The Original by FUNPEC-RP

Neural networks and dimensionality reduction to increase predictive efficiency for complex traits

G.N. Silva[1], I.C. Sant'Anna[2], C.D. Cruz[3], M. Nascimento[4], C.F. Azevedo[4] and L.S. Glória[5]

[1]Fundação Universidade Federal de Rondônia, Departamento de Matemática e Estatística, Ji-Paraná, RO, Brasil
[2] Instituto Agronômico – IAC, Centro de Seringueiras e Sistemas Agroflorestais, Votuporanga, SP, Brasil
[3] Universidade Federal de Viçosa, Departamento de Biologia Geral, Viçosa, MG, Brasil
[4] Universidade Federal de Viçosa, Departamento de Estatística, Viçosa, MG, Brasil
[5] Universidade Estadual do Norte Fluminense Darcy Ribeiro, Campos dos Goytacazes, RJ, Brasil

Corresponding author: G.N. Silva
E-mail: gabi.silva@unir.br

Genet. Mol. Res. 21 (1): gmr18982
Received October 11, 2021
Accepted January 11, 2021
Published February 27, 2022
DOI http://dx.doi.org/10.4238/gmr18982

ABSTRACT. The study of complex traits using large databases of molecular markers has reshaped genetic breeding programs as it allows the direct incorporation of information from a large number of molecular markers for the prediction of genomic values. However, the large number of markers can lead to problems of computational demand, multicollinearity, and dimensionality. We evaluated the use of Multilayer Perceptron Neural networks to resolve this problem and propose a new dimensionality reduction method called Probe Subset Selection Methodology, for the prediction of genetic values, in Genome Wide Selection studies. We used a simulated F1 population for 12 quantitative traits, including different modeling structures, average degrees of dominance and heritability. The Multilayer Perceptron Neural Networks, together with the proposed Probe

Subset Selection Methodology, provided more accurate predictions than the RR-BLUP methodology and reduced the root mean square error from 577.249 to values below 24. The use of computational intelligence in breeding programs is a promising tool for prediction purposes, since epistasis and dominance were not limiting factors for the proposed Multilayer Perceptron Neural Network method.

Key words: Artificial Intelligence; Subset selection; Dominance; Epistasis; Prediction

INTRODUCTION

The need for greater genetic gain drove researchers to formulate various breeding strategies in order to obtain more accurate phenotypic observations. Meuwissen (2001) proposed the use of Genome Wide Selection (GWS), a methodology that allows the incorporation of molecular information directly in the prediction of the genetic merit of individuals.

However, methodologies based on GWS present some challenges regarding their applicability. The first is related to the large number of molecular markers that are not associated with QTLs (controlling loci of quantitative characteristics), which makes the number of molecular markers almost always much higher than the number of individuals evaluated and causes dimensionality and multicollinearity issues (Cruz, 2013). In this context, Gianola et al. (2011) and Azevedo et al. (2015) have recommended the use of statistical methods that integrate both the selection of covariates and the regularization of the estimation process.

Another challenge imposed by this methodology is related to the genetic model associated with the quantitative character, which differ from each other, depending on the statistical model employed (Odilon Junior, 2013). Most genetic models include only the additive portion of the genetic value, while neglecting dominance and epistatic interactions, which are important effects for the improvement of prediction accuracy (Sant’anna et al., 2021).

Aiming to overcome these challenges, various researchers have proposed the use of techniques based on computational intelligence, such as Artificial Neural Networks (ANN), since their results depend on learning rather than on the distribution of the variables themselves, so that they can capture nonlinear relationships between markers from the data itself (Long et al., 2011; Cruz and Nascimento 2018).

Multilayer Perceptron Neural Network (MLPNN) are a particular class of ANNs that have properties that make them attractive to genetic breeding programs for the purpose of predicting genetic values and have been used for many authors (Sant’Anna et al., 2015; Silva et al., 2016; Barbosa et al. 2021). In general, these studies conclude that the application of MLPNN in genomic selection is powerful for capturing complex interactions in comparison with semiparametic and linear regressions.

However, the problem of working with genomic selection is related to the high number of markers, which increases the chance of a high correlation between markers and also leads to less precision and a great computational demand for MLPNN training (Cruz and Nascimento, 2018).

Sant'anna et al., (2021) reported that a subset of markers can be used for training and demonstrated that by reducing the search space, ANN can improve the learning process and increase the predictive power of the model.

In the light of the foregoing, we propose a new dimensionality reduction methodology, based on the selection of variables, called Probe Subset Selection and evaluated the efficiency of the Multilayer Perceptron Neural Network (MLPNN) method for prediction in a simulated population, considering different scenarios with difficult situations, for breeding based on sexual reproduction, with the inclusion of dominance, epistasis and environmental effects. The results were compared with those achieved by the RR-BLUP method.

