UNIVERSIDADE ESTADUAL PAULISTA – UNESP

CÂMPUS DE JABOTICABAL

APPLICATION OF ARTIFICIAL NEURAL NETWORKS TO GENOME-ENABLED PREDICTION IN NELLORE CATTLE

André Mauric Frossard Ribeiro

Animal Scientist

2019

UNIVERSIDADE ESTADUAL PAULISTA – UNESP

CÂMPUS DE JABOTICABAL

APPLICATION OF ARTIFICIAL NEURAL NETWORKS TO GENOME-ENABLED PREDICTION IN NELLORE CATTLE

André Mauric Frossard Ribeiro

Advisor: Prof. Dr. Henrique Nunes de Oliveira

Thesis presented to the *Faculdade de Ciências Agrárias e Veterinárias – Unesp*, Campus of Jaboticabal in partial fulfillment of requirements for the degree of Ph.D. in Genetics and Animal Breeding.

2019

R484a

Ribeiro, André Mauric Frossard

Application of artificial neural networks to genome-enabled prediction in Nellore cattle / André Mauric Frossard Ribeiro. -- Jaboticabal, 2019

36 p.

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal

Orientador: Henrique Nunes de Oliveira

1. Genomic selection. 2. Machine learning. 3. Zebu. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: APPLICATION OF ARTIFICIAL NEURAL NETWORKS TO GENOME-ENABLED PREDICTION IN NELLORE CATTLE

AUTOR: ANDRÉ MAURIC FROSSARD RIBEIRO
ORIENTADOR: HENRIQUE NUNES DE OLIVEIRA

Aprovado como parte das exigências para obtenção do Título de Doutor em GENÉTICA E MELHORAMENTO ANIMAL, pela Comissão Examinadora:

Prof. Dr. HENRIQUE NUNES DE OLIVEIRA
Departamento de Zootecnia / FCAV / Unesp - Jaboticabal

Prof. Dr. NICOLA VERGARA LOPES SERÃO (Videoconferência)
Departament of Animal Science-Iowa State University / Ames/Iowa

Ana Fabrícia Braga Magalhães
Pesquisadora Dra. ANA FABRÍCIA BRAGA MAGALHÃES
Instituto de Zootecnia / Sertãozinho/SP

Maria Eugênia Z. Mercadante
Pesquisadora Dra. MARIA EUGÊNIA ZERLOTTI MERCADANTE
Instituto de Zootecnia / Sertãozinho/SP

Pós-doutorando GERARDO ALVES FERNANDES JÚNIOR
Departamento de Zootecnia / FCAV / UNESP - Jaboticabal

Jaboticabal, 29 de julho de 2019

AUTHOR'S CURRICULAR DATA

André Mauric Frossard Ribeiro – born on March 22nd of 1986 in Santos-SP, son of Araken Frossard Ribeiro and Teresa Francisca Mauric Ribeiro. Started his undergraduate education in Animal Science at *Universidade Federal de Viçosa*, in May 2006 and concluded it in July 2010. From 2010 until 2012, he did an exchange internship in Iowa-USA and Wisconsin-USA. He started the Master degree in 2012 in Genetics and Breeding at *Universidade Federal de Viçosa* and concluded in 2015. He started his Ph.D. course in Genetics and Animal Breeding at School of Agricultural and Veterinarian Sciences - São Paulo State University, Campus of Jaboticabal in August 2015 under Prof Dr. Henrique Nunes de Oliveira. During the Ph.D., he did an internship at Iowa State University under Dr. Nick Serão.

I dedicate this work to my brother Leandro, who always has been and always will be with me.

Until one day bro.

ACKNOWLEDGMENTS

A Faculdade de Ciências Agrárias e Veterinárias - Câmpus de Jaboticabal, por ter me proporcionado a oportunidade.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior -Brasil (CAPES) -Código de Financiamento 001. Também pela concessão da bolsa.

A Associação Nacional de Criadores e Pesquisadores pela oportunidade de realizar esse trabalho.

Ao meu orientador Henrique Nunes de Oliveira pelo tempo, confiança e paciência durante todos esses anos.

Aos membros da banca do Exame Geral de Qualificação, Prof. Dr. Roberto Carvalheiro, Prof. Dr. Fernando Sebastián Baldi Rey, Profa. Dra. Lucia Galvão de Albuquerque e Dr. Francisco Ribeiro de Araújo Neto, pelas sugestões que contribuíram e acrescentaram a este trabalho.

Aos membros da banca e defesa de Tese Dra. Maria Eugênia Zerlotti Mercadante, Dra. Ana Fabrícia Braga Magalhães e Dr. Gerardo Alves Fernandes Júnior pela participação na banca de defesa e grande contribuição. Especialmente ao Prof. Dr. Nicola Serão pelos seus eternos ensinamentos, preciosos conselhos e pela sua confiança em mim.

A todos meus amigos do PPGMA Andrés, Bianca, Baltazar, Samuel, Lúcio, Anderson, Tiago, Lucas, Laysa, William, Patrícia, Taíse, Gabi, Bruna, Rafael, Tonussi, Ivan, Daiane, Diércles, Diogo, Sirlene, Gustavo, Juan, e Sabrinas pela amizade e os cafezinhos durante estes anos do Doutorado.

Aos meus colegas da salinha 16 e 17, Cher, Ligia, Francisco, Hayala, Marisa, Andrea, Donicer e Camila, que participaram diretamente e indiretamente de toda essa trajetória na FCAV.

Aos meus amigos da UFV, Darlene, Carol, Renata e toda a galera da ZOO6.

A toda a galera do basquete “Team Lino”.

Aos meus pais por todo apoio em todas minhas decisões, pelos inúmeros conselhos e por serem meus exemplos. Por serem a coisa que eu mais me orgulho de ter.

A Laura pelo amor infinito e por suportar meus defeitos, tolerar meus humores e, principalmente, por me entender (ou tentar entender).

A Lolla por ser minha alegria quando chego em casa e me fazer esquecer de todos meus problemas.

As minhas irmãs Carol e Gabi por serem exemplos de mulheres fortes e determinadas. Por estarem sempre preocupadas com o irmazão delas e ter a certeza que sempre vão estar do meu lado independente de qualquer coisa.

Ao meu tio Toni pelo conhecimento e sabedoria, e sempre me incentivar a busca-los.

Ao meu irmão Leandro, pela maior amizade do mundo e sempre me proteger mesmo nos momentos onde não havia mais forças. A pessoa mais corajosa e guerreira que eu conheci. Tenho tantos momentos para agradecer que dizer obrigado é muito pouco, por isso tudo em minha vida, não só essa Tese, agradeço e dedico ao meu irmão guerreiro.

Muito OBRIGADO!

SUMMARY

LIST OF TABLES

LIST OF FIGURES

APLICAÇÃO DE REDES NEURAIS ARTIFICIAIS NA PREDIÇÃO GENÔMICA AMPLA EM BOVINOS NELORE

Resumo – Nos últimos anos, o rápido desenvolvimento de tecnologias de sequenciamento de alto rendimento permitiu a genotipagem em larga escala de milhares de marcadores genéticos. Diversos modelos estatísticos foram desenvolvidos para predizer os valores genéticos para traços complexos usando as informações de marcadores moleculares em alta densidade, pedigrees ou ambos. Esses modelos incluem, entre outros, as redes neurais artificiais (RNA) que têm sido amplamente utilizadas em problemas de previsão em outros campos de aplicação e, mais recentemente, para predição genômica. O objetivo deste trabalho foi avaliar o desempenho de redes neurais artificiais na predição genômica de bovinos Nelore. Para isso foram testadas diferentes arquiteturas de rede (1 a 4 neurônios em camada oculta), 5 estratégias para seleção de animais com base na acurácia do EBV a serem declaradas para a rede de treinamento como entrada e avaliação de matrizes de relacionamento (NN_G (G como entrada); NN_GD (combinados G com D); e N_Guar (Guar como entrada)) a serem utilizados como entrada para predição genômica em características de peso corporal de bovinos Nelore em relação a modelos de regressão lineares bayesianos hierárquicos (BayesB). . Para isso, utilizou-se o dEBV de 8652 animais genotipados para peso corporal aos 120 dias, 240 dias, 365 dias e 455 dias. Esses animais foram divididos pela acurácia do EBV em população de treinamento e na validação. Todas as estratégias foram repetidas 5 vezes e a correlação entre dEBV e dEBV previsto foi usada como a medida de precisão dos modelos testados. Não havia evidências de que redes mais complexas (com mais neurônios) produzissem melhores previsões quando usamos NN_G ou NN_GD. Possivelmente, isso ocorreu porque o dEBV para o peso corporal foi estimado sob um modelo aditivo de herança em que o mérito genético aditivo tem uma relação linear com os efeitos do SNP. Tanto para NN_G quanto para NN_GD, quanto maior o número de animais na maior população de treinamento, maior foi a capacidade de previsão das características do peso corporal. No entanto, ao avaliar o cenário com o mesmo tamanho da população treinada, podemos observar que os modelos de treinamento com animais com maior acurácia do EBV apresentaram maior capacidade

preditiva. Assim, as redes neurais artificiais não são apenas impactadas pelo número de animais no grupo de treinamento, mas também pela precisão do EBV desses animais. Além disso, todos os modelos de redes apresentaram melhores previsões quando comparados com BayesB, para cenários com poucos animais na população treinada, e podem ser uma ferramenta importante para programas ou características que possuem poucos animais genotipados. Também empregamos o agrupamento K-means para relações genômicas aditivas entre todos os animais genotipados para dividir os animais em grupos de treinamento e validação, com o objetivo de aumentar as relações dentro do grupo e diminuir entre grupos para a validação cruzada. O método de agrupamento K-means foi aplicado a uma matriz de dissimilaridade contendo elementos de um menos a relação genômica aditiva entre pares de animais para dividir o animal genotipado em quatro grupos. Os resultados mostram uma variação considerável na precisão entre os grupos. Em geral, as características de peso corporal com maiores valores de herdabilidade (p365 e p455) apresentaram maior precisão de predição. O grupo utilizado como população de referência com menor relação genômica com as populações-teste mostrou que as redes neurais apresentaram pior capacidade de predição quando comparadas às populações de treinamento com maior grau de parentesco com os grupos teste. Assim, podemos concluir que a capacidade de predição genômica de RNA ao usar a matriz G como entrada é dependente do grau de relação genômica entre a população de treinamento e a população de referência.

