50 birds mainly for consumption and sale (Magothe *et al.*, 2012). The birds are left to scavenge during the day and are confined in shelter of moderate cost at night (King'ori, 2010). In intensive system, birds are fully confined in constructed shelters or runs and provided with commercial or home-made feed rations and health care (Magothe *et al.*, 2012). They are enclosed throughout in a deep litter and slatted floors type of housing.

Feeds and feeding

Lack of proper feeds and water results in reduction of birds' immunity thus making them prone to diseases and parasites. In free-range systems birds are left to scavenge for insects, food waste, greens grass, leafy vegetables and any scattered grains (King'ori *et al.*, 2004). They are occasional given supplement feeds of different kinds, but their requirements for minerals, trace elements and vitamins are likely to be met because of the nature of feed they consume. Feed composition and availability vary, depending on the season, locality and production system. The nutrient requirement in poultry are energy and protein, as well as

EGERTON UNIVERSITY
LIBRARY

vitamins, minerals and water. Feeding in poultry changes depending on the age and status (chicken, grower, egg layer, broody hen) of the bird (King'ori *et al.*, 2010). It is always recommended to adopt supplementing chicken feeds with locally available feedstuff in order to cut on the cost of production. Feed quality should be assessed always for safety measures of birds and to avoid exposing them to aflatoxins. Several studies have been done on different feeds and the results have been used in feeding and supplementation of IC e.g. kienyeji mash.

Health

The chicken industry world-wide has been incurring substantial losses through mortality and morbidity due to infectious diseases worldwide (van der Most *et al.*, 2011). Poultry are mostly affected by New Castle Disease, chronic respiratory disease, coccidiosis, fowl pox, salmonellosis and fowl typhoid (Omiti and Okuthe, 2009). Conventional control of this poultry diseases involves high-cost sanitary measures, biosecurity, vaccination programs, medical treatments and genetic resistance (Khobondo *et al.*, 2017). These control mechanisms sometimes result in genetic resistance to disease which varies among breeds with respect to their ability to withstand and survive pathogenic infection (Rahman *et al.*, 1997). Biosecurity measures are not always observed by farmers in rural communities, making disease control to be difficult to control (Khobondo *et al.*, 2017).

2.4 Genetic improvement

Various genetic improvement strategies to enhance the productivity of IC over the past years include breed substitution, crossbreeding/upgrading and selection within population (Ngeno, 2015).

2.4.1 Crossbreeding

Crossbreeding of IC with exotic chicken through cockerel or pullets exchange program is one of the many genetic improvement programs adopted in Kenya, an objective of increasing egg and meat production, farmer's income and protein intake (Magothe and Kahi, 2010). This improvement strategy can be utilized to create new breeds that combine desirable traits from two or more breeds and this take a shorter time period as pure breeding to achieve desired results. Crossbreeding as a strategy is widely used in order to take advantage of heterosis and breed complementarity. Crossbreeding is still experiencing a lot of challenges despite its importance. This challenges are associated with incompatibility of crossbreds with low-input production systems, lack of clear breeding goals and operational breeding programs to ensure availability of breeding stock for the farmers (Ngeno, 2011).

2.4.2 Breed substitution

This improvement strategy am at substituting IC with exotic chicken breeds with an objective of having chickens with faster growth and higher egg production. This strategy has been reported to have contributed to disappearance and displacement of the indigenous breeds (Hanotte *et al.*, 2010). This strategy is faced with challenges related to adaptability of the introduced exotic chickens to the prevailing conditions of production and opposition from global move on conservation of indigenous genetic resource on substituting local breeds with exotic breeds (Ngeno, 2015).

2.4.3 Within breed selection

Pure breeding (within breeding) utilizes selection based on variation among individuals within the population. It takes a longer time to achieve genetic progress, although, the genetic progress is permanent and cumulative unlike in crossbreeding (FAO, 2010). Genetic improvement strategy should focus on within breeding selection to account for failures of crossbreeding (Okeno *et al.*, 2012). It's important to note that previous improvement efforts on genetics have been mainly through crossbreeding (Muasya *et al.*, 2015), although it recorded limited improvements. Crossbreeding also were associated with challenges to high cost of production inputs, availability of breeding stock and unavailability of inputs (Bett *et al.*, 2011).

2.5 Breeding programs

A breeding program is the planned breeding of a group of animals or plants, usually involving several individuals and extending over several generations with an aim of making genetic improvement. Before introduction of genomic selection Poultry breeding programs have been based mainly on observed or measured phenotypes and pedigree information. Breeding program involves characterization of production systems which is the first step to be undertaken for the purpose of identifying threats and opportunities for the holistic improvement of IC breeding program (Okeno *et al.*, 2011). This plays a key role in understanding the production system and management practices of farmers which are important in developing improvement strategies. Definition of a breeding goal is the second step in developing a breeding program. Breeding goal is based on selection schemes to be considered in deciding on the breeding program to be adopt (Okeno *et al.*, 2012). The genetic improvement gain generated from the breeding program is disseminated to the commercial

farmers by selling the birds not selected in the nucleus to farmers in the commercial tier for breeding purposes.

Phenotype based breeding programs have not yielded good results because of lack of clear breeding objectives and lack of constant supply of breeding stock to farmers (Ngeno, 2011). They are also faced with the challenge of generation lag and the difficulty of estimating dominant effects (Ibanez-Escriche and Gonzalez-Recio, 2011). To overcome these challenges adopting GS models with ability to easily accommodate non-additive effects is recommended (Bolormaa *et al.*, 2015). The main goal of adopting GS in breeding program is the selection of commercially important traits that can only be directly recorded for one sex at an early stage of life resulting to an increase in genetic response per year and a decrease in the inbreeding rate (Wolc, 2014).

2.6 Genomic selection

Genomic selection is a breeding strategy whereby phenotypes are predicted based on genome-wide markers of genotypes using conventional statistical model. Unlike MAS which utilizes only the fewer markers which have large effects on the trait of interest used to predict breeding values, genomic selection uses whole-genome high-density markers to predict the breeding values of each genotyped individual (Goddard *et al.*, 2007). Thus genomic-based selection programmes have advantages over MAS because of its ability to allow direct selection for sex-limited traits, sex-influenced traits, traits that are expensive to measure or traits measured at slaughter (Fulton, 2012). Contrary to MAS, GS uses all estimated effects of marker loci regardless of whether they are significantly associated with the phenotype or not to predict GEBV of each individual in the population under selection (Singh and Singh, 2015). According to Mukhopadhyay *et al.* (2013), GS is a two-step process i.e. estimation of the marker effects that are in linkage disequilibrium (LD) with QTL in training population. The training population consists of animals with both phenotypic and genotypic information, then the obtained marker effects are used to predict the GEBV of individuals in the validation (test) populations based on their SNPs only. Accuracy of prediction is then verified with the observed phenotypic records of the testing population.

The genome sequence for poultry was first available in 2004 through a project funded by the US National Institutes of Health. This enabled sequencing of the chicken genome, using DNA from an inbred Jungle Fowl from the University of California, Davis (International Chicken Genome Sequencing Consortium, 2004). Since then, the development of SNP panels for poultry has been swift with a second build released in 2006 that corrected

some of the deficiencies found with the first version (Fulton, 2012). The advent of whole-genome sequencing and high-density SNP genotyping technologies has resulted in increased marker density and facilitated GS prediction in a number of domesticated animals including chickens.

Economic feasibility in genomic selection is one of the factors to be considered due to the anticipated cost of genotyping large samples. Obtaining higher density SNP panels for GS is an expensive process, therefore to address cost of genotyping, new genotyping strategies have been proposed, such as combining the high-density genotyping of selected individual for breeding with the low-density genotyping of selection candidates (Ibanez-Escriche and Gonzalez-Recio, 2011). Also, low-density genotypes can be used and missing genotypes imputed using the genotype information from relatives (Heidaritabar, 2016). Therefore, application of GS in breeding programmes as proposed by Meuwissen *et al.* (2001) coupled with development of new methodologies and statistical analysis tools to integrate genomic information for better outcome of prediction accuracy are recommended.

2.7 Prediction accuracy of genomic estimated breeding values

Genomic prediction technique is applied early in breeding programs as a way of increasing overall selection pressure in a breeding program (McDowell, 2016). Prediction accuracy act as a measure of genomic prediction models, since it assesses how reliable a future phenotype of target individuals can be predicted. The accuracy to a larger extent depends on the parameters such as sample size of the reference, its genetic structure and choice of a statistical model to capture the relationship between the genetic architecture of a trait and the underlying marker calls in a panel of high-density marker data (McDowell, 2016).

Genomic selection to large extent focuses on traits with the low accuracy of selection (heritabilities) and traits measured in one sex like milk production (Ibanez-Escriche and Gonzalez-Recio, 2011). To obtain acceptable accuracy for low heritable traits, a larger number of genotypes and phenotypes should be used (Calus and Veerkamp, 2007). According to Akanno *et al.* (2014) multi-generational training population of 775 to 1000 individuals for moderate to highly heritable traits and not less than 1000 individuals for lowly heritable traits would achieve a considerable level of accuracy. Therefore, accuracies of GEBV can be improved with denser SNP panels (Meuwissen and Goddard, 2010) and recruiting large population.

2.8 Strategies to improve the accuracy of genomic estimated breeding values

To retain the good level of accuracy, breeders are encouraged to use denser panels, enlarge the size of the reference population, or re-estimate marker effects frequently (Zhang *et al.*, 2011). Achieving higher accuracy of genomic estimated breeding values (GEBVs) requires the reference population to be very large of both genotyped and phenotyped individuals (Wang *et al.*, 2018). For the best result, a reference population with genotypic and phenotypic data for training the prediction model should have a multi-generational data structure with greater than 2000 individuals (Akanno *et al.*, 2014). Although the accuracy of the GEBVs depends on marker density, it does not persist across all generations. Genomic selection methodologies have been benchmarked to show superiority over different available models in accuracy improvements. This exploration of different models has led to adoption of artificial intelligence concept using machine learning in animal breeding in an effort of improving on the accuracy of prediction.

2.9 The transition from traditional breeding schemes to genomic selection era

In phenotypic breeding schemes, the aim was to improve economically important traits by adopting selection based on pedigree information and recorded phenotypes. The superior individuals are then selected according to the rank order of their estimated breeding values (EBVs) obtained from the best linear unbiased prediction models (BLUP) (Zhang *et al.*, 2011). Even though conventional predictive models have been applied successfully for many quantitative traits in ongoing breeding programs, making genetic progress is still difficult for low heritability traits and traits expressed at late stages of life (Heidaritabar, 2016). With advances in molecular genetics in particular identification SNP, better understanding of quantitative trait genetic architecture has been made possible. Detection of QTL has been successful due to their association between genetic markers and phenotypes. Discovery of QTLs controlling important traits has made marker-assisted selection (MAS) possible leading to increased rates of genetic gain (Heidaritabar, 2016). Unfortunately, the success of MAS is limited when traits are controlled by single genes, minor traits and different SNPs associated with different traits.

2.10 Conventional genomic selection models, ability and limitations

Marker and phenotype information availability has made it possible for GS to be used to establish associations between markers and phenotypes (Wang *et al.*, 2018). This is aimed at utilizing genome-wide markers to compute genomic estimated breeding values

(GEBV) to be applied in prediction of performance. The success of genomic selection in making genomic predictions is related to the choice of the prediction method of the GEBV. Several statistical models have been adopted for predicting GEBV, such as the Best Linear Unbiased Prediction (BLUP) (Henderson, 1975), Stepwise Regression; Ridge Regression–BLUP (RR-BLUP) and Bayesian Estimation (Meuwissen *et al.*, 2001). These conventional predictive models, however, are faced with statistical challenge related to high dimensionality of marker data and typically make strong linear assumptions which in turn limit their accuracies because they do not have ability to capture complex non-linear relationships within genotypes and between genotypes and phenotypes (Ma *et al.*, 2018). The conventional statistical models often used to predict GEBVs provide different assumptions on the number of individual animals used and marker effect. Therefore, these methods generally differ in assumption on the models associated with a quantitative character (Rusende *et al.*, 2008). For instance, RR-BLUP, assumes normality with non-zero variance. Any violation of these assumptions might cause deviations between the observed and predicted accuracies which in turn affect accuracy of prediction. Conventional genomic predictions tools also assume linearity of SNP effects leading to limited accuracy and biased GEBVs.

To counter on these challenges, machine learning models such as artificial neural network (ANN) and deep convolutional neural network (DCNN) are recommended because they have many advanced features. Deep multi-layered neural network architecture is one of these features with an ability to accommodate noisy data, linear and non-linear relationships between variables and associations between explanatory variables (Gonzalez-Camacho *et al.*, 2012).

2.11 Genomic selection: from linear model to complex models

Artificial intelligence (AI) is a branch of Computer Science which has demonstrated its importance in scientific field through machine learning methods (ML). These ML methods can automatically perform data analysis and can find a hidden intrinsic pattern from big data which are not possible for a human (Islam, 2017). It provides a new avenue to model noisy data and have been suggested as an attractive, non-linear alternative to traditional statistical analyses (Acevedo *et al.*, 2009). This is attributed to their ability to model complex and non-linear relationships without having to fulfil the strict assumptions required by conventional parametric model (Favaro *et al.*, 2014).

According to Lee *et al.* (2017) genomic-based prediction requires modeling of the association between markers and phenotypes. Models such as ANN and Deep GS with ability

to capture non-linearity and interactions between marker features have the capacity to achieve higher accuracy of GEBVs. Considering the nature of biological data which are often high-dimensional and noisy, traditional ML methods such as support vector machine, random forest and Bayesian network which depend on hand-turned features are not able to provide a suitable technique to handle such a data (Islam, 2017). Conventional ML application in genome-enabled prediction has received a lot of attention both in animal (VanRaden *et al.*, 2009) and plant breeding (Crossa *et al.*, 2010). Therefore, this provides an opportunity for study of the application of ML models in genomic predictions of IC.

2.12 Machine learning in genomic enabled prediction

Although estimation of breeding values has evolved from being based on traditional pedigree and phenotype information to genomic information, the main concern to breeders has been on developing predictive models that go beyond additive thus improves on traditional genomic models (Lopes, 2016). Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. It focuses on the development of computer programs that can access data and use it to learn through experience. ML is an alternative approach for prediction and classification, capable of dealing with the dimensionality problem in a computationally flexible manner (Abdollahi-Arpanahi *et al.*, 2020). There are many types of ML methods classes, i.e. those that perform classification, regression, clustering, dimension reduction and deep learning

2.12.1 Artificial Neural Networks

Artificial Neural Networks (ANN) are mathematical models which have been described as additional models in performance prediction in the field of animal breeding (Gianola *et al.*, 2011). It mimics the way the biological neural network of human brain works, has ability to learn from experience to improve its performance and can adapt themselves to change in the environment (Njubi *et al.*, 2010). The neurons process the presented inputs and matching outputs in a supervised manner and can facilitate the combination of different input types which are important in accommodating highly non-linear biological data (Hosseinia *et al.*, 2007). The first application of ANN in the field of animal breeding was in a research done by Macrossam *et al.* (1999), where they used ANN as a method for optimizing mating allocation to maximize production traits in the Australian dairy industry. Since then, application of machine learning is increasing slowly, owing to their flexibility in prediction, forecasting and taking care of interaction between parameters. Even with this model

promising predictive ability, its application in animal breeding is still scarce (Gorgulu, 2012), considering huge dataset available to be analyzed.