MATERIAL AND METHODS

Data simulation

The comparative study on prediction methods, based on simulated data, consisted of the evaluation of an F1 population, coming from divergent parental line genomes, with 500 individuals and genotyped in relation to 1000 codominant markers. Although low, the number of markers used does not compromise the comparative objectives of the present study. The phenotypic values of the individuals were generated according to equation (1):

$$F_i = G_i + E_i \qquad \text{(Eq. 1)}$$

where: G_i is the genetic effect given by the sum of genetic effects in each locus, and E_i is an environmental effect.

The phenotypic value expressed by a given individual, for quantitative characteristics controlled by 100 gene loci, was obtained by adopting two models: the additive-dominant model (Equation 2) and epistatic model (Equation 3).

$$Y_i = \mu + \sum_{j=1}^{100} p_j \alpha_j + E_i \qquad \text{(Eq. 2)}$$

$$Y_i = \mu + \sum_{j=1}^{100} p_j \alpha_j + \sum_{j=1}^{99} p_j \alpha_j \, \alpha_{j+1} + E_i \qquad \text{(Eq. 3)}$$

where: $\alpha_j = a_i + d_i$and d_i/a_i = degrees of dominance, with $\mu + a_j$, $\mu + d_j$ and $\mu - a_j$, coded with 1, 0 or -1, for the genotypic classes AA, Aa and aa, respectively; p_j is the contribution of locus j to the manifestation of the trait under consideration, generated by binomial distribution, with parameters n = 99 and p =q= 0.5.

For the epistatic model (Equation 3), the first summation refers to the contribution of the individual locus through its additive and dominant effects; the second summation represents the multiplicative effects due the epistatic interactions between pairs of loci: a_j is the multiplicative effect of the favorable allele in locus j, and j+1 is the contribution of locus j to the manifestation of the trait under consideration.

Quantitative traits were simulated by considering three degrees of dominance (d/a = 0, 0.5, and 1) and two broad sense heritability levels (h^2 = 35 and 70), which represented three gene activities: additive, dominance and epistatic, thereby totalling twelve scenarios (Table 1).

Table 1. Simulated scenarios composed of a combination of traits, heritability, and dominance degree.

Traits	Heritability (%)	Model	d
T1 - D0H35Ad	35	additive	0
T2 - D0H35Ep	35	epistatic	0
T3 - D0H70Ad	70	additive	0
T4 - D0H70Ep	70	epistatic	0
T5 - D60H35Ad	35	additive-dominant	0.6
T6 - D60H35Ep	35	epistatic	0.6
T7 - D60H70Ad	70	additive-dominant	0.6
T8 - D60H70Ep	70	epistatic	0.6
T9 - D120H35Ad	35	additive-dominant	1.2
T10 - D120H35Ep	35	epistatic	1.2
T11 - D120H70Ad	70	additive-dominant	1.2
T12 - D120H70Ep	70	epistatic	1.2

Establishment of the number of molecular markers to be selected

In order to solve possible dimensionality problems and or the occurrence of multicollinearity due to the greater number of marks in relation to the number of genotyped individuals, a methodology for reducing dimensionality based on subset (selection of variables) was proposed, named Probe Subset Selection Methodology. This method was developed in two stages. The first was used to establish the appropriate size of the subsample of markers, and the second, to identify the most important markers to compose the final subset to be used for further prediction analysis.

Aiming to determine the optimal number of markers, as suggested by Sant'Anna et al. (2021), the Stepwise regression was used for the most complex characteristic (T10), with epistatic effects, dominance and low heritability. The maximum number of markers was determined based on the combination of three criteria: the determination coefficient (R^2) – best possible values –, the root mean square error (RMSE) – lowest possible values and the condition number (CN) of the correlation matrix. CN is an important measure used to identify the existence of multicollinearity in matrices of variables, through the analysis of the eigenvalues of the X^TX matrix, where X refers to the independent variables of the model (Montgomery and Peck, 1981). CN is expressed by Equation (4):

$$C = \frac{max\ (\lambda_1,\lambda_2,\dots,\lambda_p)}{min\ (\lambda_1,\lambda_2,\dots,\lambda_p)} \qquad \text{(Eq. 4)}$$

where: λ_i is the eigenvalue for the i-th variable, i=1, ..., p.

The CN evaluation was carried out as suggested by Montgomery and Peck (1981): if $CN < 100$, then there were no multicollinearity problems; if $100 < CN < 1000$, then the multicollinearity was moderate to severe; and if $NC > 1000$, then severe multicollinearity was considered.