Palavras-chave: Máquina de aprendizagem, Seleção Genômica, Zebu

APPLICATION OF ARTIFICIAL NEURAL NETWORKS TO GENOME-ENABLED PREDICTION IN NELLORE CATTLE

Abstract - In recent years, the fast development of high-throughput sequencing technologies has enabled large-scale genotyping of thousands of genetic markers. Several statistical models have been developed for predicting breeding genetic values for complex traits using the information on dense molecular markers, pedigrees, or both. These models include, among others, the artificial neural networks (ANN) that have been widely used in prediction problems in other fields of application and, more recently, for genome-enabled prediction. The objective of this work was to evaluate the performance of artificial neural networks in the genomic prediction of complex trait in Nellore cattle. For this, we has been tested different network architectures (1 to 4 neurons on hidden layer), 5 strategies to select animals based on their EBV accuracy to be declared for the training network as input and evaluation of relationship matrices [NN_G (G as input); NN_GD(combined G with D), and N_G_{uar} (G_{uar} as input)] to be used as input for genomic prediction in body weight traits in Nellore cattle relative to hierarchical linear Bayesian regression models (BayesB) . The dEBV of 8652 animals genotyped for body weight at 120 days, 240 days, 365 days, and 455 days was used. Animals were divided into training population and validation by the predicted EBV accuracy. All strategies were repeated five times, and the correlation between dEBV and predicted dEBV was used as the accuracy measure of the models tested. There was no evidence that more complex networks (with more neurons) produced better predictions when we used NN_G or NN_GD. Possibly, this was because dEBV for body weight trait was estimated under an additive model of inheritance in which additive genetic merit has a linear relationship with SNP effects. For both NN_G and NN_GD, the higher the number of animals in the larger training population was the prediction ability for body weight characteristics. However, when evaluating the scenario with the same size of the training population, we observed the training models with animals with higher accuracy of EBV presented greater predictive ability. Thus, artificial neural networks are not only impacted by the number of animals in the training group but also by the accuracy of the EBV of these animals. Also, all network models

presented better predictions when compared with BayesB, for scenarios with few animals in the training population, and maybe an important tool for programs or traits that have few animals genotyped. We also employed K-means clustering to additive genomic relationships among all genotyped animals to partition animals into training and validation groups, to increase within-group and decrease between-group relationships for cross-validation. The K-means clustering method was applied to a dissimilarity matrix containing elements of one minus the additive genomic relationship between pairs of animals to partition the genotyped animal into four groups. The results show considerable variation in accuracy between groups. In general, body weight traits with higher heritability values (p365 and p455) presented higher prediction accuracy. The group used as the reference population with the lowest genomic relationship with the test populations showed the neural networks showed worse prediction ability when compared to training populations with higher kinship degree with the test groups. Thus we can conclude that the ability of genomic prediction of ANN when using the matrix G as input is dependent on the degree of genomic relationship between the training population and the reference population.

Keywords: Genomic Selection, Machine Learning, Zebu

CHAPTER 1 - GENERAL CONSIDERATIONS

INTRODUCTION

Most of the economically important traits is expected to be controlled by an infinite number of loci, and each locus has an infinitely small additive effect (Fisher, 1918). In this context, Meuwissen et al. (2001) proposed a novel approach to the genome-based prediction of complex traits. This methodology has been called genome-wide selection (GWS) and consists of the simultaneous use of hundreds or thousands of markers, covering all the genome, so that all genes of a quantitative trait are in linkage disequilibrium with at least a part of the markers. Marker effects are estimated as a regression of the phenotype on the genotype in training data sets, i.e., the animals in the population with both phenotypic and genotypic (or genomic) information and these estimates are used to predict enomic estimated breeding values (GEBV) for all individuals with genomic data without any phenotypic information (prediction population).

The fast and continued development of high-throughput genotyping and sequencing technologies has enabled large-scale genotyping of thousands of genetic markers (e.g., single nucleotide polymorphisms, SNPs) in genomes of plant and animal species. It has made it possible for genomic selection (GS) to offer new possibilities for improving the efficiency of animal breeding methods and program.

Beef cattle production in Brazil is very important worldwide, being Nellore the most used beef breed. Although significant genetic progress has been achieved for growth traits in Nellore in the last decades through conventional selection, progress for reproduction, meat quality, and feed efficiency traits has been less significant during the same period. Thus, genomic prediction in beef cattle offers a great promise to predict genetic merits of selection candidates for traits that are difficult and costly to measure, and traits that are measured too late in life and/or by sacrificing potential candidates for reproduction, such that candidates for selection cannot have breeding value with high accuracy at the time when selection decisions are made (Miller, 2010).

The “traditional” parametric prediction models (i.e., RR-BLUP, Bayes A, B, Cπ and BLASSO) based on additive inheritance has been successfully applied to the

improvement of agricultural and livestock species, but often present incompatibility of the real genetic architecture of complex traits, which may involve non-additive effects, such as several types of interactions between genes and genotype-environment interactions (Yang et al., 2010). In this context, there has been an increasing interest in the use of semi-parametric and non-parametric methods for the genomic-enable prediction of quantitative traits by accounting for non-additive and non-linear effects as well as genotype-environment interactions (Gianola et al., 2006; de Campos et al., 2010). An example is artificial neural networks (ANN), which provide a powerful learning technique on complex features by predicting future results based on training data (Shaneh and Butler, 2006).

The ANN is computational models developed in the domain of artificial intelligence to emulate a biological nervous system in the form of a mathematical information processing system. In general, ANN is interconnected artificial neurons organized in several layers that mimic the structure of the human brain (Pereira and Rao, 2009) and can be used to model complex patterns and prediction problems. The ANN is characterized as universal function approximators and are able to capture both linear and nonlinear relationships between inputs and outputs, and can adaptively learn complex functional forms, through a series of transformations (i.e., activation functions) driven by parameters. Multilayer feed-forward is the most common architecture used in ANN, which consists of neurons assembled into layers

The great advantage of ANN is their universal learning ability, which is obtained during a training phase, such that they can learn properties (patterns) from, for example, genomic data, without the need to explicitly define a genetic model. After successful training, the information is stored in the synaptic weights of the ANN connections and is dynamically adapted in a process that is not centralized but distributed in parallel. Thus, an ANN can approximate solutions to problems to the same class that was not explicitly trained, even when the problem is non-linear (Kriesel, 2007). The properly specified ANN can be a powerful tool capable of learning about the complex traits affected by cryptic genomic interactions and for predicting results.

When ANN are training, we often presume perfect data quality. However, it rarely occurs, and data inaccuracy can highly influence performance of the model. In addition, when data quality varies, the potential performance of predictive models can

help a decision-maker to design an appropriate information system in terms of predictive accuracy. Thus, this study aims to evaluate the performance in the use of ANN for genomic prediction in body weight traits in Nellore cattle and compare with a hierarchical linear Bayesian regression models (Bayes B). Also, ANN architectures were explored, for that we test 1 to 4 neurons in the hidden layer to find the architecture with best results.

LITERATURE REVIEW

A BRIEF OVERVIEW OF GENOMIC PREDICTION

Beginning in the 1970s, the development of different molecular marker systems drastically increased the total number of polymorphic markers available to animal and plant breeders. Initially, molecular markers were used to detected quantitative trait loci (QTL) in domestic animals and then integrated into the traditional phenotypic selection by applying marker-assisted selection (MAS). However, MAS has not shown the expected impact in livestock due to several reasons as the need to conduct crosses experiment to create an extensive linkage disequilibrium, only a limited amount of genetic variation for the trait are capture, many QTL and associations could not be replicated.

Meuwissen et al. (2001) proposed and described a novel approach to selected candidates with the assumption that using high marker density will be enough for responsible genes of a trait to be in linkage disequilibrium with these markers. In this approach, named Genomic Selection, each marker effect is estimated and then with the sum of these effects is obtained the breeding value of any animal. For that, firstly, the effects of markers are estimated in a reference population that have been phenotyped and genotyped. After that, the Genomic Breeding Values of candidates for selection without known phenotype could be predicted.

Genomic Selection has been becoming more feasible with large-scale and cheap genotyping methods, and substantial gain have been reached for several species, mainly in dairy cattle, around the world. Its adoption has a huge potential to

enhance genetic improvement programs in livestock since it can increase the rates of genetic gain by shortening the generation interval, selecting candidates early in life, it also enables to select for traits that are expensive or difficult to measure or expressed just in one sex. Several studies have demonstrated the superiority of marker-based models over pedigree-based models for breeding value predictions for several traits both animal and plant (VanRaden, 2008; Hayes et al., 2009; de los Campos et al., 2009; Crossa et al., 2010 and 2011).

There is an infinite number of models and approaches to genomic prediction; most used models are SNP-based regression where phenotypes are regressed on thousands of SNPs, applying different prior distributions to marker effects. In the beginning, the whole-genome selection was implemented using linear regression methods (Meuwissen et al., 2001) as:

$$f(x_{i1}, x_{i2}, \ldots, x_{ip}) = \mu + \sum_{j=1}^{p} x_{ij}\beta_j$$

where μ is an intercept, x_{ij} is the genotype of the ith individual at the jth marker (j = 1, . . ., p), and β_j is the corresponding marker effect. However , with development of high-density SNP panels, several shrinkage or regularization methods, e.g., Bayes A, Bayes B, Bayes C, BayesCpi, Bayesian Lasso, Bayes R, have been proposed for dealing with the problem of large number of markers (*p*) and small number of records (*n*) in SNP regression models, avoiding multicollinearity and overfitting among predictors, thus making the model feasible.

However, these parametric statistical methods used in genomic prediction tend to make strong assumptions about functional forms and the statistical distribution of marker effects (Gianola et al., 2006). Besides that, most of these methods are based on additive inheritance presenting incompatibility with the real genetic architecture of complex traits, which may involve non-additive effects, such as various types of gene interactions and genotype-environment interactions. In this context, semi-parametric and non-parametric methods, e.g., Reproducing Kernel Hilbert Spaces Regression (RKHS), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forests (RF) and Boosting, have been proposed and studied in complex traits by accounting for non-additive and non-linear effects, being an alternative to increasing prediction accuracy (Gianola et al., 2006; de Campos et al., 2010).

This review explores how ANN work and their application in breeding and genetics. The ANN is characterized as universal function approximators and can capture both linear and nonlinear relationships between inputs and outputs and can adaptively learn complex functional forms, through a series of transformations (i.e., activation functions) driven by parameters.

ARTIFICIAL NEURAL NETWORKS (ANN)

The ANN is on the domain of artificial intelligence. They are mathematical methods inspired by a biological neural network (Mammals' brain) consists of billions of interconnected neurons that have the ability to handle complex tasks, such as face recognition, body motion planning, and muscles activities control (Jiang et al., 2017). Due to their ability to handle intrinsically nonlinearities, artificial neural networks are widely applied to such scenarios.

BIOLOGICAL NEURONS

The fundamental unit of neural networks, both biological and artificial, is the neuron. The biological neurons are composed of three main parts: dendrites, a cell body or soma, and an axon (Figure 1). The neurons are connected with the use of axons and dendrites; these connecting regions are referred to as synapses. Dendrites are composed of several thin extensions that form the dendritic tree; their fundamental purpose is to receive continuous stimuli (signals) from several other pre-synaptic neurons or the external environment and transfer the information to the cell body. In the cell body, all signals came from dendrites are combined and processed, producing an activation potential that indicates if the neuron can trigger an electric impulse along its axon (da Silva et al., 2017).

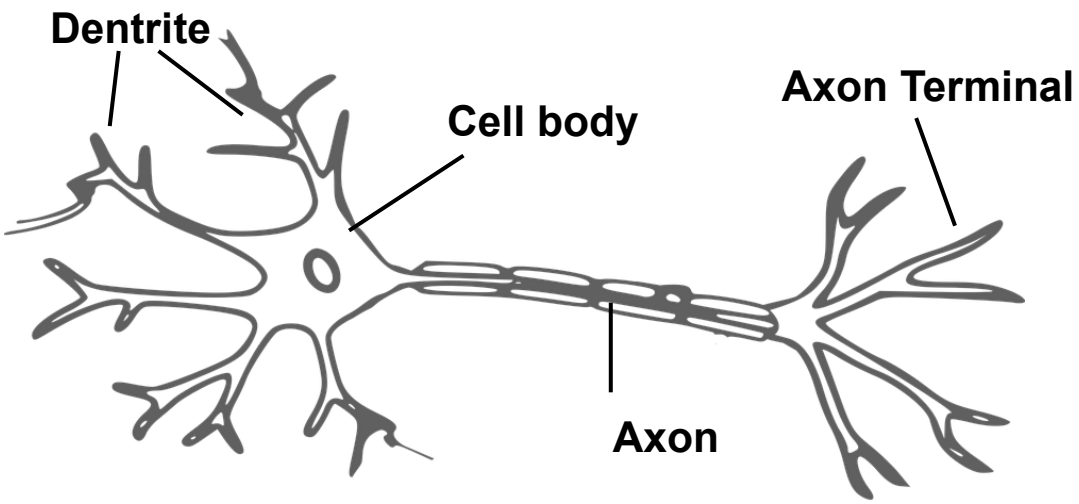

Figure 2. Illustration of the synaptic connection between neurons.

