Acta Scientiarum
http://www.periodicos.uem.br/ojs/
ISSN on-line: 1807-8621
Doi: 10.4025/actasciagron.v47i1.69089

Two-step genomic prediction using artificial neural networks - an effective strategy for reducing computational costs and increasing prediction accuracy

Maurício de Oliveira Celeri[1], Cynthia Aparecida Valiati Barreto[1], Wagner Faria Barbosa[1], Leísa Pires Lima[2], Lucas Souza da Silveira[1], Ana Carolina Campana Nascimento[1], Moysés Nascimento[1*] and Camila Ferreira Azevedo[1]

[1]Departamento de Estatística, Universidade Federal de Viçosa, Campus Universitário, Av. Peter Henry Rolfs, s/n, 36570-900, Viçosa, Minas Gerais, Brazil. [2]Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais, Campus Muriaé, Muriaé, Minas Gerais, Brazil. *Author for correspondence. E-mail: moysesnascim@ufv.br

ABSTRACT. Artificial neural networks (ANNs) are powerful nonparametric tools for estimating genomic breeding values (GEBVs) in genetic breeding. One significant advantage of ANNs is their ability to make predictions without requiring prior assumptions about data distribution or the relationship between genotype and phenotype. However, ANNs come with a high computational cost, and their predictions may be underestimated when including all molecular markers. This study proposes a two-step genomic prediction procedure using ANNs to address these challenges. Initially, molecular markers were selected either directly through Multivariate Adaptive Regression Splines (MARS) or indirectly based on their importance, identified through Boosting, considering the top 5, 20, and 50% of markers with the highest significance. Subsequently, the selected markers were employed for genomic prediction using ANNs. This approach was applied to two simulated traits: one with ten trait-controlling loci and heritability of 0.4 (Scenario SC1) and the other with 100 trait-controlling loci and a heritability of 0.2 (Scenario SC2). Comparisons were made between ANN predictions using marker selection and those without any marker selection. Reducing the number of markers proved to be an efficient strategy, resulting in improved accuracy, reduced mean squared error (MSE), and shorter adjustment times. The best ANN predictions were obtained with ten markers selected by MARS in SC1, and the top 5% most relevant markers selected using Boosting in SC2. As a result, in SC1, predictions using MARS achieved over a 31% increase in accuracy and a 90% reduction in MSE. In SC2, predictions using Boosting resulted in more than a 15% increase in accuracy and an 83% reduction in MSE. For both scenarios, computational time was up to ten times shorter with marker selection. Overall, the two-step prediction procedure emerged as an effective strategy for enhancing the computational and predictive performance of ANN models.

Keywords: multivariate adaptive regression splines; boosting; artificial neural network; genetic breeding.

Received on July 28, 2023.
Accepted on December 11, 2023.

Introduction

Advancements in sequencing and genotyping techniques have enabled the direct use of DNA information for the selection of genetically superior individuals. Within this context, Genome-Wide Selection (GWS), as initially proposed by Meuwissen, Hayes, and Goddard (2001), has demonstrated exceptional efficiency in predicting genomic breeding values (GEBVs). This methodology facilitates early individual selection, ultimately enhancing genetic gain per unit of time (Crossa et al., 2017; Voss-Fels, Cooper, & Hayes, 2019). GWS methods can analyze and estimate the effects of thousands of Single Nucleotide Polymorphism (SNP) markers on phenotypes. However, only markers in linkage disequilibrium (LD) with quantitative trait loci (QTLs) are relevant for predicting GEBVs and explaining genetic variations in traits of interest.

The primary statistical methods employed in GWS typically assume distributions for marker effects in the models or rely on implicit regressions or dimensionality reduction techniques (Resende, Silva, & Azevedo, 2014). Nonetheless, there is a growing interest in nonparametric methods, including Artificial Neural Networks (ANNs), for handling GWS data. Specifically, the Multilayer Perceptron Networks (MLP) class of

ANNs has shown promising results in genetic breeding studies conducted by researchers (Crossa et al., 2017; Cruz & Nascimento, 2018; Rosado et al., 2020; Sousa et al., 2022). ANNs operate akin to biological neurons, acquiring knowledge through experiences and capable of capturing all available information, including complex feature architectures, to generate decision-making criteria (Long et al., 2010; Howard, Carriquiry, & Beavis, 2014; Cruz & Nascimento, 2018; Montesinos López, Montesinos López, & Crossa, 2022; Xu et al., 2022).

Despite the potential of ANN procedures in genomic analysis, numerous markers can limit their applicability. High-dimensional scenarios often result in substantial computational costs and challenges in the learning process since many resources (markers) represent unnecessary aspects of the ANN search space (Long, Gianola, Rosa, & Weigel, 2011; Ehret, Hochstuhl, Gianola, & Thaller, 2015). In other words, molecular markers not in LD with the QTLs related to the trait of interest can introduce unnecessary complexity when applying ANNs. However, selecting a subset of markers potentially associated with the traits of interest can mitigate the high-dimensionality problem, leading to improved learning and enhanced predictive power of ANN models (Long et al., 2010; Crossa et al., 2017; Sant'Anna et al., 2020a; Silva et al., 2022; Aono et al., 2022).

Among the proposed methods for dimensionality reduction, those based on principal components and independent components merit mention (Azevedo, Resende, Silva, Lopes, & Guimarães, 2013; James, Witten, Hatie, & Tibshirani, 2013; Long et al., 2010; Resende et al., 2014; Costa et al., 2020; Paixão et al., 2022; Fialho et al., 2023).