Probe Subset Selection Methodology

The Probe Subset Selection Methodology proposed in this work consists of identifying the most important variables through a procedure inspired by the genetic algorithm (Holland, 1975). The genetic algorithm is based on the coding of a set of all possible solutions for a given problem, so that, through an exhaustive search and inspired by evolutionary biology (heredity, mutation, natural selection, etc.), the best solution will be determined (Goldberg, 1989; Raymer et al., 2000).

The use of the genetic algorithm in its original form for studies of genomic selection is impracticable, as it would mean testing all possible combinations (C_{1000}^{x}) between the molecular markers, where x refers to the optimal number of markers previously established. Even the selection of a few markers (small x) would lead to an exorbitant number of possible combinations and would require a huge amount of time and computational resources. The proposed Probe Subset Selection must be applied in two stages: in the first, a finite set of solutions is established, and in the second stage, the recombination of the best solutions is performed to obtain a single solution.

In the first step, a number n of subsamples or "probes" - representing a sample of explanatory variables or molecular markers - is randomly removed from the original set of variables. In this study, 20,000 subsamples (probes) involving n markers that had been used in the regressions were arbitrarily considered. At the end of the first stage, an index is calculated to represent the relative importance (IR_i) of each molecular marker according to Equation (5) below:

$$IR_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \beta_{ij} R_j^2 \text{, para i = 1,2, ..., m} \qquad \text{(Eq. 5)}$$

where: n_i: number of times that a given Mi marker participated in the 20,000 polls carried out; β_{ij}: regression coefficient associated with the Mi marker included in the regression model in a given probe S_j; R_j^2: coefficient of determination obtained by the regression adjusted in the j-th probe; m: total number of markers studied (in the study, equal to 1000).

The second step consists of the recombination of the results and the selection of those with the best performance to be used in a new regression. This procedure is carried out based on the ranking provided by the indexes IR_i, which allows the selection of the n best variables for the calculation of the new multiple regression and the new coefficient of determination.

Genome Wide Selection – RRBLUP

The RR-BLUP additive model was adopted, as described by Meuwissen et al. (2001) (Equation 6):

$$y = Wb + Xm + e \qquad \text{(Eq. 6)}$$

where: y is the vector of phenotypic observations; b is the vector of fixed effects; m is the vector of random marker effects, and e refers to the vector of random errors, $e \sim N(0, I\sigma_e^2)$; W and X are matrices of incidence for b and m, respectively. Individual genomic estimated breeding values (GEBVs) were estimated by the following Equation (7).

$$GEBV = \hat{y}_j = \hat{\mu} + \sum_i x_{ij}\hat{m}_i \qquad \text{(Eq. 7)}$$

where: Xij is the line of the incidence matrix that allocates the genotype of the j-th marker for each individual (i), 1, 0, -1 for genotypes AA, Aa, aa, respectively, for biallelic and codominant markers, and $\hat{m}_i$ is the effect of the i-th marker estimated by RR-BLUP.

Multilayer Perceptron Neural Network (MLPNN)

The MLPNN architecture used was backpropagation, with three hidden layers and considering one to four neurons in each layer. The selected molecular marker matrix [M1 M2 ... Mn] was considered as input information, so that the desired output was the true genotypic value - a known value, since the populations were generated via simulation. In the output layer, MLPNN returned the predicted value for each individual ($Y_{Network}$) (Figure 1).

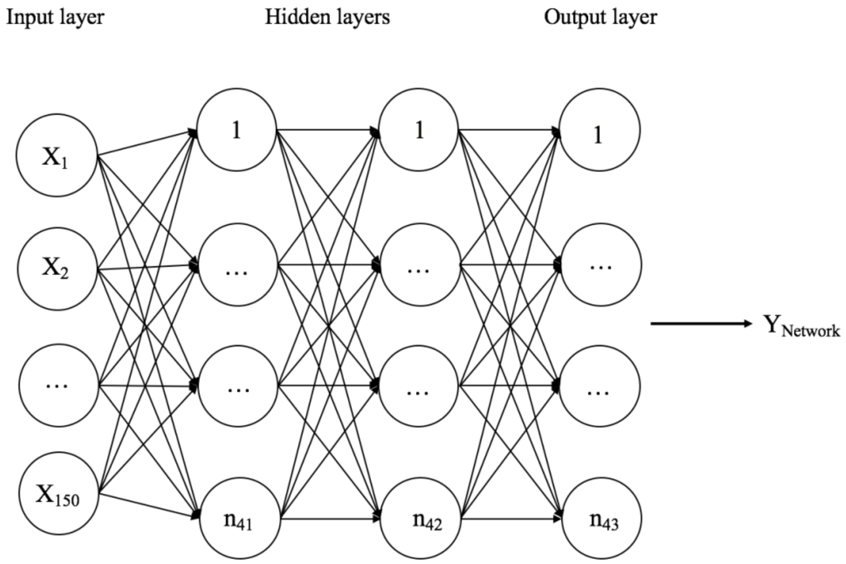

Figure 1. Multilayer Perceptron Neural Network (MLPNN) layout. Inputs X_1, ..., X_{150} in the input layer refer to the 150 markers considered in the analyses. Three hidden intermediate layers (ni1, ni2 and ni3) consisting of i neurons (i = 1, ..., 4). The MLPNN returns the vector of predicted values ($Y_{Network}$).