Therefore, one of the essential functions performed by a neuron is the combination of signals received from the previous neurons. If the combination of the received signals is above the excitation threshold of the neuron, an electric impulse is produced and propagated. When the impulse reaches the end of an axon, neurotransmitters are released into the synapses, and the process continues into the next neurons. The functionality of a neuron is dependable of its synaptic weighting, which is also dynamic and dependent on the cerebral chemistry (Hodkin and Huxley, 1952). This change is how learning takes place in living organisms.

ARTIFICIAL NEURONS

In 1943, the neurophysiologist Warren McCulloch and the mathematician Walter Pitts introduced the first models of an artificial neuron (McCulloch and Pitts, 1943), based on the logistic threshold algorithm. McCulloch-Pitts neurons are able to separate Booleans inputs, but any learning process to update weights were established.

McCulloch-Pitts neurons (Figure 2) consists of *n* input signals, which receive the values x_1, x_2, x_3..., x_n, which are coupled to the weights w_{k1}, w_{k2}, w_{k3},..., w_{kn} which values can be positive or negative and *k* is the neuron used. The weights are the adjustable parameters that change as the training sets are presented to the network. The effect of a particular synapse in the postsynaptic neuron is given by $x_i w_{ki}$, the sum

of this combination ($\sum_{i=1}^{n} x_i w_{ki}$) decides whether or not the neuron triggers, by comparing this sum at the threshold of the neuron. The activation of the neuron is obtained through the application of an activation function, which activates or not the output, depending on the value of the weighted sum of its inputs. The most commonly used activation functions are the linear function (purelin), which is recommended for linearly separable problems, the log-sigmoid function, and the hyperbolic tangent function, depending on the characteristics of the data.

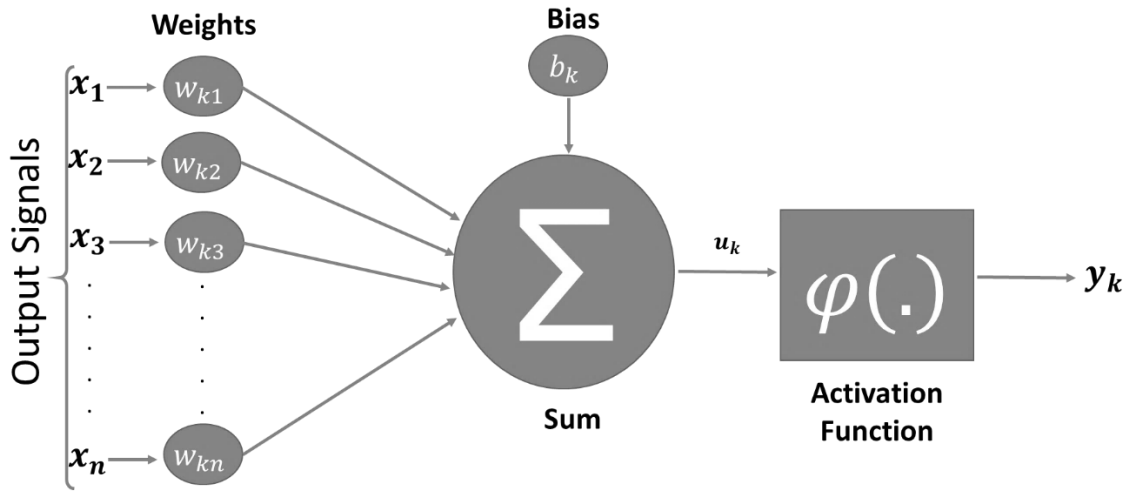

Figure 2. Example of an artificial neuron.

Thus, considering Figure 2, it is verified that the artificial neuron is formed by seven basic elements:

1. Input signals: Signals from the external environment representing values assumed by the variables of a specific application;
2. Synaptic Weights: Values that weighted each of the input variables of the network, quantifying the relevance of a certain neuron;
3. Linear Combiner: Aggregation of all input signals that were weighted by the respective synaptic weights;
4. Activation Threshold: The variable that specifies the appropriate threshold so that the result produced by the linear combiner can generate a trigger value towards the output of the neuron;

5. Activation Potential: The result obtained by the difference between the linear combiner and the activation threshold;
6. Activation function: Responsible for defining the output of the neuron, given the values of the vectors of weight w = (w_{k1}, w_{k2}, w_{k3},..., w_{kn})t and inputs x = (x_1, x_2, x_3..., x_n)t. There serve several import functions, the most commonly used activation functions are the linear function (purelin), which is recommended for linearly separable problems, the log-sigmoid function, and the hyperbolic tangent function, depending on the characteristics of the data;
7. Output signals: The final value produced by the neuron in relation to a particular set of input signals.

The following two expressions summarize the result produced by the artificial neuron:

$$u = \sum_{i=1}^{n} w_i x_i - \theta$$

$$y = \varphi(u)$$

We can summarize the operation of an artificial neuron through the following steps:

1. Presentation of a set of values representing the neuron variables;
2. Multiplication of each entrance of the neuron by its respective synaptic weight;
3. Obtaining the activation potential produced by the weighted sum of the input signals subtracted by the activation threshold;
4. Application of an activation function;
5. Compilation of the output from the application of the neural activation function in relation to its activation potential.

ARTIFICIAL NEURAL NETWORK ARCHITECTURES AND TRAINING PROCESSES

An artificial neural network is defined as a data processing system consisting of a large number of simple highly interconnected processing elements (artificial neurons)

in an architecture inspired by the structure of cerebral cortex of the brain (Tsoukalas and Uhring, 1997).

The architecture of an artificial neural network defines how its several neurons are arranged, or placed, in relation to each other. These arrangements are structured essentially by directing the synaptic connections of the neurons.

An artificial neural network can be divided into three parts, called layers:

1. Input layer: Layer responsible for receiving input signals. These inputs are usually normalized within the limit values produced by activation functions;
2. Hidden layers: Layer composed of neurons which are responsible for extracting patterns associated with the process or system being analyzed;
3. Output layer: Layer composed of neurons responsible for producing and presenting the final network output

There are several classes of ANN. Classified according to their learning mechanism. However, the most used and studied classes of the network are the single layer feedforward network and the multilayer feedforward network.

The SLFN (Figure 3) has just one input layer and a single neural layer, which is also the output layer. Thus, the information always flows in a single direction, which is from the input layer to the output layer, and the number of outputs will always coincide with its amount of neurons. The networks are usually employed in pattern classification and linear filtering problems.

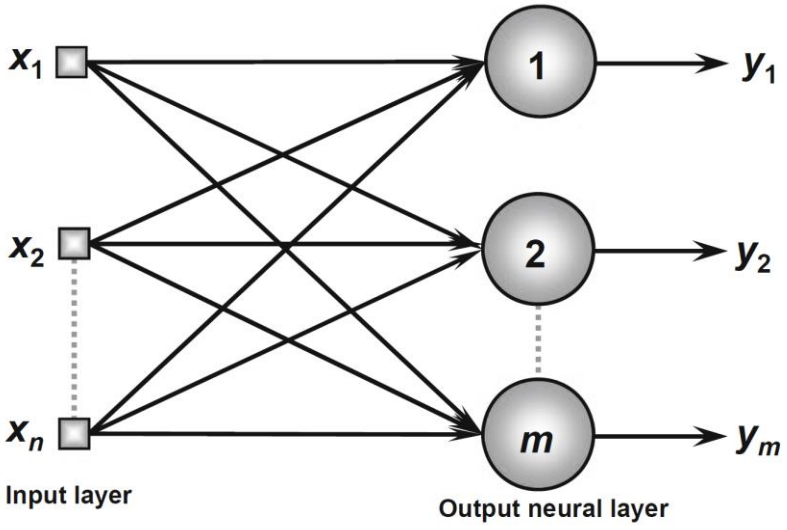

Figure 3. Example of a single-layer feedforward network (da Silva et al., 2016).

The multiple layers are composed of one or more hidden neural layers (Figure 4). They are employed in the solution of diverse problems, like those related to function approximation, pattern classification, system identification, process control, optimization, robotics, and so on.

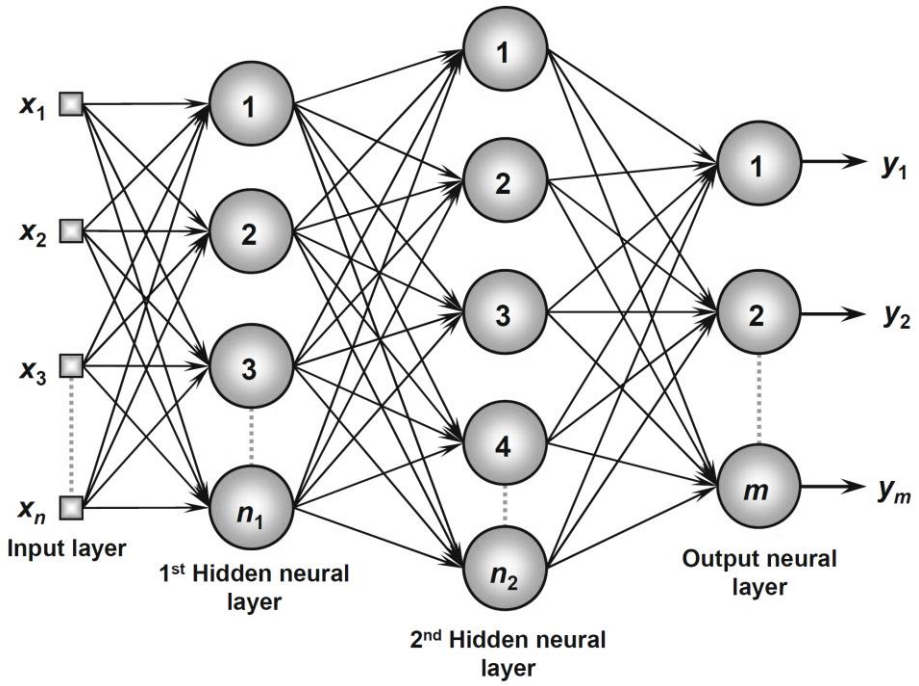

Figure 4. Example of a feedforward network with multiple layers (da Silva et al., 2016).

The architecture of ANNs most used in the genomic prediction of complex characteristics Multilayer Feedforward network, also called Multilayer Perceptron Feed-Foward (MLP) networks. These types of ANNs allow solving of a broader spectrum of problems through the addition of hidden layers, including discrete problems and nonlinear regressions, increasing the space of hypotheses that the network can represent and allowing high computational power. The choice of the activation function used in the hidden MLP layer is related to the underlying function expected for the problem.

Determining the number of neurons to be used in hidden layers is an essential task in ANN architecture. A network with an insufficient number of neurons may not be able to capture complex patterns. In contrast, a network with many neurons may lead

to overfitting, the situation when an ANN is so tightly fitted to the training set that it is difficult to generalize and to decrease predictive capacity.