2.12.2 Deep learning

Deep learning (DL) is one of the current developed machine-learning techniques which has shown its promising predictive capabilities, this is attributed to the deep multi-layered neural network architecture feature (Ma *et al.*, 2018). This feature has a large number of neurons that are used to capture complex, non-linear relationships in large datasets (LeCun, *et al.*, 2015). This technique is able to address the shortcomings of conventional machine learning methods such as noisy and high-dimensionality of genomic data (Islam, 2017). The genetic analysis of continuous quantitative traits does not avoid the current discussions on artificial intelligence (AI), like renewed interest in deep learning (DL) methods such as Multilayer Perceptrons (MLPs) and Deep Convoluted Neural Networks (DCNNs) (Bellot *et al.*, 2018). Deep Convoluted Neural Networks are a specialized type of neural network where inputs are associated which each other and their hidden layers consist of convolutional layers, max-pooling layers (sampling layer), fully connected layers and normalization layers. Convolutional layers are considered a strong pattern detector of local features of a given dataset (Islam, 2017). During the training phase, a DCNN automatically learns the coefficients of the filters which is the combination of the input values where the weights are the same for all genotypic markers (Bellot *et al.*, 2018). Max-pooling layer creates higher-level features that can help smooth the noisiness in training data and reduces the size of the input feature thus avoiding overfitting and makes the model become invariant to input features (Islam, 2017). Deep Convoluted Neural Networks combines the several layers of convolutions with non-linear activation functions a commonly used activation function is rectified linear unit (ReLU).

Since accuracy of genomic prediction depend on an appropriate choice of a statistical model to capture the relationship between the genetic architecture of a trait and the underlying marker calls in a panel of high-density marker data (McDowell, 2016), it is advisable to adopt ML techniques which has ability to account for non-linearity and accommodate high dimensional data thereby improving on accuracy of prediction. Improved accuracy will play a key role in livestock breeding companies who practice genomic selection since small percentage improvements in accuracy could generate much larger improvements in genetic gain over the lifetime of a breeding program.

CHAPTER THREE

PREDICTION OF INDIGENOUS CHICKEN PHENOTYPES FROM GENOTYPES: COMPARISON BETWEEN MACHINE LEARNING AND CONVENTIONAL LINEAR MODELS

Abstract

Genomic selection is a new breeding strategy which is rapidly becoming the method of choice for selection. It is useful in predicting the phenotypes based on genome-wide markers of genotypes using conventional predictive models such as ridge regression BLUP. However, these conventional predictive models are faced with a statistical challenge related to the high dimensionality of marker data, inter and intra-allelic interactions and typically make strong assumptions. Machine learning models can be used as an alternative in the prediction of phenotypes due to their ability to address these challenges. Therefore, the aim of this study was to compare the predictive ability of machine learning using deep convolutional neural network (DeepGS), conventional neural network (Artificial neural network), conventional statistical predictive model ridge regression best linear unbiased (RR-BLUP) model and combination of DeepGS and RR-BLUP (Ensemble model) in predicting body weight (BW) of indigenous chicken based on genome-wide markers. The Pearson correlation coefficient (PCC) for body weight (BW) results from this study for the four models were 0.891, 0.889, 0.892 and 0.812 for DeepGS, RR-BLUP, Ensemble and ANN respectively. This showed that DeepGS did not yield significant difference ($P>0.05$) from the other models, therefore, it can be used in complement to the commonly used conventional models. For individuals with higher phenotypic values, the PCC results showed a drastic decrease in the performance for DeepGS, RR-BLUP, Ensemble and ANN from 0.891, 0.889, 0.892, 0.845 to 0.315, 0.466, 0.342, 0.518 respectively. Therefore, focus should be put on individuals with higher phenotypic values. This study concludes that DeepGS, Ensemble, RR-BLUP and ANN models can be used interchangeably in making phenotypic predictions

Key words: *Artificial neural network, DeepGS, Genomic prediction, Machine learning*

3.1 Introduction

Before introduction of genomic selection, phenotypes have been used to select superior animals for breeding in genetic improvement programs. However, introduction of genomic selection (GS) technology as a breeding strategy tool has led to revolution in genetic improvement programs. Genomic selection is a new breeding strategy which is rapidly becoming the method of choice of selection for animal breeders (Hayes *et al.*, 2009). This is attributed to its ability to accurately identify genetically superior animals at a young age based on genomic estimated breeding values (GEBVs). Genomic selection is useful in predicting the phenotypes of quantitative traits based on genome-wide markers of genotypes using conventional predictive models. GS has ability to accurately identify genetically superior animals at a young age. The increased accuracy and persistence of prediction based on conventional statistical models at a young age make GS more efficient than progeny testing (Calus *et al.*, 2013). The latter is hindered by the long generation interval and low reproductive rate thus lowering genetic progress.

The models adopted for predictions, as key for the success of genomic selection (GS). Several statistical models have been adopted for predicting GEBV, such as the Best Linear Unbiased Prediction (BLUP) (Henderson, 1975), Stepwise Regression, Ridge Regression–BLUP (RR-BLUP) and Bayesian Estimation (Meuwissen *et al.*, 2001). However, these conventional statistical models do not consider non-linearity among explanatory variable. The association between predictors and target variables which may involve non-additive effects is also ignored, leading to biased results and low predictive abilities. The models make strong assumptions by performing linear regression analysis only thus limiting their predictive performance since they do not capture non-linear relationships within genotypes and between genotypes and phenotypes (Ma *et al.*, 2018).

Machine learning algorithms provide a new avenue to model noisy data and have been suggested as an attractive, non-linear alternative to traditional statistical models according to literature (LeCun, *et al.*, 2015; Singh *et al.*, 2016; Islam, 2017). They have shown to be superior to conventional statistical models. This is attributed to their ability to model complex and non-linear relationships among acoustic parameters without having to fulfil the strict assumptions required by the conventional parametric model (Favaro *et al.*, 2014). They also have ability to account for non-additive effect in data and can automatically "learn" complex relationships from the training dataset, without pre-defined rules (Ma *et al.*, 2018). The focus of this study was, therefore, to compare the predictive ability of machine learning using deep convolutional neural network (DeepGS), conventional neural network

(Artificial neural network) and conventional statistical predictive model ridge regression best linear unbiased (RR-BLUP) model in predicting body weight (BW) of indigenous chicken based on genome-wide markers.

3.2 Materials and methods

3.2.1 Data source

The data used for this study were obtained from a wide-genome study done by Yuan *et al.* (2018) in four Chinese indigenous chicken breeds. The study, used 394 birds from four indigenous breeds consisting of two typical low-body-weight breeds (Chahua chicken and Silkie chicken) and two intermediate and high body-weight breeds (Beard chicken and Langshan chicken).

3.2.2 Phenotyping

The birds were randomly selected from the original conserved population and bred by performing artificial insemination of hens with sperm pools. The obtained fertilized eggs were incubated and the chicks were of an approximately equal number of cocks and hens. The hens and cocks of the different breeds were phenotyped for body weight. Live body weight (BW) was measured at hatch and every week until 12 weeks of age.

3.2.3 Sample collection, DNA extraction and genotyping

Blood samples were collected from 394 birds at 12 weeks of age. The DNA extraction was done using the phenol-chloroform method and diluted to 50 ng/ml. Genotyping was performed using Illumina 60K Chicken SNP BeadChip (Groenen *et al.*, 2011). Quality control was conducted on all birds across four breeds using customized scripts in R software version 3.3.0 (R Core Team, 2017). Single-nucleotide polymorphisms (SNPs) filtering were done using the following criteria: individual samples were excluded with call rates < 0.9 and minor allele frequency (MAF) < 0.05. After imposing the quality control checks, a total of 46211 SNPs remained.

3.2.4 Genotype preprocessing and coding

The 46211 markers had some missing genotypes therefore, missing markers were imputed using A.mat function of the RR-BLUP package in R software. Markers with 50% missing genotypes were not imputed, leaving 26698 markers for data analysis. Genotypes were coded as {0 1 2} based on R code script by Eva KF Chan (http://www.evachan.org/rscripts.html).

3.2.5 Data analysis

Conventional Neural Network

The conventional neural network model adopted for this research were a feed-forward artificial neural network (ANN). The data were split into training, validation and test data with 80% used for training, 10% being used for validation and remaining 10% used for testing. Sampling were random in order to avoid any selection bias in the dataset. The Prediction (testing) was done with the remaining data sample that were not used during the training and validation phase. Data preprocessing and analysis were done using the R software, while the construction and training of neural networks were performed using MATLAB. A three-layered, fully-connected, feed-forward artificial neural network was built using a MATLAB software function "feedforwardnet", with an 18-32-1 architecture (18 nodes in the first hidden layer, 32 nodes in the second hidden layer and one node as an output layer) (Ma *et al.*, 2018). Error optimization was done using back-propagation algorithm (Rumelhart, *et al.*, 1986). Nodes in one layer was fully connected to all nodes in the next layer to form a multilayer ANN architecture.

Conventional statistical model (Ridge Regression-Based Linear Unbiased Prediction (RR-BLUP)

Ridge Regression-Based Linear Unbiased Prediction (RR-BLUP) were adopted as a conventional statistical model. This model was built using the standard linear regression formula, given the genotype matrix as Z $(n \times p)$ and the corresponding phenotype vector as Y $(n \times 1)$;

$$Y = \mu + Z_g + \ell$$

where, n = individuals, p = markers, μ = mean of phenotype vector Y, g $(p \times 1)$ = a vector of marker effects and $\ell(n \times 1)$ = vector of random residual effects. The model was implemented using the function of mixed.solve in R software.

Deep convolutional neural network model (DeepGS)

The DeepGS model was constructed using the deep learning technique as described in a project done by Ma *et al.* (2018) available at GitHub (https://github.com/cma2015/DeepGS). Where an 8-32-1 architecture with one input layer, one convolutional layer with 8 neurons, one sampling layer, 3 dropout layers, 2 fully-connected layers and one output layer were constructed. The input layer receives the genotypic markers of a given individual in the $1 \times p$ matrix, where p is the number of

genotypic markers. The first convolutional layer filters the input matrix with 8 kernels (weights) which are each 1 × 18 in size with a stride size of 1× 1, followed by a 1 × 4 max-pooling layer with a stride size of 1 × 4. The output of the max-pooling layer is passed to a dropout layer with a rate of 0.2. The first fully-connected layer with 32 neurons is used after the dropout layer to join together the convolutional characters with a dropout rate of 0.1. Rectified linear unit (ReLU) activation function is applied in the convolutional and first fully connected layers. First fully-connected layer output is then fed to the second fully-connected layer with one neural and a dropout rate of 0.05. Using a linear regression model, the output of the second fully-connected layer is finally connected to the output layer which presents the predicted phenotypic value of the analyzed individual.

The Deep GS model was trained on the training set and validated on the ten-fold cross-validation. Parameters were optimized with the back-propagation algorithm, by setting the number of epochs to 6,000, the learning rate to 0.01, the weight decay to 0.00001 and the momentum to 0.5. The loss function minimized was taken as the mean absolute error (Mae) index (Ma *et al.*, 2018);

$$Mae = \frac{1}{m}\sum_{k=1}^{m} | predict_k - obs_k |$$

where, m is the number of individuals in the training dataset, and $predict_k$ and obs_k represent the predicted and observed phenotypic values of the k^{th} individual, respectively. Deep GS were implemented using the central processing unit (CPU) based on Deep Learning framework MXNet.

Ensemble model

An ensemble GS model (E) was constructed using the ensemble learning approach by linearly combining the predictions of DeepGS (D) and RR-BLUP (R), using the formula described by Ma *et al.* (2018).

$$Predict_E = (W_D \times predict_D + W_R \times predict_R) / (W_D + W_R)$$

where $Predict_E$ = predictions of Ensemble (E), WD = weight of DeepGS, WR = weight of RR-BLUP, $predict_D$ = predictions of DeepGS (D), $predict_R$ = prediction of RR-BLUP

For each fold of cross-validation, parameters (W_D and W_R) were optimized on the corresponding validation dataset using the particle swarm optimization (PSO) algorithm.

3.2.6 Statistical analysis

For each BW trait of a single breed, boxplots were generated by R version 3.3.0 (R-Core Team, 2017) to screen for outliers. To determine predictive ability of model, the Pearson's correlation coefficient (PCC) and corresponding significance level of p-value = 0.05 was calculated using R programming language for predicted and actual BW. Also the significance level of the difference between paired samples were examined using the student's t-test using R software.

3.3 Results

The results on Figure 1 show the regression-based predictive ability for body weight based on the four models (DeepGS, rrBLUP, Ensemble and ANN). The experimental results showed that DeepGS, RR-BLUP and Ensemble model yielded a high correlation of 0.891, 0.889 and 0.892 respectively based on Pearson's correlation coefficient (PCC) between predicted phenotypes and observed body weight at week 12. Artificial neural network yielded relatively lower PCC of 0.813 in comparison with the other models. Generally, all the methods performed similarly, although DeepGS and Ensemble model was slightly better.

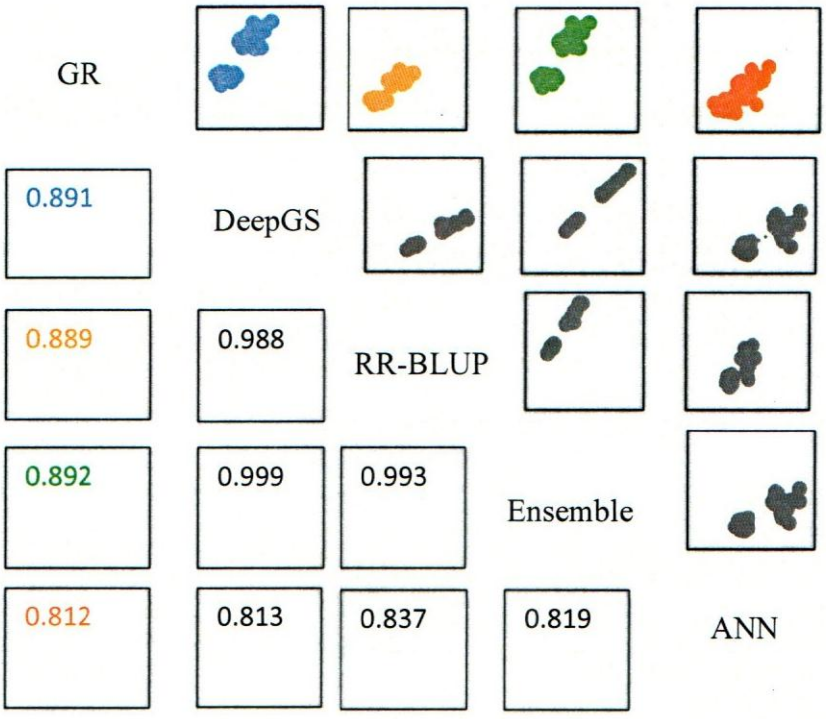


Figure 1: Pearson's correlation coefficient (PCC) analysis results between observed and predicted phenotypic values for the four GS models. The correlation results from different GS models are shown in different colours.

The absolute errors between observed and predicted phenotypic values for the four models are presented in Figure 2. The results showed that DeepGS and Ensemble models

recorded lower p-values of 1.192e-4 and 4.65e-4, respectively, on paired sample t-test compared with 0.0746 for RR-BLUP and 0.0506 for ANN.