The logistical sigmoid (logsig) and hyperbolic tangent (tansig) were the activation functions used. The MLPNN network training process was carried out using the error backpropagation algorithm (Silva et al., 2010). A fivefold cross-validation scheme was adopted. The population of 500 individuals was randomly split into five mutually exclusive subsets, and each round four of these subsets constituted the training population (totaling 80% of individuals), while the remaining subset constituted the validation population (20% of the total population).

Efficiency evaluation

The methodologies RR-BLUP and MLPNN were compared using the predictive accuracy, expressed through the root mean square error (RMSE). Predictive accuracy is the model ability to correctly predict the expected true value. The RMSE is estimated by the following Equation (8):

$$\text{RMSE} = \sqrt{\frac{\sum(\hat{Y}-Y)^2}{n}} \qquad \text{(Eq. 8)}$$

where Y is the observed phenotypic value, and $\hat{Y}$ is the estimated phenotypic value.

Computational Resources

The population simulation was implemented with the use of the GENES software system (Cruz, 2016). The statistical analyses were performed using the R software system, with the RR-BLUP package (R core team, 2018). Both the MLPNN methodology and Probe Subset Selection Methodology were implemented using the Genes software system in combination with MATLAB (Matlab, 2010).

RESULTS AND DISCUSSION

The dimensionality was reduced as proposed by Sant'Anna et al. (2021), that is, through a graphical procedure that evaluates the values of the determination coefficient (R^2), root mean square error (RMSE) and condition number (CN) for the most complex characteristic. The results indicated that the optimal number of markers to be used was approximately 150, given the higher value of R^2, low root mean square error and CN < 100 (Figure 2). In this study, a dimensionality reduction methodology called Probe Subset Selection was proposed, whose main interest was to select the 150 markers with the greatest correlation with the characteristics of interest.

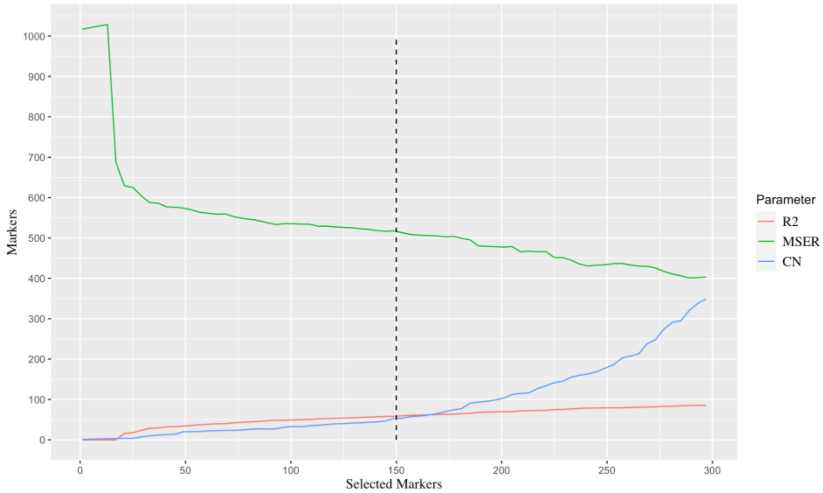

Figure 2. Graphical representation of the parameters: determination coefficient (R2) in red, root mean square error (RMSE) in green, and the condition number (CN) in blue, obtained by the Probe method, by including 1 to 300 molecular markers (from the total of 1,000) in the stepwise regression model.

Figure 3 shows the predictive accuracy (RMSE = root mean square error) obtained for the 12 scenarios evaluated, considering or not the reduction performed by the RRBLUP model. Reduced dimensionality improved the prediction accuracy (RMSE = root mean square error) for all scenarios.