Among the main networks using multiple-layer feedforward, architectures are the Multilayer Perceptron (MLP), and the Radial Basis Function (RBF), whose learning algorithms used in their training processes are respectively based on the generalized delta rule and the competitive/delta rule.

One of the most relevant features of artificial neural networks is their capability of learning from the presentation of samples (patterns), which expresses the system behavior. Hence, after the network has learned the relationship between inputs and outputs, it can generalize solutions, meaning that the network can produce an output which is close to the expected (or desired) output of any given input values.

The first phase forward-propagation occurs when the network is exposed to the training data, and these cross the entire neural network for their predictions (labels) to be calculated. That is, passing the input data through the network in such a way that all the neurons apply their transformation to the information they receive from the neurons of the previous layer and sending it to the neurons of the next layer. When the data has crossed all the layers, and all its neurons have made their calculations, the final layer will be reached with a result of label prediction for those input examples.

The learning algorithm consists of 6 steps (Figure 5):

1. Start with values (often random) for the network parameters (w_{ij} weights and b_j biases).
2. Take a set of examples of input data and pass them through the network to obtain their prediction.
3. Compare these predictions obtained with the values of expected labels and calculate the loss with them.
4. Perform the backpropagation in order to propagate this loss to each and every one of the parameters that make up the model of the neural network.
5. Use this propagated information to update the parameters of the neural network with the gradient descent in a way that the total loss is reduced, and a better model is obtained.
6. Continue iterating in the previous steps until we consider that we have a good model.

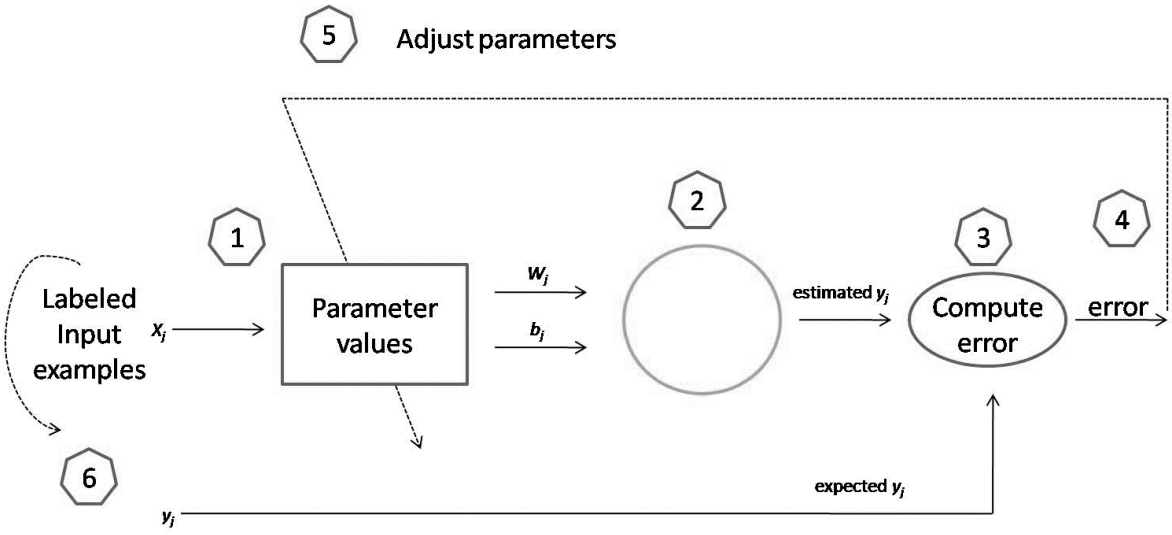

Figure 5. Steps of learning algorithm of artificial neural network.

Application of Artificial Neural Network in Animal and Plant breeding

Macrossan et al. (1999) were the pioneer on the application of ANN in animal breeding predicting first daughter yield phenotypes using dam and sire production and environment effects as an input variable. Bayesian regularized artificial neural networks (BRANN) and non-regularized MLP model were compared with linear and Bayesian linear regression. The models gave similar the correlation between predicted and observed response variables in the testing set (r=0.76), suggesting that the dairy traits analyzed little or none non-linearity. Cavero et al. (2008) described the use of ANN in the early detection and control of mastitis in Holsteins Friesians cows in an automatic system. Several network MLP architectures were tested, and milk traits (days in milk, maximal electrical conductivity, relative deviation of electrical conductivity, milk yield, and milk flow) were used as input to the network. Savegnago et al. (2011) investigated using 2 models of neural networks (MLP and RBF) on curve fitting for egg production from laying hens and compare their performance with a logistic model. They confirmed that MLP could be used as an alternative tool fitting to egg production. Ventura et al. (2012) investigated the use of ANN's to predict the genetic value of body weight at 205 days of age in Tabapuã beef cattle. The variables age of the mother at birth, season (dry or water), region, body weight at 205 days of age were used as input and the genetic value obtained by BLUP as the target of the

network. Felipe et al. (2015) showed that ANN models, after a pre-selection of variables using Bayesian Network, were the best when compared with multiple linear regression.

Gianola et al. (2011) investigated MLPs with Bayesian regularization for genome-enabled prediction of grain yield in wheat and milk production traits IN cattle. The study examined different network structures with up to 6 neurons in the hidden layer and different activation functions. Pérez-Rodriguez et al. (2012) studied the predictive ability of several parametric (Bayes A, Bayes B, and BLASSO) and semi and non-parametric models (RKHS, RBFN and BRANN) applied to production traits n wheat, showing a consistent superiority of RKHS and radial basis function neural network (RBFNN) over the parametric models. Okut et al. (2011) used BRANNs to predict body mass index in mice using SNP markers and concluded that BRANN might be a method for high-dimensional genome-enabled prediction, with the advantage of capturing non-linear relationships. Okut et al. (2013) evaluated the performance of ANN to predict expected progeny difference of marbling scores in Angus cattle. The ANN with Bayesian regularization method performed equally well for prediction of EPD as BayesCpi, based on prediction accuracy and the sum of squared errors

Ehret et al. (2015) examined different ANN architectures, as well as several genomic covariate structures as network inputs (the raw genomic marker matrix (X), genomic relationship matrix (G) and principal component scores (UD) of X) in order to assess their ability to predict milk traits in three dairy cattle. Coutinho et al. (2018) showed the superiority of the ANN over RR-BLUP in predicting GEBV in simulations with higher and lower marker densities, with higher levels of linkage disequilibrium and heritability. Tussel et al. (2013) investigated the predictive ability for litter size in pigs using different models including, Bayesian regularized neural networks (BRANN) and radial basis function neural networks (RBFNN), with different sources of genetic information. BRNN presented the best prediction accuracy in crossbred when the G matrix was used as input (r = 0.31); however, the prediction in one of the purebred had the worst performance (r = 0.02).

González-Camacho et al. (2012) compared the predictive performance of RBFNN with that of RKHS and of the additive linear model (BLASSO) on simulated data and real maize lines genotyped. RKHS and RBFNN had a slight and consistent superioty over the additive BLASSO model. Felipe et al. (2015) investigated the effect

of genotype imputation in the context of whole-genome prediction of complex traits in mice using parameters, semi-parametric and non-parametric models applied to different size of subsets of SNP. They conclude that BRANN seemed more sensitive to imputation errors; therefore, the use of imputed genotypes with this model should be carefully evaluated when using neural networks.

OBJECTIVE

In this study, different network architectures, as well as strategies of input selection to be declared for network and evaluation of relationship matrices to be used with input were tested, with the objective of evaluate the performance of artificial neural networks in the genomic prediction of growth traits in Nellore cattle.

MATERIAL AND METHODS

Phenotypic and genotypic data were provided by the National Association of Farmers and Researchers (*Associação Nacional de Criadores e Pesquisadores* - ANCP). These data sets contained information from Nellore herds located in the southeast and mid-west regions of Brazil that participate in the ANCP breeding program.

The traits evaluated were body weight adjusted to 120 (BW120), 240 (BW240), 365 (BW365) and 455 (BW455) days of age. The contemporary groups (CG) were defined according to each trait: for BW120, BW240, BW365, and BW455 defined as sex, year of birth, the season of birth and management group Each CG containing at least three animals and records outside the interval between 3 standard deviations above and below the mean of the contemporary group (CG) were removed (Table 1).

Table 1. Number of animals, phenotypic mean, standard deviations (SD), minimum values (Min), maximum values (Max) and number of contemporary groups (CG) for body weight traits in Nellore cattle.

Trait	N	Mean	SD	Min	Max	CG
BW120 (kg)	258,972	130.6	20.10	69	192	10,629
BW240 (kg)	237,072	190.2	30.42	97	284	12,885
BW365 (kg)	205,747	241.8	45.22	99	380	13,262
BW455 (kg)	175,511	283.0	53.96	116	458	12,608

The genotype data set consisted of 8,652 animals (males and females) with 960 bulls genotyped with high-density panel (HD) (Illumina Bovine HD BeadChip), 1,000 animals genotyped with medium density panel (50k) (Illumina BovineSNP50 BeadChip) and subsequently imputed to the high-density panel. In addition, 6,692 animals were genotyped with low-density panel (12k) (Clarified Nellore 2.0) and later imputed to the 50k panel and then to HD. The following criteria were used to filter genotypes at a particular locus: only autosomal SNPs were considered, and SNPs with minor allele frequency (MAF) less than 0.02, a Hardy-Weinberg equilibrium *p*-value less than 10^{-5}.

The ANN analyses were applied to the genotyped animals and the Deregressed EBV (dEBV) of the genotyped animals were used as a response variable used for genomic predictions for both ANN and BayesB model used for comparisons.

To obtain the EBV and their accuracies the (co)variance components and genetic parameters analyses were performed via the restricted maximum likelihood (REML) method developed by Patterson and Thompson (1971) and implemented in Wombat software (Mayer, 2007). Because of the more perceptible influence of the maternal additive genetic effects and the permanent maternal factors on BW120 and BW240, these were not considered for BW365 and BW450. Therefore, the genetic analysis was conducted by fitting an animal models as:

$$y = X\beta + Z_1 a + Z_2 m + Wc + e$$

where y is a vector of observations; β is a vector of fixed effects (contemporary groups and covariates); a is a vector of random direct additive genetic effects; m is a vector of random maternal additive genetic effects; c is a vector of random maternal permanent environmental effects; $X, Z_1, Z_2 \; and \; W$ are the respective incidence matrixes related $\beta, a, m \; and \; c$ to observations; and *e* is a vector of residual effects. The same model was used to obtain the estimated breeding values (EBV).

Accuracies of EBV were obtained using the following formula:

$$Acc = \sqrt{1 - PEV \; / \; \sigma_g^2}$$

where PEV is the prediction error variance, and σ_g^2 is the additive genetic variance of the trait.

Deregressed EBV (dEBV) of the genotyped animals were used as a response variable used for genomic predictions for both traditional models and ANN. dEBV was obtained by deregressed using the methodology proposed by Garrick et al. (2009), which removed parent average effects and accounted for heterogeneous variances. In addition, reliabilities of the dEBV and the weighting factor were also estimated following the methodology proposed by Garrick et al. (2009).

In the ANN analyses, we use the elements of the genomic G matrix as predictors. Because artificial neural networks have a high computational requirement, only the elements of G related to a reduced subset of the genotyped animals can be uses as input in the training population. Thus, a selective reduction of animals based on their accuracy was used to form the training population. This reduced subset was used in both parametric and non-parametric models. The five strategies were taken:

1. Animals with EBV accuracy higher than 0.95 (ACC>95).
2. Animals with EBV accuracy higher than 0.90 (ACC>90).
3. Animals with EBV accuracy higher than 0.85 (ACC>85).
4. Animals with EBV accuracy higher than 0.80 (ACC>80).
5. Animals with EBV accuracy higher than 0.75 (ACC>75).