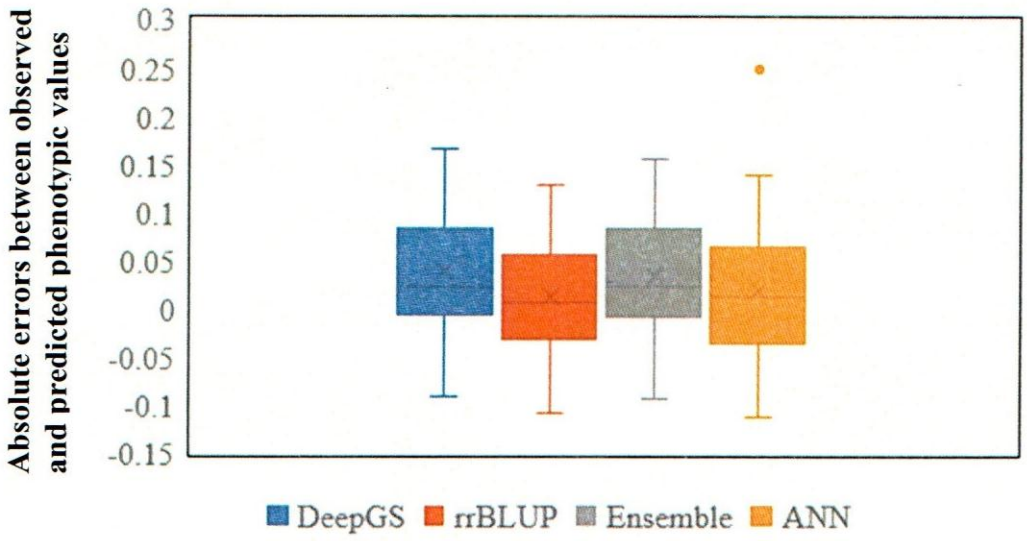


Figure 2: Box plot of the absolute errors between the observed and predicted phenotypic values.

Mean separation was also done to ascertain if the means for body weight obtained from the four models were significantly different from the actual BW mean using the least significant difference (LSD) at alpha = 0.05, the results showed that the means for the four models were not significantly different as presented in Table 1.

Table 1: mean separation using the least significant difference (LSD) of the GS models

Model	Mean	N
Actual (BW)	0.34416[a]	39
RrBLUP	0.33032[a]	39
ANN	0.32424[a]	39
Ensemble	0.30929[a]	39
DeepGS	0.30419[a]	39

Means with different letters within a column are significantly different

The study also investigated how the top-ranked 200 individuals with the highest BW performed as per prediction of phenotypes is a concern. The PCC value results show a drastic

decrease in correlation between observed and predicted phenotypic values equivalent to 0.315,0.466, 0.342 and 0.518 for DeepGS, RR-BLUP, Ensemble and ANN respectively as shown in Table 2.

Table 2: **PCC between observed and predicted BW based on DeepGS and RRBLUP model**

	Observed BW	DeepGS	RR-BLUP	Ensemble	ANN
Observed BW	1	0.315	0.466	0.342	0.518
DeepGS		1	0.935	0.998	0.022
RR-BLUP			1	0.954	0.252
Ensemble				1	0.059
ANN					1

EGERTON UNIVERSITY
LIBRARY

3.4 Discussion

The aim of this study was to compare the predictive ability of machine learning (ML) using deep convolutional neural network (DeepGS), conventional neural network (Artificial neural network) and conventional statistical predictive model ridge regression best linear unbiased (RR-BLUP) model in predicting body weight (BW) of indigenous chicken based on genome-wide markers. To achieve this, data was split into training, validation and testing datasets, genotype markers were used as input features. The PCC results showed that DeepGS yielded almost the same phenotypic predictions compared to the other three linear models (Ensemble, RR-BLUP and ANN) by a significant margin (P>0.05). This is true because the mean separation results for predicted BW from the four models showed that there was no significant difference between the four models. Similar results were reported by Bellot *at al.* (2018) in the complex human genome, Ma *et al.* (2018) study also reported that DeepGS can be used as a complement to the RR-BLUP model.

The reasons why DeepGS model performed in a similar way RR-BLUP may be attributed to the number of allele frequencies of genotypes which are always unbalanced and to a large extend the coding system adopted, which always takes a simple {0 1} or {0 1 2} format. The nature of data, the trait type being predicted and the network architecture also play a crucial role in determining the output of machine learning model, this is because most

of the data are noisy which will intern affect the outcome (Bellot *at al.,* 2018). For an instant the PCC on ANN in this study was 0.812 which were way far much higher and contrary to 0.378 reported by Ma *et al.* (2018), this may be as a result of different algorithm adopted, input data set and data splitting criteria. Okut *et al.* (2013) reported correlation training set ranging from 0.776 to 0.858 almost similar to what this study reported after using different training algorithms, different activation functions and different numbers of neurons in hidden layers. Therefore, there is a need for a study to explore how related are chosen markers to functional regions related to the trait being studied since the use of unknown or non-specific markers will lead to the introduction of noise into the dataset (Besic *et al.*, 2017).

The PCC results showed that combining DeepGS and RR-BLUP (ensemble model) resulted in better performance compared to RR-BLUP corresponding to 0.892 and 0.889 respectively. This is because validation dataset of ensemble uses PSO algorithm, which has the capability of parallel searching on very large spaces of candidate solutions, without making assumptions about the problem being optimized (Kennedy and Eberhart, 1995). Therefore, the ensemble model can play the same role as DeepGS and BLUP model. This study also estimated the absolute error between observed and predicted phenotypic value where the results showed that DeepGS and Ensemble recorded lower p-values compared to RR-BLUP and ANN models. This is in consensus with what Ma *et al.* (2018) reported. Lower error indicates a good agreement between observed and predicted phenotypic value. Therefore, the DeepGS and Ensemble model can be used in complementarily for phenotypic predictions.

This study also investigated how the top-ranked 200 individuals performed. The results indicated that neither DeepGS nor RR-BLUP performed particularly well on body weight prediction since the PCC results were 0.315 and 0.466 respectively. This is attributed to the model overfitting or underfitting caused by too much noise in the genotypes used. This was similar to results of the study done by Ma *et al.* (2018) that reported almost similar decreasing trend on the prediction of top-ranked 400 individuals with the longest grain length. Therefore, it's important to concentrate on individuals with high phenotypes because they play an important role in selection just like selection in a traditional breeding program where an individual with highest phenotypic values is given priority as replacement stock.

3.5 Conclusion

The results from this study finds out that the margin of difference between the tested models was negligible and there was no statistical difference between the four models. The

small differences observed might have arisen due to the nature of data used, different training algorithm, network architecture and a number of traits being analyzed. Therefore, DeepGS, Ensemble, RR-BLUP and ANN models can be used interchangeably in making phenotypic predictions.

CHAPTER FOUR

RANKING INDIVIDUALS BASED ON PREDICTED PERFORMANCE USING MEAN NORMALIZED DISCOUNTED CUMULATIVE GAIN VALUE

Abstract

Accuracy of genomic prediction always relies on an appropriate choice of a statistical model to capture the relationship between the genetic architecture of a trait and the underlying marker calls in a panel of high-density marker data. However, the ranking problem has become an important subject in machine learning (ML) models, due to its widespread applications in many decision-making processes because the measures of rank quality are usually based on sorting, which is not directly optimizable. To counter this, mean normalized discounted cumulative gain value (MNDCGV), a standard quality measure in information retrieval with capabilities of ranking individuals according to breeding values has been proposed. Few studies have emphasized on the ranking of individuals based on predicted phenotypic values using MNDCGVs but none have been reported in animals. The focus of this study, therefore, was, to evaluate the prediction performance of DeepGS, RR-BLUP and Ensemble GS models using MNDCGV. The MNDCGV results showed the accuracy of GEBVs estimated using DeepGS was approximately equal to 0.75~0.78, RR-BLUP 0.66~0.76 and Ensemble 0.76~0.79 as a result of top-ranked alpha increasing from 10% to 70%. The Ensemble and DeepGS model outperformed the conventional RR-BLUP model by a significant margin ($P<0.05$), therefore they can be used as a supplement to RR-BLUP. Thus, Ensemble and DeepGS models can be given a top priority as GS model and as an alternative to conventional GS models in predicting the performance and ranking of individuals with high breeding values to be used for selection purposes in indigenous chicken breeding programs. Ensemble model performed very well in ranking individuals with better performance compared to DeepGS and RR-BLUP, with improvement values of 0.01 and 0.11, respectively. Thus, Ensemble model can be given a top priority as GS model for performance improvement.

Keywords: *Cross-validation, Genomic selection, Mean normalized discounted cumulative gain value,*

4.1 Introduction

Selection criteria based on phenotypes have led to substantial genetic improvement as a result of its ability to utilizes phenotypic or pedigree information in prediction of

breeding values. However, with technological advances, there has been a growing interest in developing new statistical models and algorithms capable of predicting untested phenotypes in methods commonly used in genomic selection (Gonzalez-Camacho *et al.*, 2018). This led to the development of genome-enabled prediction models in animal breeding (VanRaden *et al.*, 2008) such as the Stepwise Regression, RR-BLUP and Bayesian Estimation. The conventional genomic prediction models predict phenotypes using all available single nucleotide polymorphism (SNP) marker data collected from a population using statistical models to predict the marker-trait associations (Meuwissen *et al.*, 2001).

The accuracy of genomic prediction depends on an appropriate choice of statistical model. The model chosen should be able to capture the relationship between the genetic architecture of a trait and the underlying marker calls in a panel of high-density marker data (Mcdowel, 2016). Therefore, models such as DeepGS with ability of incorporating interactions between marker features have the capacity to achieve higher accuracy by capturing non-additive effects and noisy data. Ma *et al.* (2018) proposed an Ensemble model which intergrades DeepGS and RR-BLUP model where parameters are optimized using the particle swarm optimization (PSO) algorithm. This model was developed by inspiration from the social behaviour of bird flocking or fish schooling (Kennedy and Eberhart, 1995). Among many different statistical models developed, no much variation in accuracy of predicted has frequently been observed and reported (Roorkiwal *et al.*, 2016).

Application of machine learning (ML) in genome-enabled predictions as a means of improving accuracy has been accelerated in recent years (Bellot *et al.,* 2018). The increased use of ML has been due to a growing interest in using semi- and non-parametric models such as, deep convolutional neural network, artificial neural network and ensemble model for genome-enabled prediction of quantitative traits to account for non-additive gene effects and higher non-linearities as well as genotype-environment interactions (Ehret *et al.*, 2015). According to Blondel *et al.* (2015) predictive accuracy of most models are typically assessed using the Pearson's correlation coefficient (PCC) between observed trait values and the predicted trait values despite poor correlations with ranking accuracy.

In order to select the most favourable individuals to be used in genetic improvement programs, it is important to correctly rank individuals from the most favourable to least favourable based on accurately predicting breeding values (Blondel *et al.*, 2015). The focus of this study therefore, was to evaluate the prediction performance of GS models using the mean normalized discounted cumulative gain value (MNDCGV) as described by (Blondel *et*

al., 2015) which have been described as rewarding more strongly models which assign a high rank to individuals with high breeding value.

4.2 Materials and methods

4.2.1 Phenotypes and genotypes source

The phenotypes and genotypes data used for this study were obtained from a study done by Yuan *et al.* (2018) from Chinese indigenous chicken breeds. The study used 394 birds where hens and cocks of the different breeds were phenotyped for body weight (BW). Phenotyping for BW was measured at hatch and every week until 12 weeks of age. For genotypes, blood samples were collected for DNA extraction for all birds used in the study using phenol-chloroform method and diluted to 50 ng/ml. Genotyping was performed using Illumina 60K Chicken SNP BeadChip (Groenen *et al.*, 2011). Quality control was done using customized scripts in R software version 3.3.0 (R Core Team, 2017).

4.2.2 Missing markers and coding

Imputation using A.mat function on the RR-BLUP package installed in R software (R Core team) were done to impute missing genotypes. This function takes care of markers with 50% missing genotypes. Genotypes were coded into {0 1 2} based on R code script.

4.2.3 Data analysis

Cross-validation was used to evaluate the prediction performance of GS models as proposed by Crossa *et al.* (2016) and Gianola and Schon (2016). Five-fold cross-validation was used for this study, in which individuals in the dataset were randomly partitioned into five groups of approximately equal size. Using genotypic and phenotypic data, the GS models were trained and validated from four groups with 90% individuals for the training set and 10% for the validation set. The trained GS model was used to predict phenotypic trait of individuals from the remaining group using only genotypic data. This process was repeated five times until each group was used once for testing. The predicted phenotypic traits values were combined for performance evaluation. The entire five-fold cross-validation experiment was repeated ten times with different seeds used to shuffle the order of individuals in the original dataset. Therefore, for each given level of alpha, this procedure produced ten different mean normalized discounted cumulative gain values (MNCGVs), and the average was used as the final result.

The predicted performance for determining individuals with high phenotypic values to be used for selection for each GS model was assessed by measuring the MNDCGVs as

described by Blondel *et al.* (2015). Given n individuals, the predicted and observed phenotypic values form a $n \times 2$ matrix of score pairs *(X, Y)*. The MNDCGVs for selecting the top-ranked k^{th} individuals were calculated in an iterative manner as follows:

$$MNV(K,X,Y) = \frac{1}{K}((K-1)MNV(K-1,X,Y)) + (\frac{\sum_{i=1}^{k} y(i,X)d(i)}{\sum_{i=1}^{k} y(i,Y)d(i)})$$

where, $d(i) = 1/(\log_2 i + 1)$ is a monotonically decreasing discount function at position i; $y(i,Y)$ is the i^{th} value of observed phenotypic values Y sorted in descending order, here $y(1,Y) \geq y(2,Y) \geq \ldots \geq y(n,Y)$; $y(i,X)$ is the corresponding value of Y in the score pairs (X, Y) for the i^{th} value of predicted scores X sorted in descending order; MNV is the mean normalized value of selecting the top individual. Thus, MNDCGV has a range of 0 to 1 when all the observed phenotypic values are larger than zero; a higher MNV (k, X, Y) indicate a better performance of the model to select the top-ranked k individuals with high phenotypic values.

4.2.4 Statistical analysis

The significance level of the difference between paired samples was examined using the student's t-test as implemented in R software. The mean separation was done to ascertain if the means for the top ranked individuals predicted by the three models were significantly different using least significant difference (LSD) at alpha=0.05.

4.3 Results

The MNDCGV was used to evaluate the performance of DeepGS, RR-BLUP and Ensemble (integrated DeepGS and RR-BLUP) for selecting individuals with high phenotypic values for body weight. With top-ranked alpha increasing from 10% to 70%, the results showed that the MNDCGV for DeepGS was approximately equal to 0.75~0.78, RR-BLUP 0.66~0.76 and Ensemble 0.76~0.79 as shown in Figure 3. The top alpha for this work did not reach 100% because the number of individuals used was 394 and when subjected to fivefold cross-validation, 79 individuals remained for performance evaluation.