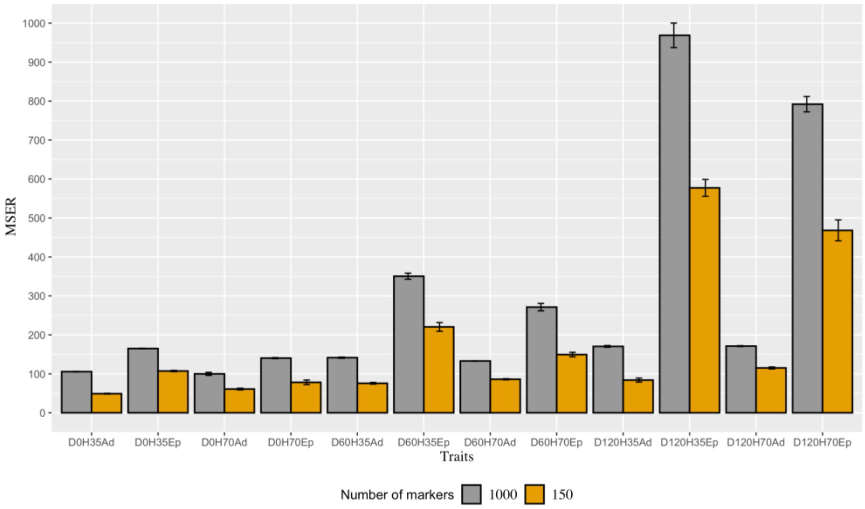

Figure 3. RMSE obtained from RR-BLUP without (1000) and with (150) reduction of the marker matrix by Probe Subset Selection Methodology, in a set of validation data involving cross-validation procedures.

Specifically for the trait D120H35Ep (T10), which is a variable considered to be highly complex due to its effects of dominance, epistasis, and low heritability, it is observed that RMSE was reduced from 960 to 570, approximately, after the application of marker selection by the Probe Subset Selection Methodology. For the trait D120H70Ep (T12), characterized by high dominance and epistasis, RMSE decreased from 790 to 470, approximately (Figure 3). Even for low complexity traits, such as D0H70Ad (T3), RMSE was reduced from 100 to 60, approximately (Figure 3). Therefore, the Probe Subset Selection Method proposed as a dimensionality reduction technique was efficient and has a promising potential for applicability aiming to improve the prediction accuracy of genomic selection methodologies and reduce the computational demand of computational intelligence techniques.

This method was built inspired by the genetic algorithm and has already been shown to be efficient in other applications that involve dimensionality reduction. Raymer et al. (2000) adopted the genetic algorithm in combination with the nearest neighbor classification rule and compared the results with classical feature selection and extraction techniques, and with the results obtained they were able to identify favorable water binding sites on protein surfaces.

Problems related to dimensionality in genomic selection studies have been reported by other authors (Azevedo et al., 2014; Sant'anna et al., 2021). Azevedo et al. (2014) proposed the use of dimensionality reduction methods to overcome the genomic selection to predict genomic breeding values for carcass traits in pigs. James et al. (2013) discussed the problems arising from the high dimensionalities considered in wide genomic selection - such as variance bias, overfitting and multicollinearity - and highlighted the possibility of using reduction and selection procedures for subsamples or penalized methods, such as Stepwise Regression, Partial Principal Components, Partial Minimum Squares and Lasso Bayesiano as alternatives. Sant'anna et al. (2021) showed that the use of stepwise regression before the use of these techniques led to an improvement in the accuracy of prediction of the genetic value, facilitating processing and analysis time due to a reduction in dimensionality.

The comparison between the RRBLUP and MLPNN prediction methods was performed only after using the proposed dimensionality reduction method. It is important to note that both methods used the same 150 markers. Several authors have already pointed out that methodologies based on computational intelligence require a lot of time and super computers. They have also found that methods for reducing dimensionality help to increase the efficiency of these methods (Hinton and Salakhutdinov, 2006; Azevedo et al., 2014; Sant'Anna et al., 2021). Sant'Anna et al., 2021 used dimensionality reduction methodologies and proved that such methods improve the efficiency of RRBLUP models and RBFNN networks and reduce the time of the analysis in a normal computer from 20h to less than one hour, after dimensionality reduction.

Table 2 presents a comparison between the RR-BLUP and MLPNN methodologies, after the dimensionality reduction performed by the Probe Subset Selection Methodology. As observed in Table 2, the prediction of genomic genetic values based on MLPNN outperformed RR-BLUP for all scenarios evaluated. Specifically, for T10 - D120H35Ep, for example, RMSE decreased from 577.249 to 24.483.

Table 2. RMSE obtained from RR-BLUP and MLPNN through selected markers (150) by Probe Subset Selection in a set of validation data involving cross-validation procedures.