The number of animals for each strategy are presented in Table 2.

Table 2. Number of Animal in each training strategy for each body weight traits.

Trait	h^2	ACC>95	ACC>90	ACC>85	ACC>80	ACC>75
BW120	0.16	94	179	256	331	403
BW240	0.18	92	178	248	309	374
BW365	0.40	152	260	323	484	1482
BW455	0.38	135	230	292	392	999

Because the size of the training population may have impact on prediction accuracy, we also evaluated the prediction models by setting the same number of animals for each accuracy strategy so that the mean accuracy of these animals was close to the reference population Table 2. Thus, the training population size for BW120, BW240, BW365, and BW455 were, respectively, 94, 92, 152, and 135 animals.

We also employed K-means clustering to additive genomic relationships among the 8652 genotyped animals to partition animals into training and validation groups, to increase within-group and decrease between-group relationships for cross-validation. The K-means clustering method was applied to a dissimilarity or distance matrix containing elements of one minus the additive genomic relationship between pairs of animals to partition the genotyped animal into four groups.

We used the Hartigan and Wong algorithm, implemented using R for K-means clustering. The maximum relationship coefficient (g_{max}) was calculated between each animal and all other animals in each of the four partitioned groups so that each animal had four g_{max} values. The density distributions of the four g_{max} values for all animals in a group were used to quantify the quality of the clustering.

Deregressed EBV (dEBV) of the genotyped animals were used as a response variable used for genomic predictions for both traditional models and ANN. dEBV was obtained by deregressed using the methodology proposed by Garrick et al. (2009), which removed parent average effects and accounted for heterogeneous variances. In addition, reliabilities of the dEBV and the weighting factor were also estimated following the methodology proposed by Garrick et al. (2009).

The marker matrix was then reparametrized in the ***W*** matrix with m markers, in which, for any marker locus *j*:

$$w_j \begin{cases} A_1A_1 = 0 - 2p_j \\ A_1A_2 = 1 - 2p_j \\ A_2A_2 = 2 - 2p_j \end{cases}$$

where j = 1, 2, …, m and p_j is the frequency of reference allele. Thus, the construction of the additive relationship (G) matrix follows:

$$G = \frac{WW'}{trace(WW')}$$

Consequently, there is appropriate parameterization of the marker matrix ***S*** for dominance deviations with m markers, in which, for each marker locus *j*:

$$S_j = \begin{cases} A_1A_1 = -2(1-p_j)^2 \\ A_1A_2 = -2p_j(1-p_j) \\ A_1A_2 = 2p_j^2 \end{cases}$$

The dominance genomic relationship matrix was calculated by:

$$G = \frac{SS'}{trace(SS')}$$

We also tested as input the genomic unified additive relationship (G_{UAR}). These methodologies proposed a correction in the calculation of the individuals 'inbreeding coefficient. Thus, the relationship between two genotypes is obtained by:

$$G_{UAR} = \frac{1}{N}\sum_i A_{ijk} = \begin{cases} \frac{1}{N}\sum_i \frac{(x_{ji} - 2p_j)(x_{jk} - 2p_j)}{2p_j\,(1-p_j)} & j \neq k \\ 1 + \frac{1}{N}\sum_i \frac{x_{ji}^2 - (1 + 2p_j)x_{jk} + 2\,p_j^2}{2p_j\,(1-\,p_j)} & j = k \end{cases}$$

In order to compare, genomic predictions were also performed using BayesB models using BGLR package in software R as follows:

$$y = 1_n\mu + Zg + e,$$

where y is the vector of dEBV for the respective trait, μ is the location parameter common to all observations, 1n is a vector of 1's, Z is the incidence matrix of direct genomic breeding values (DGV) and g is the vector of DGV and is assumed to follow a normal distribution N(0,Gσ_g^2), where G is the marker-based genomic relationship

matrix as a genomic relationship matrix and σ_g^2 the genetic variance captured by the markers; e is a vector of random residual effects and is assumed to follow a normal distribution N(0,Iσ_e^2), where I is an identity matrix; and σ_e^2 is the residual variance.

In the Bayesian framework, genomic analyses were also performed using BGLR software. The allelic substitution effect of each SNP was estimated using BayesB with π=0.997 for all body weights, which were fitted with values in the covariate codes like 0, 2 (for homozygotes) and 1 (for heterozygotes) using the following model:

$$y = 1_n\mu + Mu + e$$

where y is the vector of dEBV for the respective trait, μ is the location parameter common to all observations, 1n is a vector of 1's, μ is the overall mean, 1 is a vector of ones, u is the vector of the substitution effect, M contains the genotype (A_1A_1 = 2, A_1A_2 = 1, and A_2A_2 = 0) for each individual and each marker, and *e* is the vector of residual effects, assumed follow normal distribution N(0,**I**σ_e^2), where σ_e^2 is the residual variance.

The direct genomic value (DGV) for individual i within a validation set was derived as the sum of predicted effects of SNP posterior means overall k marker effects estimated in the training set:

$$DGV = M\hat{u}$$

where DGV is the vector of direct genomic values estimated with the marker genotypes, M is an incidence matrix that relates genotypes to individuals, and û is the vector of SNP effects which is estimated by either one of the two methods described below.

The ANN architecture used was the Single Hidden Layer Feed Forward Neural Network (SLNN, Figure 6) with the number of neurons in this hidden layer defined via model comparison, testing 1 to 4 neurons on a hidden layer. One of the NN used as input signals is a vector of additive genomic relationship from G of i^{th} animal ($g'_i = (g_{i1},$

$g_{i2, ..., } g_{i8652)}$,i= 1...,8652, NN_G) with all 8652 genotyped animals, besides that we also fitted the NN models using like input vector from G and D matrices (NN_GD).

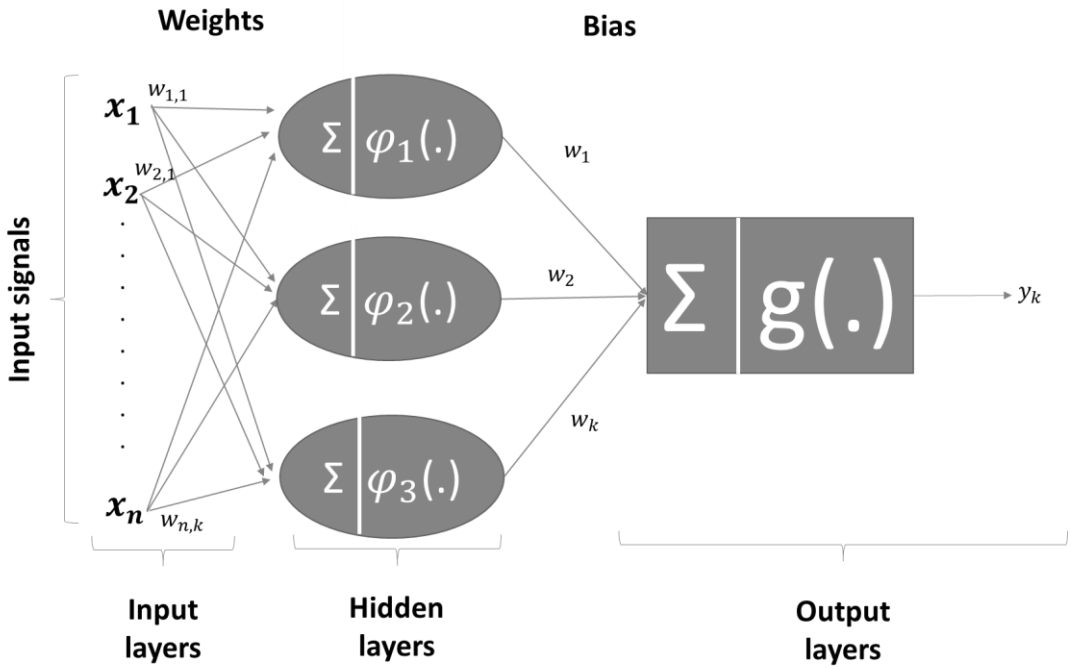

Figure 6. Illustration of Single-layer feed-forward neural network. The *x* values are values of *n* input variables, such as G and D, and the value of y_k (phenotype) as predicted by the network.

To represent the behavior of the synapses, the input terminals are coupled to the weights (w_{kj}, k = 1, ..., K neurons), and connected to the neurons in the hidden layer with their appropriate biases ($b_1^{(1)}$, $b_2^{(1)}$, ..., $b_k^{(1)}$). The input into neuron K, before activation, is expressed linearly as $b_k^{(1)} + \sum_{j=1}^{p} w_{kj} x_j$ where $b_k^{(1)}$ is a parameter of biases defined in the hidden layer. This value is transformed by a linear or non-linear activation function (φ_k) as:

$$f_k\left(b_k^{(1)} + \sum_{j=1}^{p} w_{kj} x_j\right)$$

In order to fit some nonlinear relationship between the output and input values, the hyperbolic tangent function $(\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}})$ will be used in the neurons of the

hidden layer, giving the ANN greater flexibility than the standard linear regression models (Mackay, 2003).

The output signals of the hidden layer are used as input signals to the output layer and transformed as:

$$b^{(2)} + \sum_{k=1}^{K} w_k \, f_k \left(b_k^{(1)} + \sum_{j=1}^{p} w_{kj} x_j \right).$$

where w_k is specific synaptic weight for the k^{th}neuron and b_k^2 is the parameter defined in the output layer. This value is then activated by the following function:

$$g(.) = g \left[b^{(2)} + \sum_{k=1}^{K} w_k \, f_k \left(b_k^{(1)} + \sum_{j=1}^{p} w_{kj} x_j \right) \right].$$

The ANN was training minimizing the error function which is in the function of K synaptic weights (*w's*), and consequently, these weights are iteratively updated by a back-propagation learning algorithm, to approximate the target variable. The updating is usually accomplished by back-propagating the error, which is essentially a non-linear least-squares problem. Back-propagation is a supervised learning algorithm based on a suitable error function, the values of which are determined by the target and the predicted outputs of the network. Weights in an SLNN architecture are determined by a back-propagation algorithm to minimize the sum of squares of errors using gradient-descent methods. During the training, weights and biases in the ANN are successively adjusted based on the input and the desired output. Each iteration of feed-forward in an SLNN constitutes two sweeps: forward activation to produce the desired output, and backward propagation of the computed error to update the values for the weights and biases. The forward and backward sweeps are repeatedly performed until the ANN solution agrees with the desired value to within a pre-specified tolerance.

The learning back-propagation algorithm used Bayesian regularization proposed by Mackay (1992 and 1995) to avoid the overfitting. This Bayesian approach is accomplished in two steps:

1. Obtain conditional posterior modes of the elements in θ, assuming σ_e^2 and σ_θ^2 are known. These modes are obtained by maximizing:

$$p(\theta/y, \sigma_e^2, \sigma_\theta^2) = \frac{p(\theta/y, \sigma_e^2)\, p(\theta/\sigma_\theta^2)}{p(\theta/\sigma_{e,}^2 \sigma_\theta^2)}$$

2. Update the variance components σ_e^2 and σ_θ^2 by maximizing $p(y/\sigma_{e,}^2 \sigma_\theta^2)$.

where θ = (w1,… ,wk; b_1^1 ,… , b_k^1 ; b_1^2 ,… , b_k^2)´ is the vector of synaptic weights and biases, and p(θ|σ_θ^2) = MN(0,σ_θ^2I) being a priori distribution and where σ_θ^2 is the common variance for all elements of θ

For each of these partitions, models were fitted to the training set data, and prediction accuracy was evaluated in the testing data set. Accuracy was assessed using means of Pearson's correlation between predictions and observations and by the predictive mean squared error.