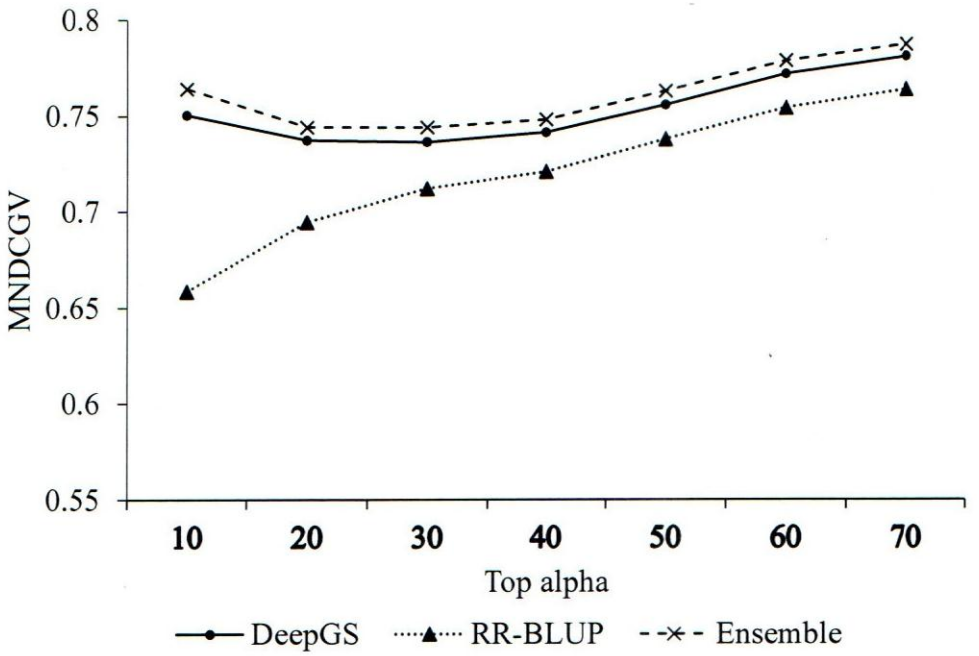


Figure 3: Mean normalized discounted cumulative gain values (MNDCGV) curves for DeepGS, RR-BLUP, and Ensemble model with top-ranked alpha increasing from 10% to 70%

The mean separation was done to using LSD at alpha=0.05 to ascertain if the means for the top ranked individuals predicted by the three models were significantly different. The results showed that MNDCGVs for DeepGS and Ensemble were significantly higher than those of RR-BLUP as shown in Table 3.

Table 3: Mean separation using least significant difference (LSD) *for GS models*

Model	Mean	N
Ensemble	0.76129[a]	7
DeepGS	0.75332[a]	7
RR-BLUP	0.72018[b]	7

Means with the same letter are not significantly different.

This study also aimed at investigating the performance improvement (absolute increment) for the three GS models. Integrated (Ensemble) model substantially improved the prediction performance over DeepGS and RR-BLUP. The absolute MNDCGV improvement (absolute increment) at the top-ranked level of alpha = 10% of Ensemble (0.76) over RR-BLUP (0.66) was 0.11 and for Ensemble (0.76) over DeepGS (0.75) was 0.01 shown in Table 4 and with P-values corresponding to 0.01 and 0.000107 respectively. Figure 4 shows the trend of the absolute increment of in MNDCGV of DeepGS and the ensemble GS models over RR-BLUP. When compared with RR-BLUP the median results showed that Ensemble model improved the MNDCGVVs by 0.03 and DeepGS improved by 0.02 corresponding to 3% and 2% respectively.

Table 4**: Prediction performance based on MNDCGV for Ensemble, DeepGS and RR-BLUP with top-ranked alpha increasing from 10% to 70%**

		MNV		Ensemble *vs* DeepGS	DeepGS *vs* RR-BLUP	Ensemble *vs* RR-BLUP
Top α	DeepGS	RR-BLUP	Ensemble	MNV improvement	MNV improvement	MNV improvement
10	0.75	0.66	0.76	0.01	0.09	0.11
20	0.74	0.69	0.74	0.01	0.04	0.05
30	0.74	0.71	0.74	0.01	0.02	0.03
40	0.74	0.72	0.75	0.01	0.02	0.03
50	0.76	0.73	0.76	0.01	0.02	0.03
60	0.77	0.75	0.78	0.01	0.02	0.02
70	0.78	0.76	0.79	0.01	0.02	0.02

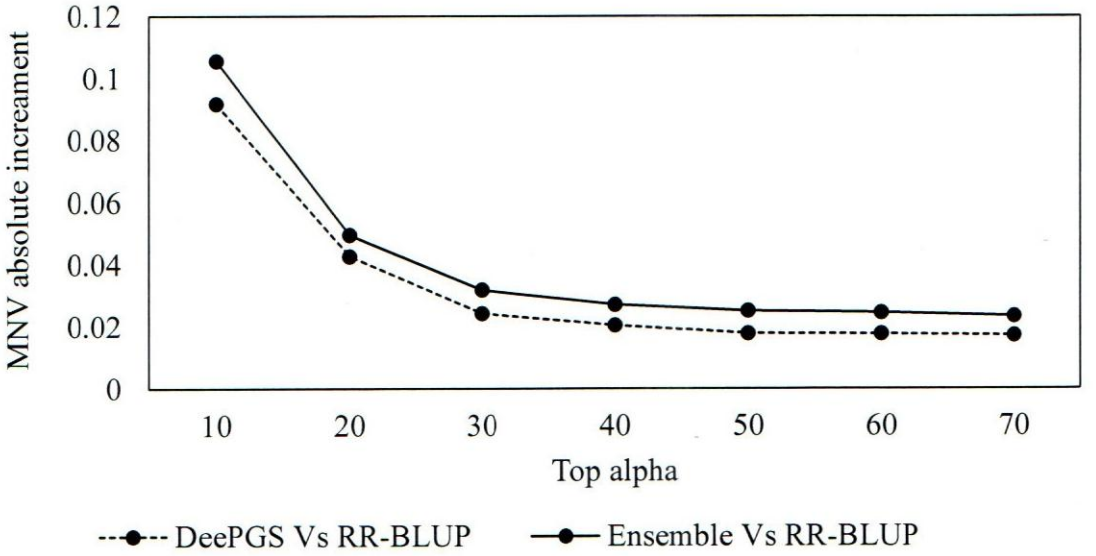


Figure 4: The absolute increase in MNDCGV of DeepGS and the ensemble GS models over RR-BLUP evaluated using fivefold cross-validation with five replicates

4.4 Discussion

The ranking problem has become an important topic in machine learning, partly due to its widespread applications in many decision-making processes because the measures of rank quality are usually based on sorting, which is not directly optimizable. To counter this, mean normalized discounted cumulative gain value (MNDCGV), a standard quality measure in information retrieval with capabilities of ranking individuals according to breeding values (performance) has been proposed (Blondel *et al.*, 2015). It has the ability of rewarding models which assign a high rank to individuals with high breeding values. It is also important to select a small number of individuals because selected individuals contribute their genetic materials to the next generation. If too many candidates are selected, selection intensity becomes low and it is not possible to obtain a good improvement of the target trait in the next generation (Blondel *et al.*, 2015). The focus of this study, therefore, was, to evaluate the prediction performance of GS models using the MNDCGV.

The MNDCGV results showed that the DeepGS model was approximately equal to 0.75~0.78, RR-BLUP 0.65859~0.763621 Ensemble 0.764274~0.7871 with top-ranked alpha increasing from 10% to 70%. The Ensemble model and DeepGS model outperformed the conventional RR-BLUP model by a significant margin ($P<0.05$). The mean separation for MNDCGV results for the models showed that there was no significant difference between Ensemble and DeepGS model but RR-BLUP was significant different from the other two models. Therefore, Ensemble and DeepGS can be used as a supplement to RR-BLUP in predicting the performance of individuals with high phenotypes to be used for selection

purposes in indigenous chicken breeding programs. A similar trend of results was reported in the phenotypic prediction of a wheat study done by Ma *et al.* (2018). The reason behind this difference in performance ranking is that both Ensemble, DeepGS and RR-BLUP models capture different aspects of the relationship between phenotypes and genotypes, this is attributed to the fact that the models used different algorithms to build regression-based models (Ma *et al.*, 2018). Also, the integrated model (Ensemble) and DeepGS had an advantage over conventional models (RR-BLUP) because of their ability to handle categorical variables and missing values without prior imputation and estimate variable importance and interactions (Blondel *et al.*, 2015).

The MNDCGV improvement at the top-ranked level of alpha = 10% results showed that integrated model (Ensemble) performed very well in ranking individuals with better performance compared to DeepGS and RR-BLUP, corresponding to 0.01 for DeepGS and 0.11 for RR-BLUP. This confirms a trend observed by Ma *et al.* (2018) and Blondel *et al.* (2015) in their study on wheat phenotypic prediction and a ranking approach to genomic selection respectively. This is attributed to the fact that ensemble model uses a particle swarm optimization (PSO) algorithm, which has the capability of parallel searching on very large spaces of candidate solutions, without making assumptions about the problem being optimized (Kennedy and Eberhart, 1995). This is also supported by Ma *et al.* (2018) argument on the combination of predictions of DeepGS and RR-BLUP to contribute to better performance.

Performance improvement was also evaluated by comparing the Ensemble and DeepGS model with respect to RR-BLUP. The median results showed that Ensemble model improved the MNDCGV over RR-BLUP by 0.03 and DeepGS improved by 0.02. This further affirms the fact that Ensemble model is superior and robust in performance improvement compared to DeepGS and RR-BLUP. Therefore, Ensemble model can be given a top priority as the GS model and as an alternative to conventional (ridged regression) GS models. The performance of ridge regression model with respect to MNDCGV was not good for this study, these egos what Ma *et al.* (2018) reported but contradicts what Heslot *et al.* (2012) suggested on ridged regression model being the best models for ranking individuals.

4.5 Conclusion

The performance evaluation results for this study showed that the Ensemble and DeepGS model performed better than RR-BLUP. Therefore, for phenotypic ranking, this work advocates the application of Ensemble and DeepGS model to be used as a supplement

to RR-BLUP in ranking of top performing individuals with high phenotypic values to be used for selection purposes. Also, for performance improvement this study suggests use of ensemble model because it has shown to be robust, powerful and effective in comparison to other GS models. Application of MNDCGVs in ranking top performing animals opens up a new avenue for selecting superior animals to be used for genetic improvement programs.

CHAPTER FIVE

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

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. This is because of its ability to accommodate noisy, non-linearity in data set and they make decisions based on prior knowledge. 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. 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. These results imply that 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*

5.1 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). Even though 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 their flexibility in classification, pattern 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 pressure thus 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 practices is 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.

5.2 Materials and methods

5.2.1 Data source

The phenotypic and genotypic data used for this study 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.

5.2.2 Imputation of missing genotypes and coding

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.

5.2.3 Statistical analysis

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 was not used during the training phase. Data preprocessing and analysis was performed using the R software.

In order 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 comparison of the different neural network types using identical datasets to determine the most suitable architecture for predicting a phenotypical trait based on genotypes.

5.2.4 Feedforward artificial neural network

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 projects 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 5 and 6 respectively.

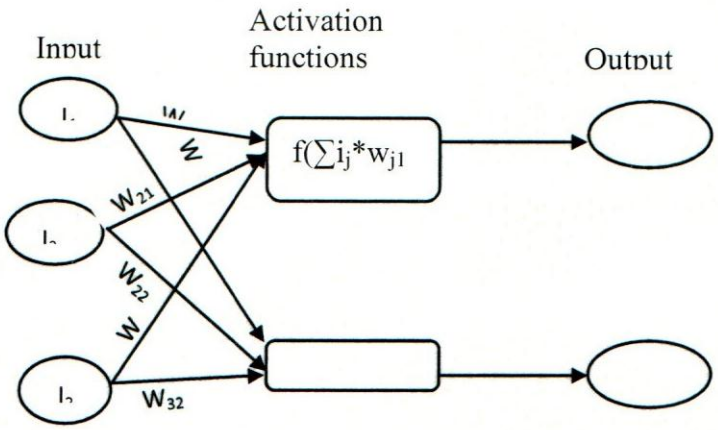


Figure 5: Single feed-forward network

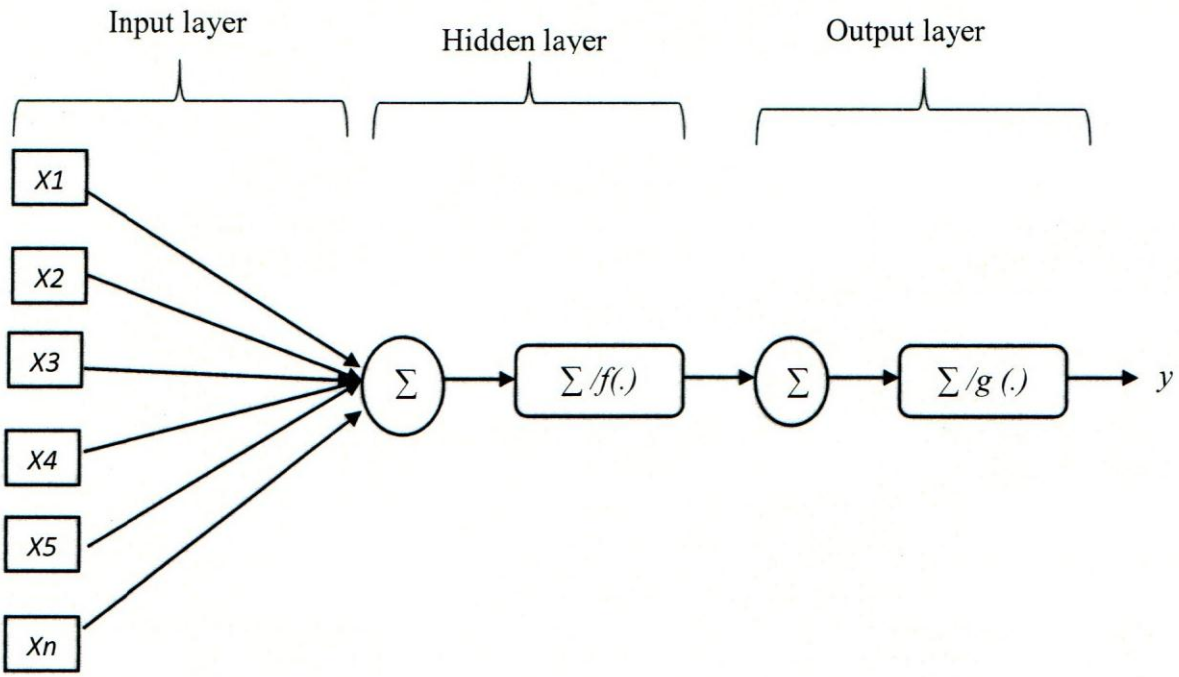


Figure 6: Fully connected feed-forward network with one hidden layer

5.2.5 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 function: 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 = the total number of hidden neurons: w_{2t} = vector of weights: q_i^t = the 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 samples

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:

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

5.3 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 shown in Table 5. 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 the sense that the network performs a multiple linear regression, in which the weights of the hidden layer can be interpreted as regression coefficients (Ehret *et al.*, 2015).

Table 5: **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 7.