	RMSE	
Traits	**150 RR-BLUP**	**150 MLPNN**
T1 - D0H35Ad	49.092 ± 2.294	4.696 ± 0.044
T2 - D0H35Ep	107.264 ± 3.601	10.718 ± 0.186
T3 - D0H70Ad	60.946 ± 5.060	3.327 ± 0.145
T4 - D0H70Ep	78.218 ± 13.380	9.044 ± 0.800
T5 - D60H35Ad	75.732 ± 4.872	5.039 ± 0.068
T6 - D60H35Ep	220.485 ± 24.846	15.105 ± 0.594
T7 - D60H70Ad	86.116 ± 3.249	3.729 ± 0.063
T8 - D60H70Ep	149.463 ± 12.954	12.425 ± 0.281
T9 - D120H35Ad	83.985 ± 11.687	6.195 ± 0.473
T10 - D120H35Ep	577.249 ± 48.375	24.483 ± 0.740
T11 - D120H70Ad	115.167 ± 5.722	4.512 ± 0,134
T12 - D120H70Ep	468.195 ± 59.989	19.697 ± 0.354

Other authors had already reported the superiority of neural networks in prediction studies (Sant'anna et al., 2015; Silva et al., 2016; Silva et al., 2017; Sant'anna et al., 2019; Sant'anna et al., 2021). Sant'anna et al. (2019) evaluated the genome-enable prediction by Radial Basis Neural Networks (RBFNN) model compared to RRBLUP and obtained greater prediction accuracy for the RBFNN model. For traits with $h^2 = 35\%$, RMSE decreased from 49.092 to 4.696, by adopting MLPNN over RR-BLUP, in the scenario with additive effects (T1). For traits with $h^2 = 70\%$, RMSE decreased from 60.946 to 3.327, by adopting MLPNN over RR-BLUP, in the scenario with additive effects (T3) (Table 2).

In general, the MLPNN method is less affected by the different genetic architectures evaluated (Figure 4). Figure 4a shows that, with the inclusion of dominance effects, RMSE increased, on average, from 74 to 311, approximately, using RR-BLUP, and from 7 to 14, approximately, using MLPNN, in agreement with the results obtained by other authors (Sant'anna et al., 2019; Sant'anna et al., 2021). Sant'Anna et al. (2019) carried out studies using simulated populations with different levels of dominance and heritability and verified the predictive superiority of the RBF neural networks when compared with the G-BLUP method, given the lower RMSE values obtained.

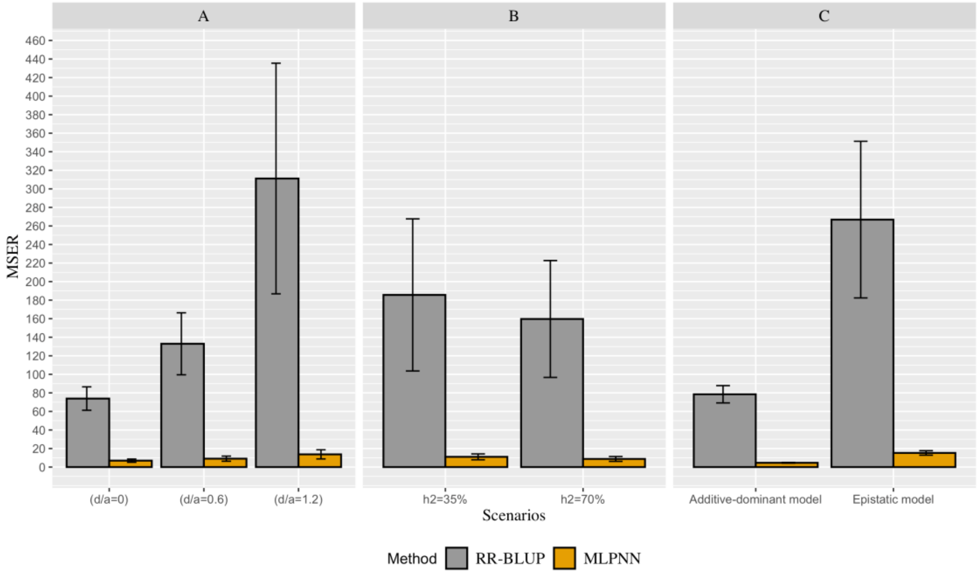

Figure 4. RMSE obtained from MLPNN and RR-BLUP through selected markers (100) by Probe Subset Selection in a set of validation data involving cross-validation procedures, considering: (a) dominance levels - no dominance (d/a = 0), moderate dominance (d/a = 0.6) and high dominance (d/a = 1.2); (b) Heritability – h2 = 35% and h2 = 70%; (c) Genetic model – additive-dominant and epistatic model.

Greater impact on RMSE is observed with the inclusion of dominance effects in the control of low-heritability traits. For the traits T1-D0H35Ad, T5-D60H35Ad and T9-D120H35Ad, the prediction accuracy ranged from 4.696 to 5.039 and then, to 6.165 with MLPNN. It ranged from 49.092 to 75.732 and then, to 83.985 with RR-BLUP, by including dominance in the scenario with additive effects (Table 2).