RESULTS AND DISCUSSION

The Figure 7 and Figure 8 shows the Pearson's correlation coefficients between predicted dEBV and the target dEBV for prediction strategy using different network architecture for all body weight trait using NN_G and NN_GD. The single panels show how the dependency of the average Pearson's correlation coefficients on network architecture neurons in the hidden layer.

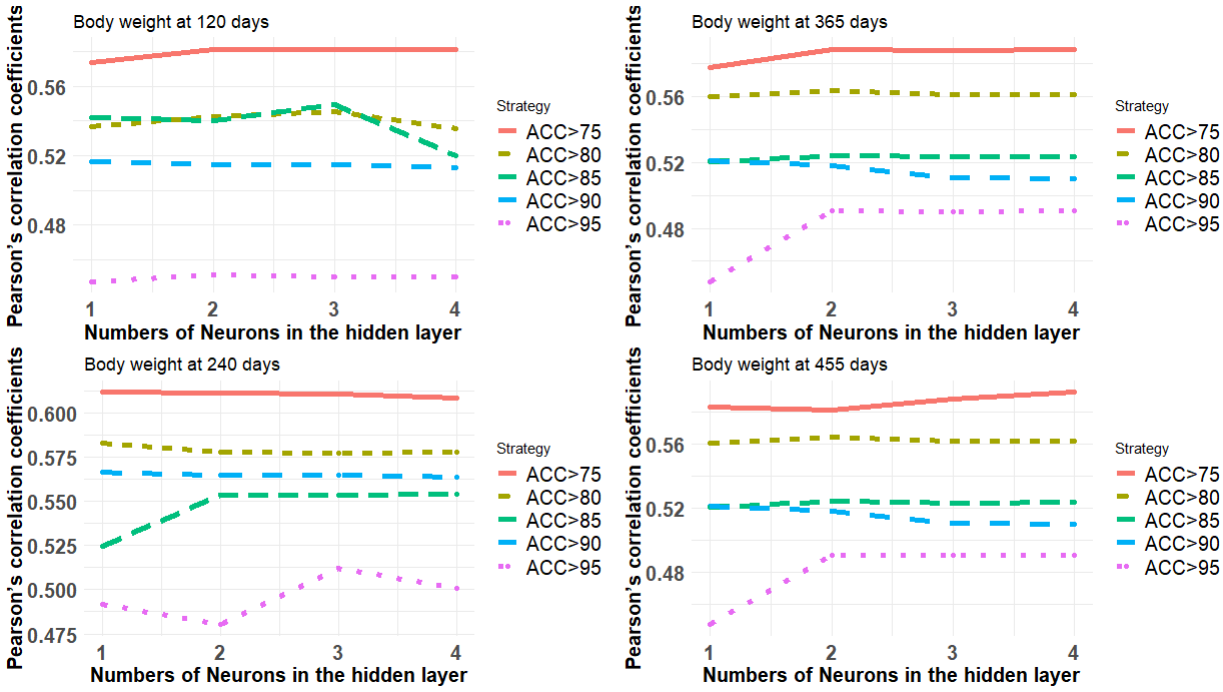

Figure 7. Pearson's correlation coefficients between predicted dEBV and the target dEBV for all body weight trait using G Matrix (NN_G) as input, according to the number of neurons in the hidden layer and prediction strategy Comparison of predictive abilities for all scenarios using artificial neural networks with G matrix as input (NN_G).

The prediction performance of NN_G models (Figure 7) did not show strong changes comparing architecture with different numbers of neurons in the hidden layer. However, for most of the body weight traits with the strategy with less animals on population training (ACC>95) showed the best results when we used 2 or more neurons on a hidden layer. Here, we can observe that independent of the number neurons in the hidden layer, the accuracy for strategies with more individuals in the training population had better result. However, in order to compare these results with other models, we evaluated just the architecture with higher correlation for each strategy.

Results for Pearson correlation coefficient obtained using the NN_G models were more consistent than used NN_GD (Figure 8), mainly for strategies with larger population reference (ACC>75 and ACC>80).

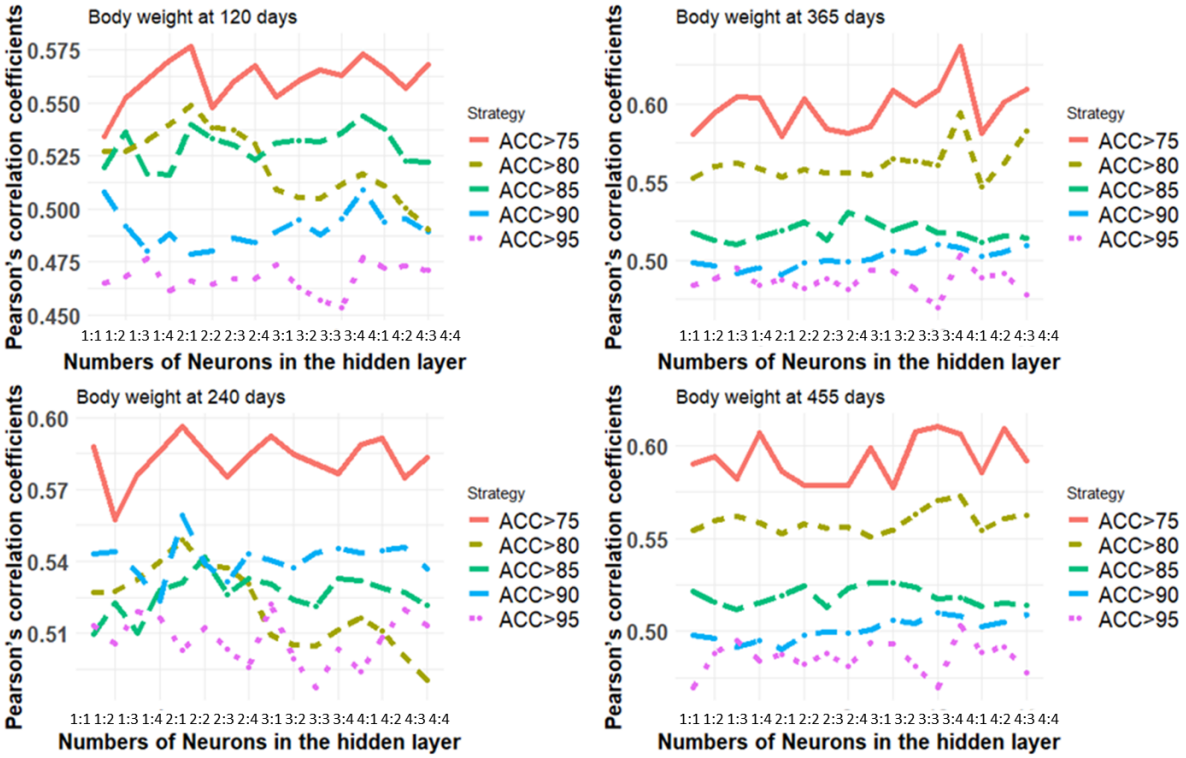

Figure 8. Pearson's correlation coefficients between predicted and target dEBV for all body weight traits using G (NN_G) and D (NN_GD) matrix as input, according to the number of neurons in the hidden layer and prediction strategy (accuracy of the animals in the training set).

In Figure 9a, we clearly observed that with strategies that have a larger training population in our NN_G, they obtained better accuracy of predictions. However, when we set the number of animals for all strategies (Figure 9b), maintaining the mean EBV accuracy, the strategies with animals with high EBV in the training population obtained better results. This may be due to the fact that animals with higher accuracy have a better kinship relationship with the evaluated animals. In this way, the ANNs evaluated is not influenced by the size of the training population but is also being influenced by choice of information used as input.

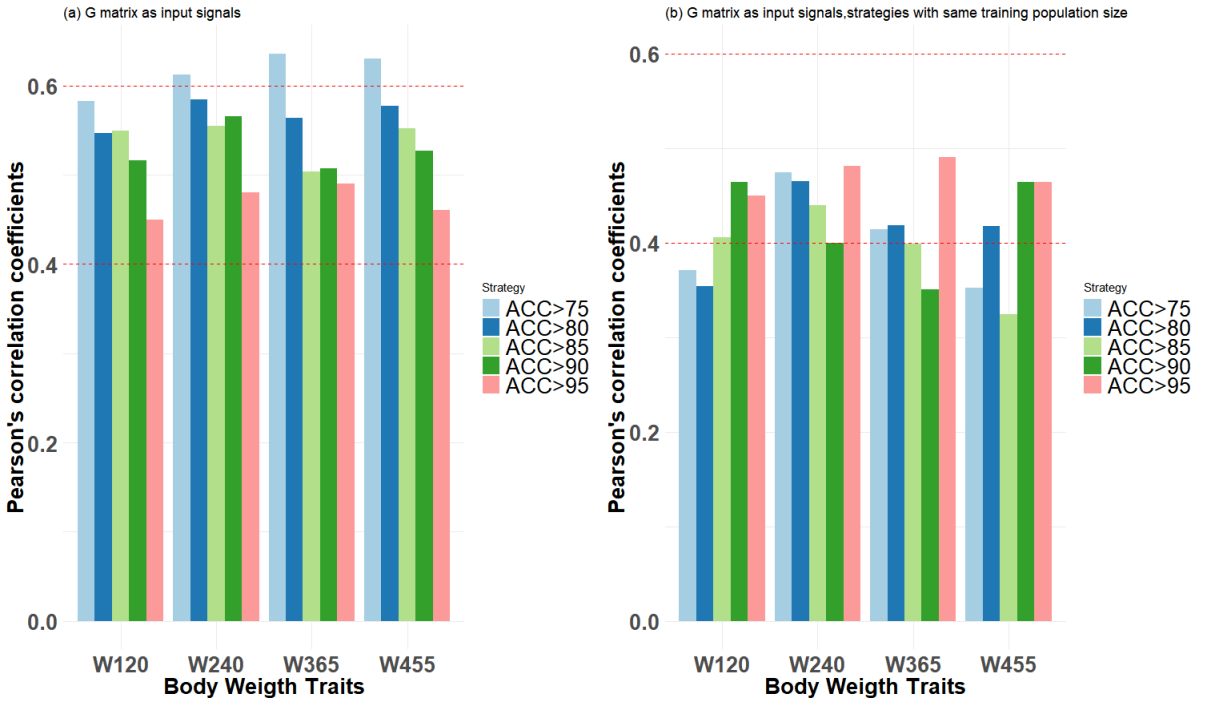

Figure 9. Pearson's correlation coefficients between dEBV and predicted dEBV according to body weight trait for all strategy using G matrix (NN_G) as input for (a) subsets with different sizes in the training population and (b) subsets with the same size for each trait.

The choice of the type of information to be used as the input seems to have further impacted the NN_GD model (Figure 10), mainly for higher traits with higher heritability (W365 and W455). Because dEBV was used as a target output during the training of all networks, no large differences were found between the NN_G and NN_GD model. In this way, the use of the phenotype or the corrected phenotype may better capture non-additive effects, such as dominance.