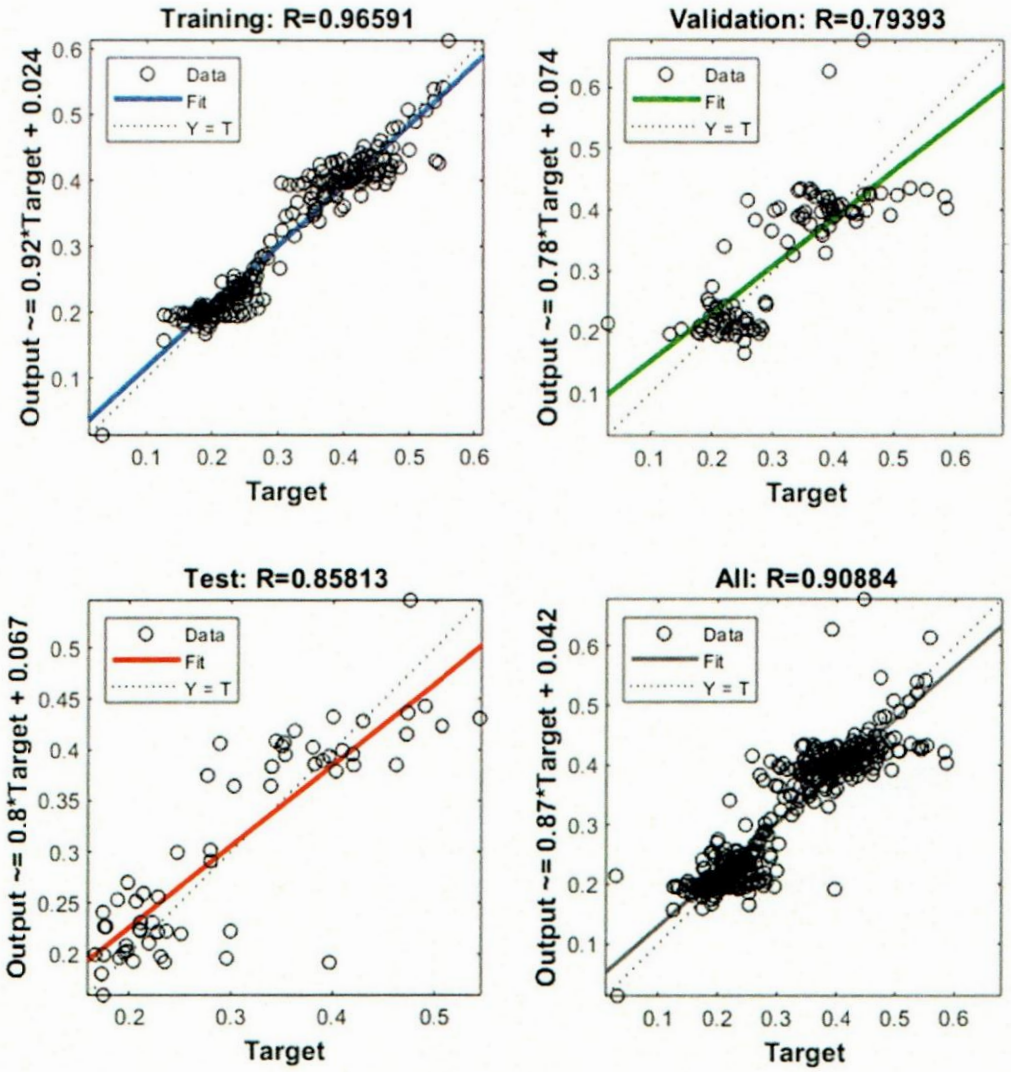


Figure 7: Pearson's correlation coefficients for the training, validation and testing data set related to 10 neuron architecture

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 8.

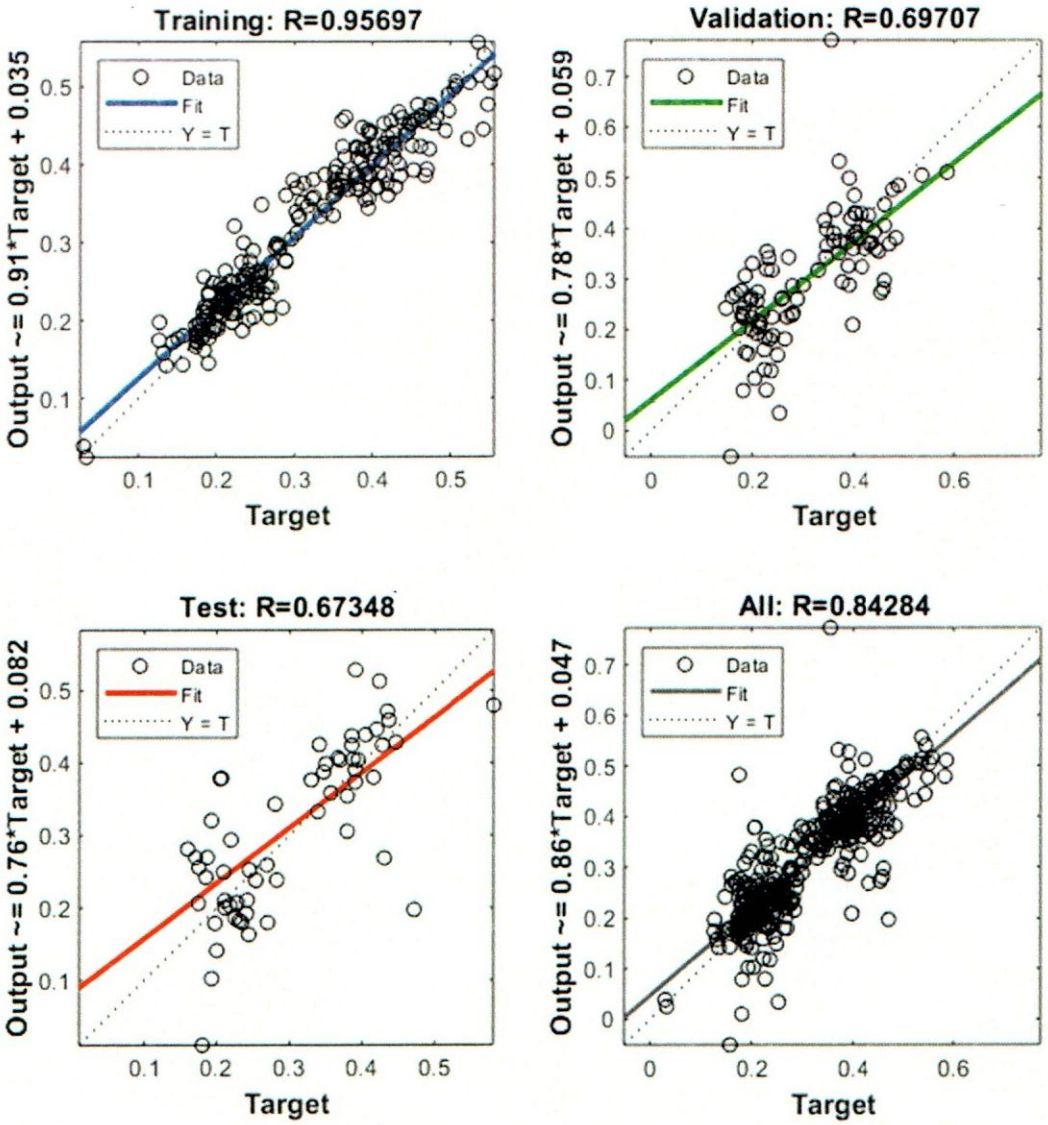


Figure 8: Correlation coefficients for the training, validation and testing data set related to 16 neuron architecture

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 9.

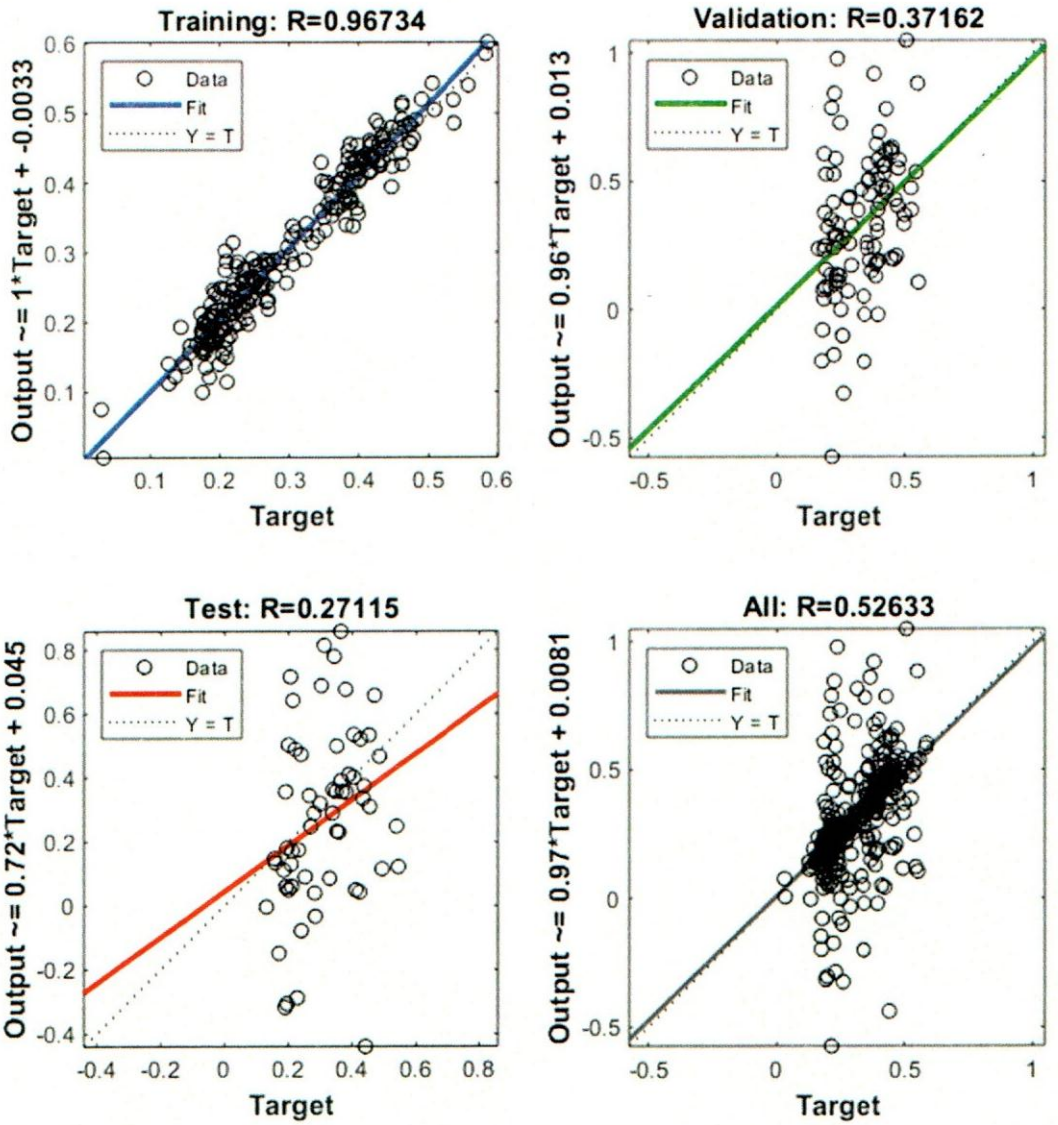


Figure 9: Correlation coefficients for the training, validation and testing data set related to 32 neuron architecture

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 was 0.77 for R and 5.72E-3 for the MAE as shown in Figure 10. Computational time was long during the training process this is as a result of all input information passing through a single neuron.

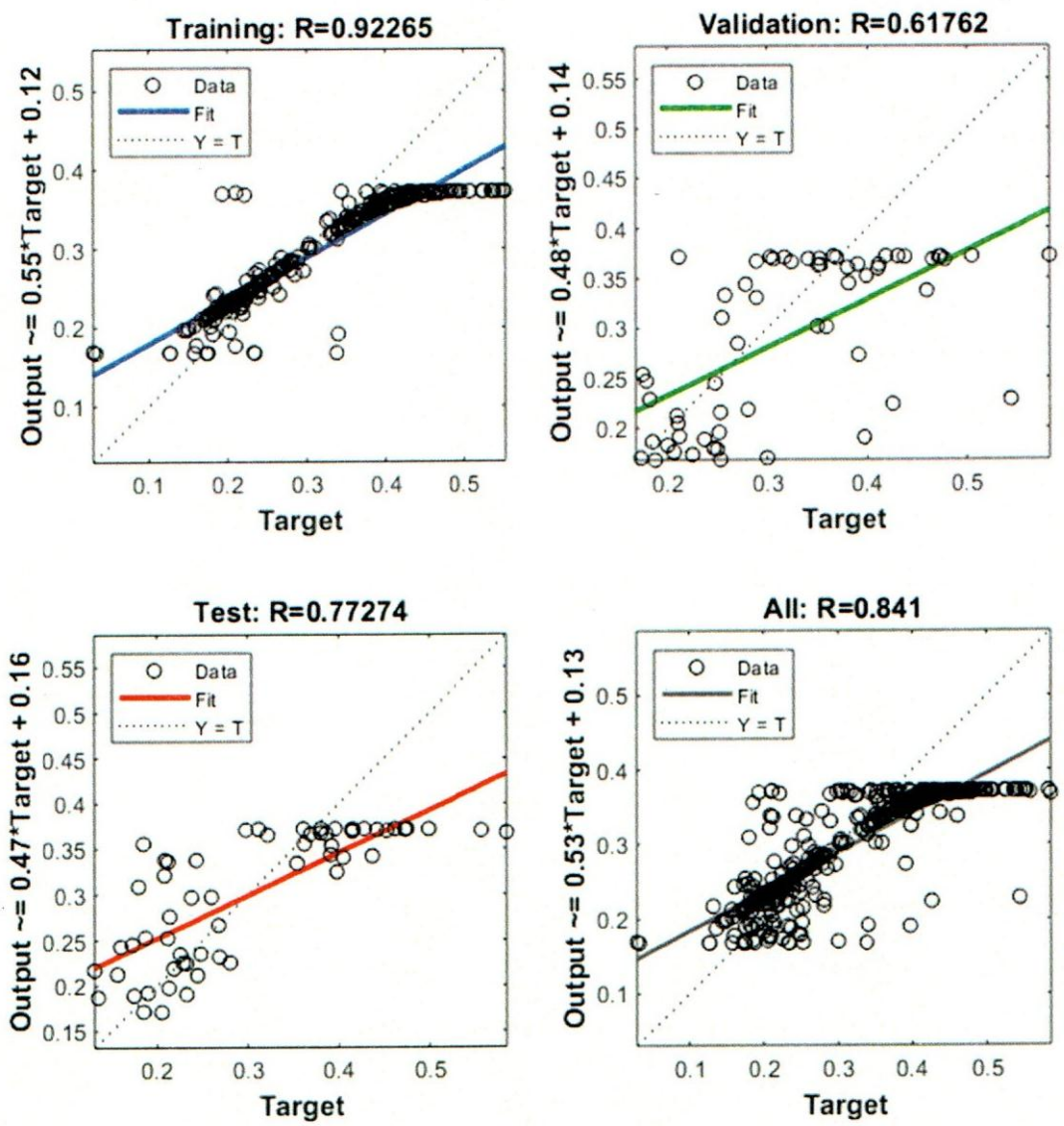


Figure 10: Representation of linear model correlation coefficients for the training, validation and testing data set related to one neuron architecture model

5.4 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 models, 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 underfitting 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. This 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 and the 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 the data 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 a correlation 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.