The study of heritability is very important for the success of breeding programs, as it assists in determining the genetic variation of individuals and its effect on segregating generations, which allows identifying, from the phenotypic values, the individuals with desirable genotypic values and the highest concentration of favorable alleles (Cruz, 2012). However, difficulties in predicting low heritability characters have been reported by several authors (Goddard, 2009; Hayes et al., 2009; Cruz, 2012; Almeida Filho et al., 2016; Glória et al., 2016). Almeida Filho et al. (2016) carried out preliminary studies to detect the contribution of dominance in phenotypic prediction in pine breeding in simulated populations. They also obtained lower accuracy for low heritability traits when dominance effects were included and concluded that dominance reduces the overall precision of prediction models.

It is worth mentioning that, although affected by low heritabilities, MLPNN networks were more accurate than RR-BLUP. In the scenarios with higher heritability, the

RMSE average changed from 11 to 9 with the use of MLPNN, and from 185 to 160, approximately, when using RR-BLUP (Figure 4b). Artificial neural networks - through their networks of artificial neurons, activation functions and learning algorithms - are believed to be able to capture the effects of disturbing factors neglected by other methodologies and thus provide more efficient results (Gianola et al., 2011; Gonzalez-Camacho et al., 2012; Nascimento et al., 2013; Bhering et al., 2015; Silva et al., 2016).

Sant'Anna et al. (2021) studied simulated populations with heritabilities of 30% and 60% and obtained inferior predictive accuracy for lower heritability traits by including dominance effects and demonstrated that Radial Basis Function neural networks outperformed RR-BLUP. Glória et al. (2016) studied the effects of markers and heritability estimates based on genome prediction through Bayesian regularized neural networks and concluded that neural networks are promising quantitative tools for genomic prediction studies.

As dominance effects, the epistatic interactions also hinder the practice of breeding and selection, since superior phenotypes can be attributed to populations with a high number of genetically distinct individuals (Vencovsky, 1973). The inclusion of epistatic effects has penalized the predictive accuracy of both methodologies by increasing the average of RMSE values from 5 to 15, approximately, using MLPNN, and from 80 to 265, approximately, using RR-BLUP (Figure 4c). Specifically, for T1-D0H35Ad, the RMSE value was 4.696 and, with the inclusion of epistatic effects (T2-D0H35Ep), the RMSE value increased to 10.718 with MLPNN and from 49.092 to 107.264 when using RR-BLUP, respectively (Table 2).

For the most complex traits, the predictive accuracy declined even more. For T9-D1.2H35Ad, the RMSE value was 6.195 and, when including epistatic effects (T10-D1.2H35Ep), the RMSE value increased to 24.483 with MLPNN, and from 83.985 to 577.249 when using RR-BLUP, respectively (Table 2). Almeida Filho et al. (2016) highlighted the difficulties of modeling epistasis effects and that such effects could be disturbing and affect the accuracy of the models.

Such results demonstrate that the MLPNN neural network was able to capture the effects of epistasis on the characteristics of interest in this study, through its neuron networks, unlike RR-BLUP, which neglects such effects, since its model only involves the term, which in its equation refers to the area of incidence of the dose effects of the studied markers rather than the interaction between them (Resende et al., 2014). In computational intelligence techniques, the inputs are also represented by the information of X_m markers. However, the hidden layers used become indispensable to capture effects, in addition to those related to the additive action of the information (Haykin, 1999; Moore, 2006).

Gonzalez-Camacho et al. (2012) carried out studies with simulated data and found that the computational intelligence method of RBF Neural Networks can capture epistatic effects. Beam et al. (2014) used Bayesian Neural Networks to detect epistasis in genetic association studies for several simulated scenarios and concluded that neural networks are a powerful technique for association studies, with the ability to capture epistatic effects. Barbosa et al. (2021) demonstrated that Multilayer Perceptron are efficient for predicting genetic values in the presence of additive-dominant and epistatic gene control in simulated populations presenting different levels of heritability.

CONCLUSIONS

The use of the proposed Probe subset selection methodology for selecting molecular markers in this study proved to be an effective strategy to improve the prediction accuracy of RR-BLUP. By adopting the selected markers, MLPNN presented predictive accuracy, expressed by mean square error, higher than that presented by RR-BLUP and successfully allowed incorporation of additive, additive-dominant, and epistatic effects into prediction models.

ACKNOWLEDGMENTS

This study is a tribute (in memorian) to Professor Fabyano Fonseca e Silva, for his brilliant contribution in spreading the Genomic prediction methodologies. His passion for science, genetics, plant and animal breeding, statistics, and his enthusiasm to propose new ideas and the search for excellence impacted the scientific career of many students. We acknowledge Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – financing code 001 – for their financial support.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Almeida Filho JE, Guimarães, JFR, Silva FF, Resende MDV, et al. (2016). The contribution of dominance to phenotype prediction in a pine breeding and simulated population. *Heredity*. 117: 33-41.