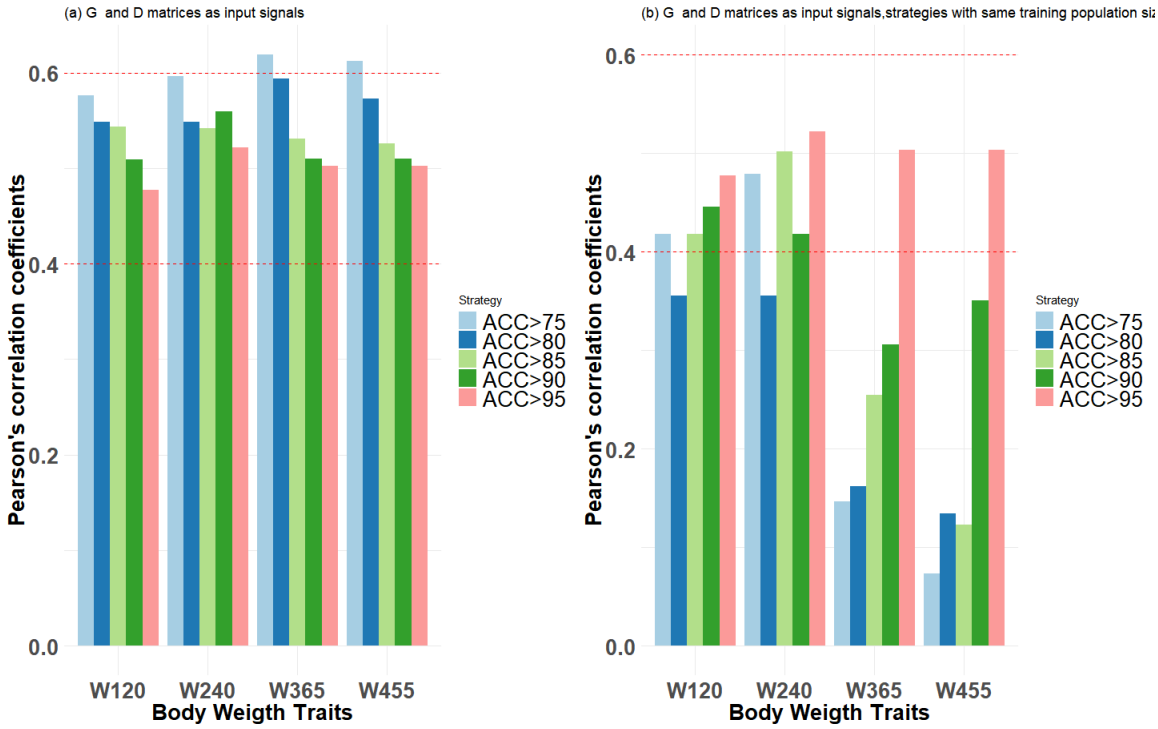

Figure 10. Pearson's correlation coefficients between dEBV and predicted dEBV according to body weight trait for all strategy using G (NN_G) and D (NN_GD) matrix as input for (a) subsets with different sizes in the training population and (b) subsets with the same size for each trait in the training population.

The use of the G_{UAR} matrix as input (NN_GUAR; Figure 11) did not show any improvement in the results, which are very close to those obtained by both NN_G and NN_GD.

Regardless of the model used (Table 3), being ANN or BayesB, with an increase of the training population, the prediction accuracy was higher. When comparing strategies with a small training population (ACC>95 and ACC>90), BayesB presented lower predictive ability than the ANN models, but with the increase of the training population, BayesB predictive ability matched and even surpassed the models ANN, mainly for body weight with higher heritabilities (BW365 and BW455). Thus, a feasible use for artificial neural networks, are situations where the number of animals is small for the training of the prediction model.

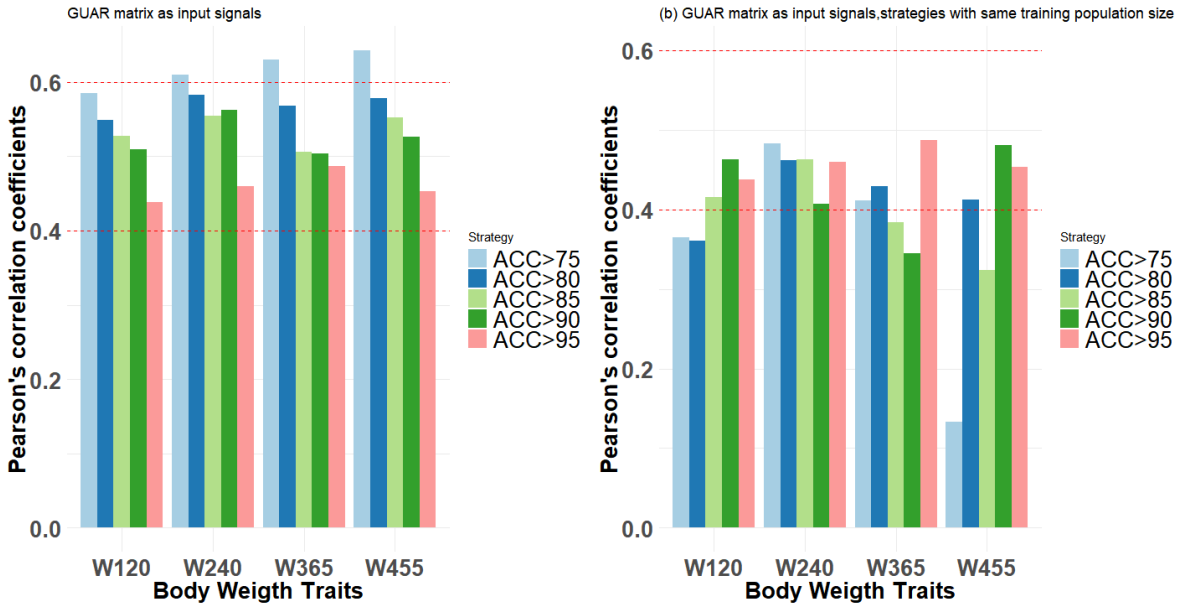

Figure 11. Pearson's correlation coefficients between dEBV and predicted dEBV according to body weight trait for all strategy using NN_GUAR matrix as input for (a) subsets with different sizes in the training population and (b) subsets with the same size for each trait in the training population.

Table 3. Empirical accuracies of genomic predictions obtained for four body weight traits of Nellore cattle based on different methods.

Trait		ACC>95	ACC>90	ACC>85	ACC>80	ACC>75
BW120	NN_G	0.449	0.515	0.550	0.546	0.582
	NN_GD	0.470	0.500	0.543	0.548	0.576
	NN_GUAR	0.437	0.512	0.527	0.549	0.585
	BayesB	0.183	0.191	0.344	0.488	0.601
BW240	NN_G	0.480	0.565	0.555	0.584	0.612
	NN_GD	0.522	0.559	0.541	0.541	0.596
	NN_GUAR	0.450	0.561	0.554	0.582	0.611
	BayesB	0.174	0.206	0.459	0.497	0.613
BW365	NN_G	0.491	0.507	0.503	0.564	0.636
	NN_GD	0.502	0.511	0.53	0.594	0.653
	NN_GUAR	0.486	0.504	0.505	0.567	0.631
	BayesB	0.242	0.289	0.679	0.684	0.752
BW455	NN_G	0.460	0.537	0.551	0.577	0.631
	NN_GD	0.502	0.513	0.526	0.572	0.615
	NN_GUAR	0.453	0.535	0.554	0.591	0.642
	BayesB	0.222	0.23	0.467	0.538	0.698

Accuracy of DGV with K-means

Table 4 shows the number of individuals, g_{ij} and g_{max} within and between the K-means clustered groups.

Table 4. The number of individuals and the averages ± standard deviation within and between-group additive genomic relationships (g_{ij}), maximum within and between-group relationships (g_{max}) for four partitioned groups after K-means clustering.

Groups	**K1**	**K2**	**K3**	**K4**
Number	1742	1568	1660	3828
g_{ij} within groups	0.04 ± 0.068	0.042 ±0.05	0.054±0.06	0.014+0.031
g_{ij} between groups	-0.009 ± 0.068	-0.009 ±0.027	-0.012±0.29	-0.011 ± 0.027
g_{max} within groups	0.44 ± 0.13	0.421 ± 0.11	0.404 ± 0.082	0.35 ±0.11
g_{max} between groups	0.287 ± 0.114	0.198 ±0.11	0.26 ± 0.10	0.24 ± 0.13

Table 4 shows that g_{max} values within a group are much larger than the average of the g_{max} values of a group with the other three groups. The greatest difference between these g_{max} values was for group K2, which had the lowest g_{max} with the other groups (0.198).

Table 5 and Figure 12 expands on the information in Table 4 by showing the the average of the g_{max} values of a group with the other three groups and density distribution of g_{max} of each individual in a particular group with all animals different groups, respectively. Group K2 presented the lowest g_{max} values between the particular groups showing a smaller genomic relationship with the other groups. In contrast, the group K4 presented the highest g_{max} values, mainly with the group K1 and K3.

Table 5. Averages ± standard deviation within and between each particular group for additive genomic relationships (g_{max}).

Groups	K1	K2	K3	K4
K1	0.44±0.13	0.15±0.11	0.21±0.08	0.24±0.12
K2	0.20±0.11	0.42±0.11	0.196±0.104	0.215±0.11
K3	0.192±0.101	0.15±0.09	0.404±0.084	0.18±0.10
K4	0.201±0.131	0.135±0.094	0.154±0.096	0.350±0.113

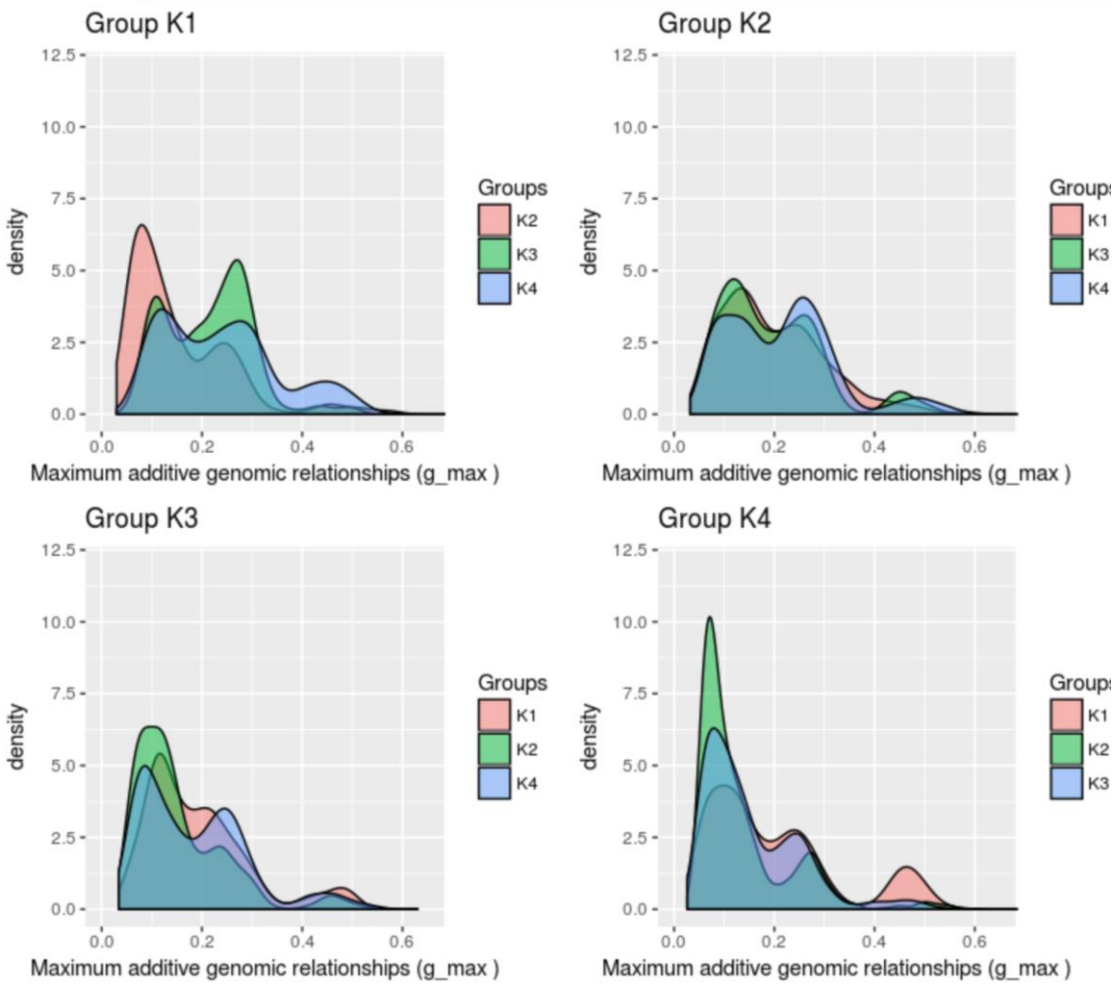

Figura 12. The density distribution of the maximum additive genetic relationships (gmax). The density distribution of the maximum additive genetic relationships (gmax) between each individual in a particular group and all animals in the different groups formed by K-means clustering.