5.5 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 also reveals the fact that increase of dimensionality of network architecture leads to model underfitting or overfitting as proven by the results obtained from the complex non-linear models (with 16 and 32 neurons).

CHAPTER SIX

GENERAL DISCUSSION, COCLUSIONS AND RECOMMENDATIONS

6.1 Aim of the study

Genomic selection has received interest from animal breeders in the recent years and is rapidly becoming the method of choice of selection in dairy cattle breeding programs (Hayes *et al.*, 2009). With the availability of high throughput genotyping and sequencing technology, genomic prediction has become the standard method in breeding schemes. The genomic prediction is a process of making predictions based on observed genomic data to predict the unobserved complex phenotypic traits of individual. Genomic predictions have been based on conventional genomic predictive models, however, the conventional statistical models do not consider non-linearity among explanatory variables and association between predictors and target variables which may involve non-additive effects, leading to biased results and low EBVs. Therefore, machine learning approaches in animal breeding are emerging as a powerful tool and alternative model for making accurate predictions from complex quantitative traits data. However, its ability to predict phenotypic values from molecular data is not well studied (Pérez-Enciso *et al.*, 2019).

The overall goal of this study was to contribute to the genetic improvement of IC through selection by adopting machine learning techniques as predictive models. The specific objectives were: i) to compare the predictive ability of conventional neural network, conventional statistical models and deep convolutional neural network for body weight of indigenous chicken at 12 weeks based on genome-wide markers of genotypes, ii) to evaluate the prediction performance of conventional neural network, conventional genomic models and deep convolutional neural network, iii) to evaluate the extent of non-linearity among explanatory variables within and between genotypes and phenotypes using artificial neural network. This thesis addressed three major research questions which included; i) what are the predictive abilities of conventional neural network, conventional statistical models and deep convolutional neural network for body weight of indigenous chicken at 12 weeks based on genome-wide markers of genotypes? ii) what are the prediction performances of conventional neural network, conventional genomic models and deep convolutional neural network? and finally iii) what is the extent of non-linearity among explanatory variable within and between genotypes and phenotypes using artificial neural network?

6.2 Study Methodology

This study used data obtained from a wide-genome study done by Yuan *et al.* (2018) in four Chinese indigenous chicken breeds. The study, genotyped 394 birds and live body weight (BW) was measured at hatch and every week until 12 weeks of age. For comparison between machine learning and conventional linear models, the deep convolutional neural network (DeepGS), conventional neural network (Artificial neural network), conventional statistical predictive model (ridge regression best linear unbiased, RR-BLUP) model and combination of DeepGS and RR-BLUP (Ensemble model) were adopted in predicting body weight (BW) of indigenous chicken based on genome-wide markers. The DeepGS model was constructed using the deep learning technique as described in a project done by Ma *et al.* (2018) available at GitHub (https://github.com/cma2015/DeepGS). The conventional neural network model adopted was a feedforward artificial neural network (ANN) built using a matrix laboratory (MATLAB) software function "feedforward net", with an 18-32-1 architecture.

For the effective selection of the most favourable individuals to be used for genetic improvement programs, it is appropriate to correctly rank individuals from most favourable to least favourable based on phenotypic values rather than to accurately predicting breeding values (Blondel *et al.*, 2015). Therefore, second the objective aimed at assessing the performance predictions of different models based on mean normalized discounted cumulative gain value (MNDCGV) in ranking individuals, as described by Blondel *et al.* (2015). To assess the extent of non-linearity among explanatory variables within and between genotypes and phenotypes, 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 network were done with MATLAB. Mean absolute error (MAE) and Pearson's correlation coefficients (R) were used to measure the ANN performance.

6.3 Comparison between machine learning and conventional linear models

Genomic selection has been based on conventional linear models, however these models are faced with statistical challenges related to dimensionality of data set and do not consider non-linearity in data and association between predictors and target variables which may involve non-additive effects resulting to biased results and low predictive abilities (Ma *et al.*, 2018). Machine learning algorithms have been reported as an attractive, non-linear alternative to conventional models according to literature (Singh *et al.*, 2016; Mohaiminul, 2017).

Chapter three of this study showed that, DeepGS did not significantly ($p>0.05$) outperform RRBLUP, Ensemble nor ANN model. This is supported by recent literature which indicates that some machine learning models can be competitive with linear methods, but not in all cases (Pérez-Enciso *et al.*, 2019). This is also egod by Ma *et al.* (2018) which reported that DeepGS can be used as a complement to the RR-BLUP model. The PCC results from this study showed that Ensemble model resulted in better performance (0.892) compared to DeepGS and RR-BLUP with PCC corresponding to 0.891 and 0.889 respectively. This is attributed to the fact that Ensemble model uses particle swarm optimization (PSO) algorithm, with capability of parallel searching without making assumptions about the problem being optimized. The ANN model did not perform very well compared to the other three models (DeepGS, RR-BLUP and Ensemble) because the performance of ANN is affected by adopted architecture. The mean separation to ascertain the significance of the tested models showed that the models were statistically not different. Therefore, the DeepGS, Ensemble, RR-BLUP and ANN model can be used in complementarily for phenotypic predictions.

On the prediction of top-ranked 200 individuals with high BW, the performance showed that all the tested models performed poorly on BW prediction as shown by PCC result corresponding to 0.315 for DeepGS, 0.466 for RR-BLUP, 0.342 for Ensemble and 0.518 for ANN model. This shows that body weight predicted for individuals with high phenotypes correlate poorly with actual body weight. Since traditional breeding programs focuses on individuals with high phenotypic performance for selection, same attention should be put when selecting individuals with high breeding values to be used in genomic selection because they not only form the breeding stock for future but there predicted correlation to the true value is worrying. Therefore, this forms a basis of selection process for individuals to be adopted in a breeding program as replacement stock.

6.4 Evaluation of prediction performance of genomic models

Machine learning (ML) application in genome-enabled predictions as a way of improving accuracy has been on a rise in recent years (Bellot *et al.*, 2018). Accuracy of prediction acts as a measure of genomic models and it is affected by density of markers, size of genome and genetic architecture. It measures how reliable the future phenotypes of target trait of individual can be predicted. Predictive accuracy of most models is typically assessed using the Pearson correlation coefficient (PCC) between observed trait values and the predicted trait values despite PCC correlating poorly with ranking accuracy (Blondel *et al.*,

2015). Ranking accuracy has become a crucial topic in machine learning models as a result of the measure of rank quality which is usually based on sorting, that isn't directly optimizable. This led to development of the MNDCGV principal which has been reported to possess capabilities of ranking individuals according to breeding values (Blondel *et al.*, 2015).

The MNDCGV from this study showed that the accuracy of GEBVs estimated using DeepGS model ranged between 0.75 and 0.78, 0.66 to 0.76 in RR-BLUP and 0.76 to 0.79 in Ensemble. When subjected to mean separation to ascertain the significance of the models, the Ensemble and DeepGS model outperformed the conventional RR-BLUP model by a significant margin ($P<0.05$). Therefore, Ensemble and DeepGS can be used as a supplement to RR-BLUP in predicting the performance of individuals with high breeding values. These results are supported by Ma *et al.* (2018) study which concluded that both Ensemble and DeepGS can be used as an alternative to conventional genomic models in ranking individuals.

6.5 Extend of non-linearity among explanatory variables within and between genotypes and phenotypes

Artificial neural network over the past time has been applied in animal breeding and have been found to be prone to overfitting, although sometimes competitive with penalized linear methods (González-Recio *et al.*, 2014). In this study (chapter 5), we have demonstrated how different neural network architecture for predicting a phenotypical trait based on genotypes perform when subjected to the non-linear dataset. Since ANNs gathers knowledge by detecting the patterns and relationships in data and have ability to learn through experience and performs well for optimization and prediction in complex non-linear systems (Pérez-Enciso *et al.*, 2019). According to this study, different ANN models performed differently in terms of prediction output, for instance the neural network with 10 neurons in the hidden layer performed better with R-value of 0.86 in comparison to 0.67 and 0.27 in network with 16 and 32 neurons in the hidden layer respectively. The non-linear activation functions in the hidden layer enable an ANN to have universal approximation ability, giving it great potential and flexibility in terms of capturing complex patterns (Pérez-Rodrígue *et al.*, 2012). The decreases in R-value with an increase in dimensionality of data is attributed to the fact that increasing the architecture dimensions of the model makes it suffer from over-fitting and also makes model more computationally demanding. The linear benchmark model recorded an R-value of 0.77, meaning that linear models can as well outperform some of non-linear models with large number of neurons.

6.6 Conclusions

i. The DeepGS, Ensemble, RR-BLUP and ANN models can be used interchangeably in making phenotypic predictions

ii. The Ensemble and DeepGS model were most appropriate in phenotypic ranking of individuals because they have shown to be robust, powerful and effective GS models.

iii. Neural architecture with a single hidden layer with 10 neurons is sufficient to efficiently account non-linearity in the data sets.

6.7 Recommendations

i. For researches its, important to consider the nature of the neural network, the choice of input values and its distribution in the input data set before adopting machine learning models for better phenotypic predictions.

ii. For efficient handling of non-linear data, it's important for researches to consider neural dimensionality since the performance of a genomic model is affected by network overfitting or underfitting.

iii. Further research need to be done the application of machine learning models in prediction of other polygenic traits, since this work was restricted to a single trait (body weight)

REFERENCES

Abdollahi-Arpanahi, R., Gianola, D., & Penagaricano, F. (2020). Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genetics Selection Evolution*, 52(1), 1-15.

Acevedo, M. A., Corrada-Bravo, C. J., Corrada-Bravo, H., Villanueva-Rivera, L. J., & Aide, T. M. (2009). Automated classification of bird and amphibian calls using machine learning: A comparison of methods. *Ecological Informatics*, *4*(4), 206-214.

Akanno, E. C., Schenkel, F. S., Sargolzaei, M., Friendship, R. M., & Robinson, J. A. B. (2014). Persistency of accuracy of genomic breeding values for different simulated pig breeding programs in developing countries. *Journal of Animal Breeding and Genetics*, *131*(5), 367-378.

Awuni, J. A. (2002). Strategies for the improvement of rural chicken production in Ghana. In A. I. Khalafalla, A. Ali & W. H. Ali (Eds.), *Characteristics and Parameters of Family Poultry Production in Africa* (pp 33-37). Vienna, Austria: IAEA.

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

Bellot, P., de los Campos, G., & Perez-Enciso, M. (2018). Can Deep Learning Improve Genomic Prediction of Complex Human Traits? *Genetics*, 210(3), 809-819.

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

Bett, H. K. (2012). *Economic analysis of indigenous chicken genetic resources in Kenya*. PhD Thesis, Humboldt University of Berlin, Germany.

Blondel, M., Onogi, A., Iwata, H., & Ueda, N. (2015). A ranking approach to genomic selection. *PloS one*, *10*(6), 1-23.

Bolormaa, S., Pryce, J. E., Zhang, Y., Reverter, A., Barendse, W., Hayes, B. J., & Goddard, M. E. (2015). Non-additive genetic variation in growth, carcass and fertility traits of beef cattle. *Genetics Selection Evolution*, *47*(1), 26-38.

Calus, M. P. L., & Veerkamp, R. F. (2007). Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *Journal of Animal Breeding and Genetics*, *124*(6), 362-368.

Calus, M.P.L., Berry, D.P., Banos, G., de Haas, Y., Veerkamp, R.F., (2013). Genomic selection: the option for new robustness traits. *Advances in Animal Biosciences*. 4(3), 618–625

Case, L. A., Wood, B. J., & Miller, S. P. (2012). The genetic parameters of feed efficiency and its component traits in the turkey (*Meleagris gallopavo*). *Genetics Selection Evolution*, *44*(1), 2-4.

Crossa, J., de Los Campos, G., Perez, P., Gianola, D., Burgueno, J., Araus, J. L., ... & Arief, V. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*, 186, 713-724.

Crossa, J., Diego, J., Jorge, F., Paulino, P-R., Juan, B., Carolina, S-P., Prashant, V., Carolina, S., Cesar, P., Deniz, A., Clay, S., Matthew, R., Maria, T., Thomas, P., Carlos, G., Roberto, J. P., Peter, W., Sukhwinder S. (2016). Genomic prediction of gene bank wheat landraces. *G3 (Bethesda)*, 6(7), 1819-1834.

Ehret, A. Hochstuhl D, Gianola D, Thaller G. (2015). Application of neural networks with back-propagation to genome-enabled prediction of complex traits in Holstein-Friesian and German Fleckvieh cattle. *Genetics, Selection, Evolution.* 47(1), 22-30.

FAO (2010). Chicken Genetic Resources Used in Smallholder Production Systems and Opportunities for their Development. In P. Sorensen (Ed.), *Smallholder Poultry Production*. Rome, Italy., 5.

Favaro, L., Briefer, E. F., & McElligot, A. G. (2014). Artifcial Neural Network approach for revealing individuality, group membership and age information in goat kid contact calls. *Acta Acustica United with Acustica*, 100(4), 782-789.

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

Fulton, J. E. (2012). Genomic selection for poultry breeding. *Animal Frontier*, 2(1), 30-36. *Genetic and Molecular Research*, 14(3), 10888-10896.

Gianola, D. and Schon, C.C. (2016). Cross-validation without doing cross-validation in genome-enabled prediction. *G3 (Bethesda),* 6(10), 3107-3128.

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

Gianola, D., R. L. Fernando, and A. Stella, (2006) Genomic-assisted prediction of genetic value with semi-parametric procedures. *Genetics* 173, 1761-1776.

Goddard, M. E., & Hayes, B. J. (2007). Genomic selection. *Journal of Animal Breeding and Genetics*, *124*(6), 323-330.

Goddard, M. E., & Hayes, B. J. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. *Journal of Animal Breeding and Genetics*, *128*(6), 409-421.

Gonzalez-Camacho, J. M., de Los Campos, G., Perez, P., Gianola, D., Cairns, J. E., Mahuku, G., & Crossa, J. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. *Theoretical and Applied Genetics*, 125(4), 759-771.

Gonzalez-Camacho, J.M., Ornella, L., Parez-Rodriguez, P., Gianola, D., Dreisigacker, S., Crossa, J. (2018). Applications of Machine Learning Methods to Genomic Selection in Breeding Wheat for Rust Resistance. *Plant Genome*, 11(2), 170104-170119.

González-Recio, O., Rosa, G. J., & Gianola, D. (2014). Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livestock Science*, 166, 217-231.

Gorgulu, O. (2012). Prediction of 305-day milk yield in Brown Swiss cattle using artificial neural networks. *South African Journal of Animal Science*, *42*(3), 280-287.

Groenen, M. A., Megens, H. J., Zare, Y., Warren, W. C., Hillier, L. W., Crooijmans, R. P., & Cheng, H. H. (2011). The development and characterization of a 60K SNP chip for chicken. *BMC Genomics*, *12*(1), 274-283.

Hanotte, O., Dessie, T., & Kemp, S. (2010). Time to tap Africa's livestock genomes. *Science*, 328(5986), 1640-1641.

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

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*, *92*(2), 433-443.

Heidaritabar, M. (2016). *Genomic selection in egg-laying chickens*. Doctoral dissertation, Wageningen University, Netherlands.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics,* 31(2): 423-447.