Azevedo CF, Resende MDV de, Silva FF, Viana JMS, et al. (2015). Ridge, Lasso and Bayesian additive-dominance genomic models. *BMC Genet*. 16: 105.

Azevedo CF, Silva FF, Resende MDV, Lopes MS, et al. (2014). Supervised independent component analysis as an alternative method for genomic selection in pigs. *J. Anim. Breed. Genet*. 131: 452-461.

Beam AL, Motsinger-Reif A and Doyle J (2014). Bayesian neural networks for detecting epistasis in genetic association studies. *BMC Bioinf*. 15: 368.

Bhering LL, Cruz CD, Peixoto LA, Rosado AM, et al. (2015). Application of neural networks to predict volume in eucalyptus. *Crop Breed. Appl. Biotechnol*. 15: 25-131.

Cruz CD and Nascimento M (2018). Inteligência computacional aplicada ao melhoramento genético. 1st edn. Editora UFV. Viçosa.

Cruz CD (2016). Genes Software – extended and integrated with the R, Matlab and Selegen. *Acta Sci. Agron*. 38: 547-552.

Cruz CD (2012). Princípios de genética quantitativa. Editora UFV, Viçosa.

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

Goldberg DE (1989). Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, New York.

Glória LS, Cruz CD, Vieira RAM, Resende MDV, et al. (2016). Accessing marker effects and heritability estimates from genome prediction by Bayesian regularized neural networks. *Livest. Sci*. 191: 91-96.

Gonzalez-Camacho JM, De Los Campos G, Pérez P, Gianola D, et al. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor. Appl. Genet*. 125: 759-771.

Haykin S (1999). Neural networks: A Comprehensive Foundation. Prentice-Hall. Englewood Cliffs, New Jersey.

Hinton GE and Salakhutdinov RR (2006). Reducing the dimensionality of data with neural networks. *Science*. 313(5786): 504-507

James G, Witten D, Hastie T and Tibshirani R (2013). An Introduction to Statistical Learning with Applications in R. Springer, Texas.

Long N, Gianola D, Rosa GJ and Weigel KA (2011). Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. *J. Ani. Breed. Genet*. 128: 247-57.

Matlab (2010). Matlab Version 7.10.0. Natick, Massachusetts: The Math Works Inc.

Moore JH, Gilbert JC, Tsai C, Chiang F, et al. (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J. Theor. Biol.* 241: 252-261.

Nascimento M, Peternelli LA, Cruz CD, Nascimento ACC, et al. (2013). Artificial neural networks for adaptability and stability evaluation in alfalfa genotypes. *Crop Breed. Appl. Biotechnol.* 13: 152-156.

Raymer ML, Punch WF, Goodman ED, Kuhn LA, et al. (2000). Dimensionality Reduction Using Genetic Algorithms. *IEEE Trans. Evol. Comput.* 4 (2): 164-171.

Resende MDV, Silva FF and Azevedo CF (2014). Estatística matemática, biométrica e computacional: Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição Sobrevivência. Editora Suprema, Viçosa, Minas Gerais.

Sant'anna IC, Silva GN, Nascimento M and Cruz CD (2021). Subset selection of markers for the genome-enabled prediction of genetic values using radial basis function neural networks. *Acta Sci. Agron*. 43: e46307.

Sant'anna IC, Nascimento M, Silva GN, Cruz CD, et al. (2019). Genome-enabled prediction of genetic values for using Radial Basis Function Neural Networks. *Func. Plant Breed. J.* 1: 1-8.

Sant'anna IC, Tomaz RS, Silva GN, Nascimento M, et al. (2015). Superiority of artificial neural networks for a genetic classification procedure. *Genet. Mol. Res*. 14: 9898-9906.

Silva GN, Nascimento M, Sant'anna IC, Cruz CD, et al. (2017). Artificial neural networks compared with Bayesian generalized linear regression for leaf rust resistance prediction in *Arabica coffee*. *Pesqui. Agropecu. Bras.* 52: 186-193.

Silva GN, Tomaz RS, Sant'anna IC, Carneiro VQ, et al. (2016). Evaluation of the efficiency of artificial neural networks for genetic value prediction. *Genet. Mol. Res*. 15: gmr.15017676.

Silva GN, Tomaz RS, Sant'anna IC, Nascimento M, et al. (2015). Neural networks for predicting breeding values and genetic gains. *Sci. Agric.* 71: 494-498.

Silva IN, Spatti HD and Flauzino RA (2010). Redes Neurais Artificiais: para engenharia e ciências aplicadas. Editora Artliber, São Paulo.

Vencovsky R (1973). Princípios de genética quantitativa. Editora: ESALQ, Piracicaba, São Paulo.