The accuracies of DGV based on training in four groups partitioned by K-means clustering and predicting the other group for weight traits are in figure 13. The results show considerable variation in accuracy between groups. In general, weight traits with higher heritability values (p365 and p455) presented higher prediction accuracy.

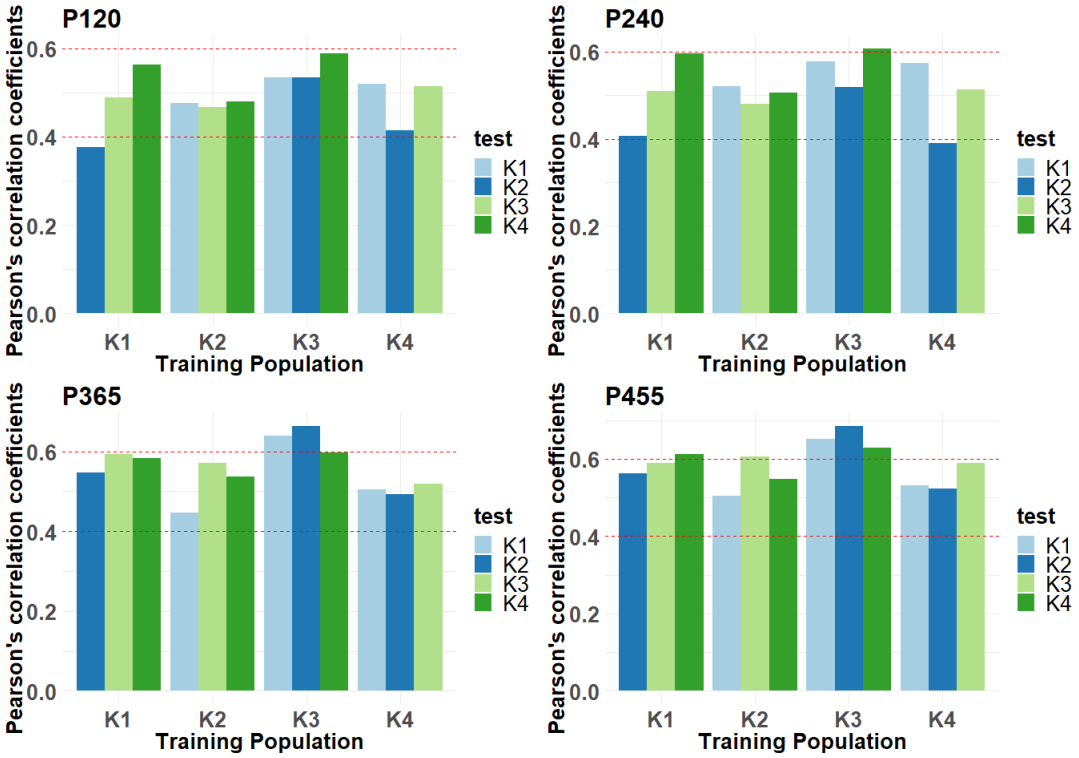

Figure 13. Pearson's correlation coefficients between dEBV and dEBV predicted for body weight traits for K-Means clustering.

We can observe that all weight traits that the group K2, when used as training population, presented the lowest prediction values. This may be due to the lower degree of genomic relationship between these groups, and this population has little informativeness to be captured by ANN. On the other hand, groups K1 and K3 were the groups that obtained the best prediction accuracy values when used as training population mainly with the group 4 where they have greater relation of genomic kinship, being thus more informative populations for the training of the neural networks.

CONCLUSION

Artificial Neural Networks are highly impacted by the size of the training population, but also, they were also influenced by the selection of the input (strategy) to be declared the training network. Artificial Neural Networks trained with animals with EBV of greater accuracy tends to be better models of prediction when compared with those trained with animals of low accuracy. Also, all network models presented better predictions when compared with BayesB, for scenarios with few animals in the training population, and may be an important tool for programs or traits that have few animals genotyped.

The K-means clustering validation confirms that the use of training populations with a higher degree of genomic kin to prediction populations improves the ability of the network to capture the informativeness of the entire population and thus better the genomic prediction of these animals.

REFERENCE

Cavero D, Tølle KH, Henze C, Buxadé C, Krieter J (2008) Mastitis detection in dairy cows by application of neural networks. **Livestock Science** 114:280-286.

Coutinho AE, Neder DG, Silva MCO, Arcelino EC, Brito SG, Carvalho Filho JLS (2018) Prediction Of Phenotypic And Genotypic Values By Blup/Gws And Neural Networks. **Revista Caatinga** 31:532-540.

Crossa J, de los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banziger M, Braun HJ (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. **Genetics** 186:713–724

Crossa J, Pérez P, de los Campos G, Mahuku G, Dreisigacker S, Magorokosho C (2011) Genomic selection and prediction in plant breeding. **J. Crop Improv**. 25:239–226

Da Silva IN, Spatti DH, Flauzino RA, Liboni LHB, Alves SFR (2017) Artificial Neural Networks: A Practical Course. Springer, Berlin

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. **Genetics** 182:375–385

de los Campos G, Gianola D, Rosa GJ, Weigel KA, Crossa J (2010) Semi-parametric genomicenabled prediction of genetic values using reproducing kernel Hilbert spaces methods. **Genet. Res**. 92:295–308

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: GSE**. 47:22

Felipe VPS, Silva MA, Valente BD and Rosa GJM (2015) Using multiple regression, Bayesian networks and artificial neural networks for prediction of total egg production in European quails based on earlier expressed phenotypes. **Poultry Science** 94: 772-780.

Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. **Trans. R. Soc**. Edinb. 52:399–433.

Garrick DJ, Taylor JF, Fernando RL (2009) Deregressing estimated breeding values and weighting information for genomic regression analyses. **Genet Sel Evol**. 41:55.

Gianola D, Fernando RL, Stella A (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. **Genetics** 173:1761-1776.

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

Gonzalez-Camacho JM, De Los Campos G, Perez P, Gianola D, Cairns JE, Mahuku G, Babu R, Crossa J (2012) Genome-enabled prediction of genetic values using radial basis function neural networks. **Theor. Appl. Genet.** 125:759–771.

Hartigan JA, Wong MA (1979) Algorithm AS 136: A k-means clustering algorithm. **Appl Stat.** 28:100-108.

Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME (2009). Accuracy of genomic breeding values in multi-breed dairy cattle populations. **Genetics Selection Evolution** 41:321-329

Hodgkin AL, Huxley AF, Katz B (1952). Measurement of current-voltage relations in the membrane of the giant axon of Loligo. **J Physiol** 116:424– 448.

Jiang Y, Yang C, Na J, Li G, Li Y, Zhong J (2017) A brief review of neural networks based learning and control and their applications for robots**. Complexity**. 1-14

Kriesel D (2007) **A brief introduction to neural networks**. Disponivel em: http://www.dkriesel.com

Mackay DJC (1992) Bayesian interpolation. **Neural Computation** 4:415 – 447

MacKay DJC (1995) Probable networks and plausible predictions—A review of practical Bayesian methods for supervised neural networks. **Network Comput. Neural Syst**. 6:469 – 505

Mackay DJ (2003) Information theory, inference and learning algorithms. Cambridge university press

Macrossan P, Hand D, Kok J, Berthold M, Abbass H, Mengersen K, Towsey M, Finn G (1999) Bayesian neural network learning for prediction in the Australian dairy industry. **Advances in Intelligent Data Analysis**. 1642:395-406.

Meyer K (2007) WOMBAT—A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). **J Zhejiang Univ Sci B.** 8: 815–821.

McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biology** 5:115–133

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. **Genetics** 157:1819 - 1829.

Miller S (2010) Genetic improvement of beef cattle through opportunities in genomics. **Revista Brasileira de Zootecnia** 39:247-255.

Okut H, Gianola D, Rosa GJM, Weigel KA (2011) Prediction of body mass index in mice using dense molecular markers and a regularized neural network. **Genetics Research** 93:189 - 201.

Okut H, Wu XL, Rosa GJM, Bauck S, Woodward BW, Schnabel RD, Taylor JF, 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.

Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. **Biometrika** 58:545-554

Pereira BDB, Rao CR (2009) Data Mining using Neural Networks: A Guide for Statisticians. Disponivel em: http://www.po.ufrj.br/basilio/publicacoes/livros/2009_datamining_using_neural_networks.pdf

Pérez-Rodríguez P, Gianola D, Weigel KA, Rosa GJM, Crossa J (2013) Technical Note: An R package for fitting Bayesian regularized neural networks with applications in animal breeding. **Journal of Animal Science** 91:3522-31

R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org. R version 2.14.0.

Savegnago PR, Nunes BN, Caetano SL, Ferraudo AS, Schmidt GS, Ledur MC, Munari DP (2010) Comparison of logistic and neural network models to fit to the egg production curve of White Leghorn hens. **Poultry Science** 90:705-711

Shaneh A, Butler G (2006) Bayesian Learning for Feed-Forward Neural Network with Application to Proteomic Data: The Glycosylation Sites Detection of the Epidermal Growth Factor-Like Proteins Associated with Cancer as a Case Study. **Canadian Conference on AI 2006: 110-121**

Tsoukalas LH, Uhrig RE (1997) Fuzzy and Neural Approaches in Engineering, J. Wiley & Sons, New York

Tusell L, Pérez-Rodríguez P, Forni S, Wu XL, Gianola D (2013) Genome-enabled methods for predicting litter size in pigs: a comparison. **Animal** 7:1739–1749

Vanraden PM (2008) Efficient Methods to Compute Genomic Predictions. **Journal of Dairy Science** 91:4414-4423

Ventura R, Silva M, Medeiros T, Dionello N, Madalena FE, Fridrich AB, Valente BD, Santos GG, Freitas, LS, Wenceslau RR, Felipe VPS, Corrêa GSS (2012) Use of artificial neural networks in breeding values prediction for weight at 205 days in Tabapuã beef cattle. **Arq. Bras. Med. Vet. Zoo**. 64:411-418.

Yang J, Benyamin B, Mcevoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM (2010) Common SNPs explain a large proportion of the heritability for human height. **Nat. Genet.** 42:565-569.