Heslot, N., Yang, H. P., Sorrells, M. E., & Jannink, J. L. (2012). Genomic selection in plant breeding: a comparison of models. Crop science, 52(1), 146-160.

Hosseinia, P., Edrisi, M., Edriss, M. A., & Nilforooshan, M. A. (2007). Prediction of second parity milk yield and fat percentage of dairy cows based on first parity information using neural network system. *Journal of Applied Sciences, 7(21)*, 3274-3279.

Ibanez-Escriche, N., & Gonzalez-Recio, O. (2011). Promises, pitfalls and challenges of genomic selection in breeding programs. *Spanish Journal of Agricultural Research, 9*(2), 404-413.

International Chicken Genome Sequencing Consortium (2004). A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature,* 432(7018), 717–722.

Kennedy, J., & Eberhart, R. (1995). Particle Swarm Optimization. In "*Proceedings of IEEE International Conference on Neural Networks*" (Vol. 4, No. 1, pp. 1942-1948).

Khobondo, J. O., Mwakubambanya, R., Wasike, C. B., & Kahi, A. K. (2017). Genetic and non-genetic sources of variation in natural antibodies titre values among indigenous chicken. *American Journal of Research Communication,*5(7), 31-45.

King'ori, A.M., (2004). *The protein and energy requirements of indigenous chickens (Gallus domesticus) of Kenya*. PhD. Thesis, Egerton University, Kenya.

Kingori, A.M., Wachora, A.M., Tuitoek, J.K. (2010). Indigenous chicken production in Kenya: A Review. *International Journal of Poultry Science*, 9(4), 309-316.

Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

Lee, S. H., Clark, S., & van der Werf, J. H. (2017). Estimation of genomic prediction accuracy from reference populations with varying degrees of relationship. *PloS One*, *12*(12), e0189775, https:// doi.org/10.1371/journal.pone.0189775

Lopes, M. S. (2016). *Genomic selection for improved crossbred performance*. Doctoral dissertation, Wageningen University, Netherlands.

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., & Ma, C. (2018). A deep convolutional neural network approach for predicting phenotypes from genotypes. *Plant*, 248(5), 1307-1318.

Macrossan, P.E., Abbass, H.A., Mengersen, K., Towsey, M., Finn, G., (1999). Bayesian neural network learning for prediction in the Australian dairy industry. In: Hand, D.J., Joost, N.K., Berthold, M.R. (Eds.). *Advances in Intelligent Data Analysis, Proceedings of Third International Symposium*, (pp.395-406). Amsterdam, Netherlands, Springer.

Magothe, T. M., Muhuyi, W. B., & Kahi, A. K. (2010). Influence of major genes for crested-head, frizzle-feather and naked-neck on body weights and growth patterns of indigenous chickens reared intensively in Kenya. *Tropical Animal Health and Production*, *42*(2), 173-183.

Magothe, T. M., Okeno, T. O., Muhuyi, W. B., & Kahi, A. K. (2012). Indigenous chicken production in Kenya: II. Prospects for research and development. *World's Poultry Science Journal*, *68*(1), 133-144.

Magothe, T., & Kahi, A. (2011). Indigenous chicken improvement in Kenya: Past efforts and future prospects. *Driving Livestock Enterprenuership Towards Attainment of Food Sufficiency and Kenya Vision, 2030*, 15-18.

McDowell, R. M. (2016). *Genomic selection with deep neural networks*. MSc Thesis, Iowa State University, Ames, Iowa.

Md. Mohaiminul Islam (2017). *Deep Learning Models for Predicting Phenotypic Traits and Diseases from Omics Data*. MSc Thesis, The University of Manitoba, Canada.

Menge, E. O., Kosgey, I. S., & Kahi, A. K. (2005). Bio-economic model to support breeding of indigenous chicken in different production systems. *International Journal of Poultry Science*, *4*(11), 827-839.

Meuwissen, T. and M. Goddard. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics,* 185(2):623-631.

Meuwissen, T.H.E., Hayes, B.J. and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*; 157(4):1819-1829.

Ministry of Agriculture, Livestock and Fisheries (MAL&F), (2017). Capacity building strategy for agriculture sector.

Mrode, R. A. (2014). Linear models for the prediction of animal breeding values. Cabi.

Mukhopadhyay, C. S., & Kumar, D (2013). Genomic Selection: Its Prospects and Challenges.

Mwacharo, J. M., Bjornstad, G., Han, J. L., & Hanotte, O. (2013). The history of African village chickens: an archaeological and molecular perspective. *African Archaeological Review, 30*(1), 97-114.

Mwalusanya, N. A., Katule, A. M., Mutayoba, S. K., Mtambo, M. M. A., Olsen, J. E., & Minga, U. M. (2002). Productivity of local chickens under village management conditions. *Tropical Animal Health and Production*, 34(5), 405-416.

Ngeno, K. (2011). *Genetic analysis of growth patterns in different ecotypes of indigenous chicken populations in Kenya*. MSc Thesis, Egerton University, Kenya.

Ngeno, K. (2015). *Breeding program for indigenous chicken in Kenya.* PhD Thesis, Wageningen University, Netherlands.

Njubi, D.M., Wakhungu, J.W., and Badamana, M.S., (2010). Use of test-day records to predict first lactation 305-day milk yield using artificial neural network in Kenyan Holstein–Friesian dairy cows. *Tropical Animal Health and Production.* 42, 639-644.

Okeno, T. O., Kahi, A. K., & Peters, K. J. (2011). Breed selection practices and traits of economic importance for indigenous chicken in Kenya. *Growth*, *2*(3.01), 1-44.

Okeno, T. O., Kahi, A. K., & Peters, K. J. (2012). Characterization of indigenous chicken production systems in Kenya. *Tropical Animal Health and Production*, *44*(3), 601-608.

Okut, H., Wu, X.L., Rosa, G.J., Bauck, S., Woodward B.W., Schnabel, R.D., Taylor, J.F., and Gianola, D., (2013). Predicting expected progeny difference for marbling score in Angus cattle using artificial neural networks and Bayesian regression models. *Genetics Selection Evolution*, 45:34

Omiti, J. and S.O. Okuthe (2009), '*An Overview of the Poultry Sector and Status of Highly Pathogenic Avian Influenza (HPAI) in Kenya: A Background Paper*', Africa/Indonesia Team Working Paper 4, Pro-Poor HPAI Risk Reduction Project.

Peixoto, L. A., Bhering, L. L., & Cruz, C. D. (2015). Artificial neural networks reveal efficiency in genetic value prediction. *Genetics and Molecular Research*, *14*(2), 6796-807.

Pérez-Enciso, M., & Zingaretti, L. M. (2019). A Guide for Using Deep Learning for Complex Trait Genomic Prediction. *Genes*, 10(7), 553.

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

Rahman, M., Sorensen, P., Jensen, H. A., & Dolberg, F. (1997). Exotic hens under semi scavenging conditions in Bangladesh. *Livestock Research for Rural Development*, *9*(3), 1-11.

Resende, M. F. R., Munoz, P., Acosta, J. J., Peter, G. F., Davis, J. M., Grattapaglia, D., & Kirst, M. (2012). Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. *New Phytologist*, 193(3); 617-624.

Roorkiwal, M., Rathore, A., Das, R. R., Singh, M. K., Jain, A., Srinivasan, S., & Hickey, J. M. (2016). Genome-enabled prediction models for yield related traits in chickpea. *Frontiers in Plant Science*, 7, 1666.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533-536.

Singh, B. D., & Singh, A. K. (2015). Marker-assisted selection. In: Singh, B.; Singh, A. K. (Eds.). *Marker-Assisted Plant Breeding*: *Principles and Practices* (pp.259-293). New Delhi, India: Springer.

Singh, R., Lanchantin, J., Robins, G., (2016). Deep Chrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, 32(17), 639-648.

Team, R. C. (2017). R: A language and environment for statistical computing.

Van der Most, P.J., de Jong, B., Parmentier, H. K., & Varhulst, S. (2011). Trade-off between growth and immune function: a meta-analysis of selection experiment. *Functional Ecology*, 25(1), 74-80.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414-4423.

Wang, X., Xu, Y., Hu, Z., & Xu, C. (2018). Genomic selection methods for crop improvement: Current status and prospects. *The Crop Journal*, 6(4); 330-3340.

Wolc, A. (2014). Understanding genomic selection in poultry breeding. *World's Poultry Science Journal*, 70(2), 309-314.

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

Zhang, Z., Zhang, Q., & Ding, X. (2011). Advances in genomic selection in domestic animals. *Chinese Science Bulletin*, *56*(25); 2655-2663.

EGERTON UNIVERSITY
LIBRARY

APPENDICES

Appendix I. Data preparation and parameters setting syntaxes for DeepGS

```
library(DeepGS)
library(rrBLUP)
data(IC_data)
Markers <- IC_data$Markers
y <- IC_data$y
cvSampleList <- cvSampleIndex(length(y),10,1)
# cross validation set
cvIdx <- 1
trainIdx <- cvSampleList[[cvIdx]]$trainIdx
testIdx <- cvSampleList[[cvIdx]]$testIdx
trainMat <- Markers[trainIdx,]
trainPheno <- y[trainIdx]
validIdx <- sample(1:length(trainIdx),floor(length(trainIdx)*0.1))
validMat <- trainMat[validIdx,]
validPheno <- trainPheno[validIdx]
trainMat <- trainMat[-validIdx,]
trainPheno <- trainPheno[-validIdx]
conv_kernel <- c("1*18") ## convolution kernels (fileter shape)
conv_stride <- c("1*1")
conv_num_filter <- c(8)  ## number of filters
pool_act_type <- c("relu") ## active function for next pool
pool_type <- c("max") ## max pooling shape
pool_kernel <- c("1*4") ## pooling shape
pool_stride <- c("1*4") ## number of pool kernerls
fullayer_num_hidden <- c(32,1)
fullayer_act_type <- c("sigmoid")
drop_float <- c(0.2,0.1,0.05)
cnnFrame <- list(conv_kernel =conv_kernel,conv_num_filter = conv_num_filter,
          conv_stride = conv_stride,pool_act_type = pool_act_type,
          pool_type = pool_type,pool_kernel =pool_kernel,
          pool_stride = pool_stride,fullayer_num_hidden= fullayer_num_hidden,
          fullayer_act_type = fullayer_act_type,drop_float = drop_float)

markerImage = paste0("1*",ncol(trainMat))
```

Training DeepGS model

```
trainGSmodel <- train_deepGSModel(trainMat = trainMat,trainPheno = trainPheno,
          validMat = validMat, validPheno = validPheno, markerImage = markerImage,
          cnnFrame = cnnFrame, device_type = "cpu", gpuNum = 1, eval_metric = "mae",
          num_round = 6000, array_batch_size= 30, learning_rate = 0.01,
          momentum = 0.5, wd = 0.00001, randomseeds = 0, initializer_idx = 0.01,
          verbose = TRUE)
```

Prediction

```
predscores <- predict_GSModel(GSModel = trainGSmodel,testMat = Markers[testIdx,],
        markerImage = markerImage)
```

Performance evaluation

```
refer_value <- runif(100)
pred_value <- sin(refer_value) + cos(refer_value)
meanNDCG(realScores = refer_value,predScores = pred_value, topAlpha = 10)
```

Training RR-BLUP model

```
rrBLUP_obj <-mixed.solve(trainPheno, Z=trainMat, K=NULL, SE = FALSE,
return.Hinv=FALSE)
```

Make predictions based on the trained model

```
rrBLUP_pred <- testMat %*% rrBLUP_obj$u + as.numeric(rrBLUP_obj$beta )
```

prepare the prediction matrix

```
test_predMat <- cbind(t(DeepGS_pred), rrBLUP_pred)
train_predMat <- cbind(testPheno, t(DeepGS_pred), rrBLUP_pred)
colnames(train_predMat) <-c("real", "DeepGS", "RR-BLUP")
```

Calculating the weight of different training model by using their predict scores

```
test_datapath <- system.file("exdata", "test_ELBPSO.RData",
                    package = "DeepGS")
load(test_datapath)

weight <- ELBPSO(rep_times = 100,interation_times = 25,weight_dimension = 2,
          weight_min = 0,weight_max=1,rate_min = -0.01,rate_max = 0.01,
          paticle_number = 10, pred_matrix = train_predMat,IW = 1,
          AF1 = 2, AF2 = 2)

ensemble_pred <- (test_predMat %*% weight)/sum(weight)
predMat <- cbind(testPheno, t(DeepGS_pred), rrBLUP_pred, ensemble_pred)
colnames(predMat) <- c("real", "DeepGS", "RR-BLUP", "ensemble")
cor(predMat)
```

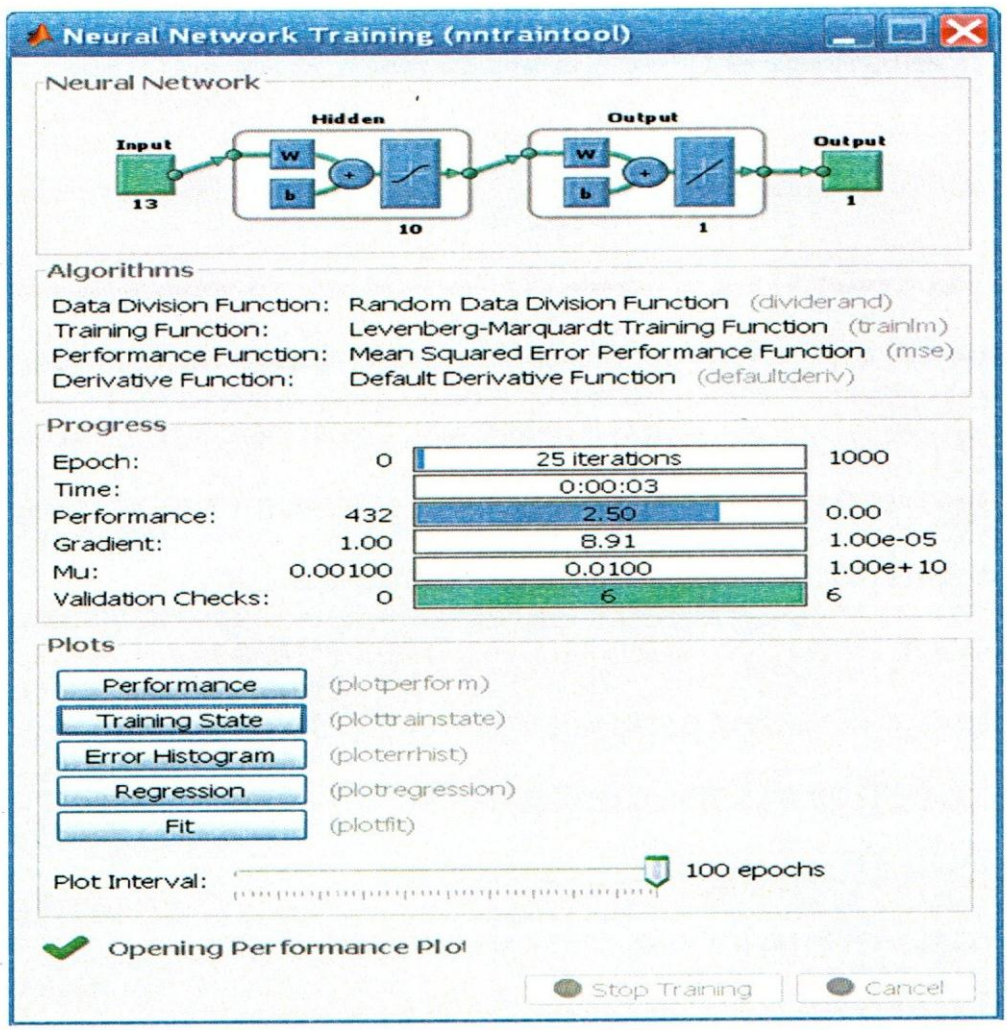
Script used in training neural network

```
% Solve an Input-Output Fitting problem with a Neural Network
% Script generated by Neural Fitting app
% Created Wed May 08 09:20:43 EAT 2019
% This script assumes these variables are defined:
%   buildingInputs - input data.
%   buildingTargets - target data.
x = buildingInputs;
t = buildingTargets;
% Choose a Training Function
% For a list of all training functions type: help nntrain
% 'trainlm' is usually fastest.
% 'trainbr' takes longer but may be better for challenging problems.
```

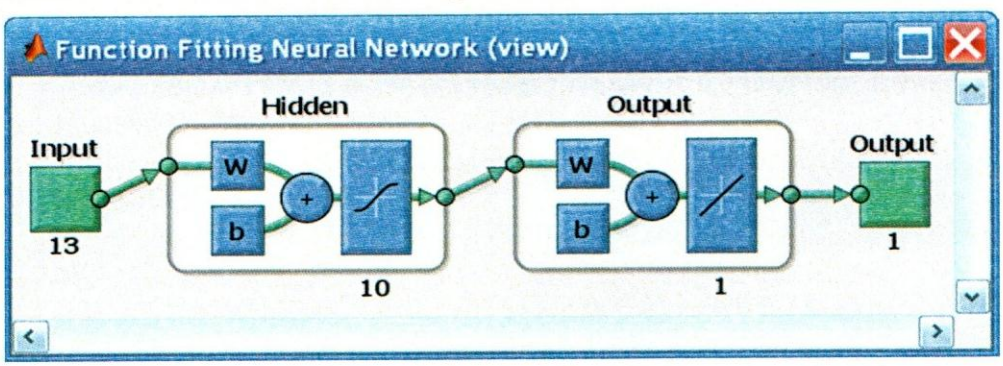
```
% 'trainscg' uses less memory. NFTOOL falls back to this in low memory situations.
trainFcn = 'trainlm'; % Levenberg-Marquardt
% Create a Fitting Network
hiddenLayerSize = 10;
net = fitnet(hiddenLayerSize,trainFcn);
% Choose Input and Output Pre/Post-Processing Functions
% For a list of all processing functions type: help nnprocess
net.input.processFcns = {'removeconstantrows','mapminmax'};
net.output.processFcns = {'removeconstantrows','mapminmax'};
% Setup Division of Data for Training, Validation, Testing
% For a list of all data division functions type: help nndivide
net.divideFcn = 'dividerand';  % Divide data randomly
net.divideMode = 'sample';  % Divide up every sample
net.divideParam.trainRatio = 70/100;
net.divideParam.valRatio = 15/100;
net.divideParam.testRatio = 15/100;
% Choose a Performance Function
% For a list of all performance functions type: help nnperformance
net.performFcn = 'mse';  % Mean squared error
% Choose Plot Functions
% For a list of all plot functions type: help nnplot
net.plotFcns = {'plotperform','plottrainstate','ploterrhist', ...
  'plotregression', 'plotfit'};
% Train the Network
[net,tr] = train(net,x,t);
% Test the Network
y = net(x);
e = gsubtract(t,y);
performance = perform(net,t,y);

% Recalculate Training, Validation and Test Performance
trainTargets = t .* tr.trainMask{1};
valTargets = t  .* tr.valMask{1};
testTargets = t  .* tr.testMask{1};
trainPerformance = perform(net,trainTargets,y);
valPerformance = perform(net,valTargets,y);
testPerformance = perform(net,testTargets,y);
% View the Network
view(net)
% Plots
% Uncomment these lines to enable various plots.
%figure, plotperform(tr)
%figure, plottrainstate(tr)
%figure, plotfit(net,x,t)
%figure, plotregression(t,y)
%figure, ploterrhist(e)
```

Training process of neural network



Artificial neural network training architecture



Appendix II. SAS output for objective one

The SAS System 11:03 Tuesday, June 1, 2019 1

The GLM Procedure

Class Level Information

Class Levels Values

method 5 ANN DeepGS Ensemble real rrBLUP

pheno 195 0.028417131 0.116502384 0.1268 0.157 0.163586284 0.1668 0.172912833 0.1731
0.178290218 0.1786 0.179133117 0.180116303 0.180773998 0.187539391 0.192207813
0.194035126 0.194476306 0.195105836 0.197764325 0.198261231 0.198634664 0.20121925
0.201313213 0.201569183 0.20260219 0.205476013 0.2062 0.206836551 0.207126678
0.207364408 0.207529129 0.208 0.209242033 0.209517494 0.2098 0.2103 0.210672342
0.210994823 0.211320909 0.213184116 0.213415238 0.214928553 0.215433812
0.216013819 0.216153936 0.217264831 0.218315572 0.218402813 0.21876251 0.2191
0.220428092 0.220446342 0.221206225 0.225403264 0.226380708 0.227546111
0.228139487 0.229131818 0.229475063 0.230692499 0.231030062 0.231465425
0.233157515 0.233191344 0.233563793 0.234621901 0.238628253 0.241086949 0.2412
0.241860007 0.242102366 0.243665142 0.246263196 0.2472 0.2577 0.259233854
0.274137522 0.2826 0.2889 0.292402613 0.298465115 0.2998 0.317 0.326088965
0.329712093 0.33 0.330354005 0.332113999 0.335289382 0.337735385 0.338099146
0.339400686 0.342017119 0.342030555 0.342848542 0.34450233 0.344598669 0.346979469
0.347757311 0.35 0.350014149 0.350920379 0.351604432 0.354 0.354235295 0.355062425
0.355974229 0.356460184 0.357 0.358047719 0.358206533 0.359338202 0.36119628
0.362526215 0.363099026 0.365732644 0.366195858 0.36634773 0.367310485 0.371139311
0.371368873 0.372692662 0.372896167 0.376700387 0.378014195 0.381764857 0.382

0.38215664 0.385427169 0.386 0.386209011 0.387055635 0.387433926 0.389063616
0.390895123 0.392519511 0.393529549 0.394364565 0.394610116 0.396234139
0.397451788 0.398140103 0.399479155 0.400366979 0.4009341 0.401708402 0.403963238
0.405731071 0.409014299 0.411549646 0.411948712 0.412 0.412386389 0.416420591
0.417541255 0.42 0.420514733 0.421870612 0.423087345 0.423712319 0.425509107
0.425757437 0.426 0.426000171 0.428 0.430067734 0.430386173 0.430834215
0.430916309 0.435 0.436 0.438181311 0.439393041 0.440796093 0.441311847 0.444
0.448116362 0.448539281 0.448616775 0.451 0.454219128 0.462 0.466 0.46982317 0.472
0.473877359 0.478 0.480848043 0.481530697 0.497 0.524 0.537714157 0.539 0.583
0.583199801

Number of observations 195

The GLM Procedure

Dependent Variable: pheno

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.04086172	0.01021543	0.93	0.4459
Error	190	2.08022181	0.01094854		
Corrected Total	194	2.12108353			

R-Square	Coeff Var	Root MSE	pheno Mean
0.019265	32.43056	0.104635	0.322644

Source	DF	Type I SS	Mean Square	F Value	Pr > F
method	4	0.04086172	0.01021543	0.93	0.4459

Source	DF	Type III SS	Mean Square	F Value	Pr > F

method	4	0.04086172	0.01021543	0.93 0.4459

The SAS System 11:03 Tuesday, June 1, 2019 3

The GLM Procedure

t Tests (LSD) for pheno

NOTE: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	190
Error Mean Square	0.010949
Critical Value of t	1.97253
Least Significant Difference	0.0467

Means with the same letter are not significantly different.

t Grouping	Mean	N	method
A	0.34416	39	real
A			
A	0.33032	39	rrBLUP
A			
A	0.32527	39	ANN
A			
A	0.30929	39	Ensemble
A			
A	0.30419	39	DeepGS

SAS output for objective 2

The SAS System 15:10 Monday, August 11, 2019 1

The GLM Procedure

Class Level Information

Class	Levels	Values
method	3	DeepGS Ensemble RR-BLUP
pheno	21	0.6585897 0.6944731 0.7119991 0.7207707 0.736325667 0.7371999 0.7377223

0.741221467 0.7439464 0.744050067 0.747987867 0.750357167 0.7540554 0.755574567
0.762912433 0.7636205 0.7642739 0.771759533 0.7786999 0.780836167 0.787139567

Number of observations 21

The GLM Procedure

Dependent Variable: pheno

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.00665560	0.00332780	5.24	0.0161
Error	18	0.01142770	0.00063487		
Corrected Total	20	0.01808330			

R-Square	Coeff Var	Root MSE	pheno Mean
0.368052	3.382424	0.025197	0.744929

Source	DF	Type I SS	Mean Square	F Value	Pr > F
method	2	0.00665560	0.00332780	5.24	0.0161

Source	DF	Type III SS	Mean Square	F Value	Pr > F
method	2	0.00665560	0.00332780	5.24	0.0161

The GLM Procedure

t Tests (LSD) for pheno

NOTE: This test controls the Type I comparison wise error rate, not the experiment wise error rate.

Alpha 0.05
Error Degrees of Freedom 18

Error Mean Square 0.000635
Critical Value of t 2.10092
Least Significant Difference 0.0283

Means with the same letter are not significantly different.

t Grouping	Mean	N	method
A	0.76129	7	Ensemble
A			
A	0.75332	7	DeepGS
B	0.72018	7	RR-BLUP

Appendix III. Publications abstract

The Authentic Knowledge: African Journal of Pure and Applied Science Research Volume 1 No.1, 2019 ISSN 2618-0626-p, 2618-0634-e

The Authentic Knowledge: African Journal of
Pure and Applied Science Research
AJPASR - ISSN 2618-0626-p, 2618-0634-e
Volume 1, Issue No.1, 2019
P.O. Box 555 (00202) Nairobi. Kenya
editor@serek.or.ke

SOCIETY OF EDUCATIONAL RESEARCH AND EVALUATION IN KENYA

Predictions of Indigenous Chicken Phenotypes from Genotypes: Comparison Between Machine Learning and conventional Linear Models

Sumukwo Chesang, Thomas Kainga Muasya & Kiplangat Ngeno*
Animal Breeding and Genomics Group, Department of Animal Science
Egerton University, P.O.BOX 536- 20115, Egerton, Kenya
Correspondence: *Email chesangsumukwo@gmail.com*; muasyaky@yahoo.com; aarapngeno@gmail.com*

Abstract

Genomic selection is a new breeding strategy which is rapidly becoming the method of choice of selection. It is useful in predicting the phenotypes of quantitative traits based on genome-wide markers of genotypes using conventional predictive models such as ridge regression best linear unbiased prediction model (BLUP). However, these conventional predictive models are faced with a statistical challenge related to the high dimensionality of marker data, inter and intra-allelic interactions and typically make strong assumptions. Machine learning models can be used as an alternative in the prediction of phenotypes due to their ability to address this challenges. Therefore, the aim of this study was to compare the predictive ability of machine learning using deep convolutional neural network (DeepGS), conventional neural network (Artificial neural network), conventional statistical predictive model ridge regression best linear unbiased (RR-BLUP) model and combination of DeepGS and RR-BLUP(Ensemble model) in predicting body weight (BW) of indigenous chicken based on genome-wide markers. The pearson correlation coefficient (PCC) results from this study for the four models were 0.891,0.889, 0.892 and 0.812 for DeepGS, RR-

International Journal of Recent Research and Review, Vol. XII, Issue 3, September 2019
ISSN 2277 – 8322

Ranking Individuals Based on Predicted Performance Using Mean Normalized Discounted Cumulative Gain Value

Sumukwo Chesang[1], Thomas Kainga Muasya[2], Kiplangat Ngeno[3]
[1,2,3]*Animal Breeding and Genomics Group, Department of Animal Science Egerton University P.O.BOX 536- 20115, Egerton, Kenya*

Abstract **- There has been a growing interest in developing new statistical models and algorithms for predicting untested phenotypes in schemes commonly used in genomic selection (GS). Accuracy of genomic prediction always relies on an appropriate choice of a statistical model to capture the relationship between the genetic architecture of a trait and the underlying marker calls in a panel of high-density marker data. However,the ranking problem has become an important topic in machine learning, partly due to its widespread applications in many decision-making processes because the measures of rank quality are usually based on sorting, which is not directly optimizable. To counter this, mean normalized discounted cumulative gain value (MNDCGV), a standard quality measure in information retrieval with capabilities of ranking individuals according to breeding values has been proposed. Few studies have emphasized on the ranking of individuals based on predicted phenotypic values using MNDCGVs improvement Ensemble model performed very well in ranking individuals with better performance compared to DeepGS and RR-BLUP, with improvement values of 0.01 and 0.11 over Ensemble model respectively.Thus, Ensemble model can be given a top priority as GS model for performance improvement.**

Keywords **- Mean normalized discounted cumulative gain value, Cross-validation, Genomic selection**

I. INTRODUCTION

Traditional breeding programs have led to substantial genetic improvement this is attributed to its abilityto utilizes phenotypic or pedigree information in prediction of breeding values. However,with technological advances, there has been a growing interest in developing new statistical models and algorithms capable of predicting untested phenotypes in schemes commonly used in genomic selection

International Journal of Veterinary Sciences and Animal Husbandry



ISSN: 2456-2912
VET 2020; 5(3): 21-26
© 2020 VET
www.veterinarypaper.com
Received: 04-03-2020
Accepted: 06-04-2020

Sumukwo Chesang
Animal Breeding and Genomics Group, Department of Animal Science Egerton University PO Box, Egerton, Kenya

Thomas Kainga Muasya
Animal Breeding and Genomics Group, Department of Animal Science Egerton University PO Box, Egerton, Kenya

Kiplangat Ngeno
Animal Breeding and Genomics Group, Department of Animal Science Egerton University PO Box, Egerton, Kenya

Application of artificial neural network (ann) to evaluate extend of non-linearity among explanatory variables within and between genotypes and phenotypes

Sumukwo Chesang, Thomas Kainga Muasya and Kiplangat Ngeno

Abstract
Artificial neural networks (ANN) have been described as an additional model for marker-based genomic predictions of complex traits in the field of animal breeding. This is because of their ability to accommodate noisy, nonlinearity in data set and they make decisions based on prior knowledge. 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. 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. The 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

Appendix IV. Research permit

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: 572908

Date of Issue: 10/March/2020

RESEARCH LICENSE



This is to Certify that Mr.. CHESANG SUMUKWO of Egerton University, has been licensed to conduct research in Nakuru on the topic: PREDICTION OF INDIGENOUS CHICKEN GENOMIC BREEDING VALUES AND PERFORMANCE USING MACHINE LEARNING for the period ending : 10/March/2021.

License No: NACOSTI/P/20/2742 Ammended

572908

Applicant Identification Number

Director General

EULIB
072988