Additionally, methods for direct selection of variables, such as Bayes B, Bayes Cπ, and Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991; Huang et al., 2020; Costa et al., 2022), and indirect selection through variable importance, such as Boosting and Random Forest (Ho, Schierding, Wake, Saffery, & O'Sullivan, 2019; Sousa et al., 2022), have been proposed. Marker selection has already demonstrated its effectiveness in genomic prediction, as demonstrated by Sant'Anna et al. (2020a), who employed stepwise regression, and Sousa et al. (2022), who used indirect selection via Bagging.

Based on the considerations outlined above, this study aims to assess the accuracy, mean squared error (MSE), and computational effort of a two-step procedure for genomic prediction. Initially, this approach involves either the direct selection of a group of SNPs using Multivariate Adaptive Regression Splines (MARS) or the indirect selection of different sets of the most relevant SNPs identified by Boosting. Subsequently, these selected markers are used for genomic prediction using ANNs. Predictions are made using simulated data, considering two distinct scenarios (SC1 and SC2). In SC1, ten loci control the trait with a heritability of 0.4, while in SC2, 100 loci control the trait with a heritability of 0.2.

Material and methods

Data simulation

The data set was simulated using the GENES software, as described by Cruz (2013). The genome comprised ten linkage groups, each spanning 20 cM and containing 200 SPNs. Thus, approximately 2000 SNPs were distributed equidistantly across the genome to ensure comprehensive coverage. For genetic linkage analysis, an F1 generation was simulated, consisting of one homozygous dominant parent and one homozygous recessive parent. The genotypes of the F1 population were used to generate an F2 mapping population comprising 1,000 individuals. Two traits were simulated, both with a zero degree of dominance (*d/a* = 0), where "a" and "d" represent the genotypic values of homozygotes and heterozygotes, respectively, and a mean value of 100. However, different narrow-sense heritabilities (h_a^2) and genetic architectures were combined to create scenarios SC1 and SC2 (Table 1). In SC1, the trait was simulated with a moderate heritability (h_a^2 = 0.40) and controlled by ten loci, with each chromosome containing a single QTL. In SC2, the trait had a lower heritability (h_a^2 = 0.20), with each chromosome harboring 10 QTLs. Thus, two genetic architectures, explaining equal parts of the genetic variance, were created, with the QTLs distributed within the regions covered by SPNs. Additionally, the proportion of genetic variation associated with the QTLs explained by the markers (r_{mq}^2) was calculated according to Goddard, Hayes, and Meuwissen (2011), as follows:

$$r_{mq}^2 = \frac{n}{n + n_{QTL}}$$

where n is the number of SNPs and n_{QTL} is the number of QTL.

Table 1. Description of the scenarios with respective genetic architectures, number of QTLs, and narrow-sense heritabilities.

Scenarios	r_{mq}^2	Genetic architecture	QTLs	Heritability (h_a^2)
SC1	0.99	1 QTL on each of the ten chromosomes	10	0.4
SC2	0.95	10 QTL on each of the ten chromosomes	100	0.2

Marker selection using multivariate adaptive regression splines

Multivariate Adaptive Regression Splines (MARS) is a nonparametric regression method used for modeling the relationship between predictive and dependent variables via basis functions, primarily in high-dimensional problems (Friedman, 1991; Huang et al., 2020).

The model proposed by Friedman (1991) is:

$$f(X) = c_0 + \sum_{i=1}^{M} c_i B_i(X) + \varepsilon,$$

where c_0 is the intercept, $B_i(X)$ is a basis function, c_i is the coefficient of B_i and M is the number of basis functions automatically defined by the MARS algorithm (Abdulelah Al-Sudani et al., 2019), and ε is the random error. The coefficients c_i $(i = 0, 1, \dots, M)$ are estimated via minimization of the residual sum of squares (RSS) (Hastie, Tibshirani, & Friedman, 2009).

MARS involves two phases: forward and backward. In the forward phase, the model incorporates pairs of basic functions that minimize the Residual Sum of Squares (RSS) (Hastie et al., 2009; Park & Kim, 2018). In the backward phase, certain basis functions are excluded to create a more parsimonious model that avoids overfitting. The MARS algorithm can select variables during model building, and this approach was employed to select SPNs, which were then used as inputs for ANN training.

Marker selection using boosting

As described by James et al. (2013), the Boosting method is a statistical learning approach that employs regression trees to adjust the residuals of an initial model. The residuals are updated in each tree, grown sequentially from the previous tree, and the response variable involves a combination of many trees, such that $\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$. The function $\hat{f}(.)$ refers to the final tree resulting from the sequential combination of the previously adjusted trees. The parameter λ is the shrinkage that controls the learning rate of the method. This method also needs to be adjusted with several splits in each tree. This parameter controls the complexity of the boosting and is known as *depth*.

In this method, the importance of variables, as described by Friedman (2001), can be used to indirectly select sets of the most relevant SPNs. The sets of markers comprised 5, 20, and 50% of the markers, equating to 100, 400, and 1,000 markers used, respectively.

Artificial neural network (ANN)

In this study, the Multilayer Perceptron Network (MLP), initially proposed by Rosenblatt (1958), was employed as the ANN class. The error backpropagation learning algorithm was chosen for adjusting weights during ANN training. The network topology varied, featuring a single hidden layer with one to 30 neurons. The logistic function served as the activation function, and the identity function was used as the output function. The training phase encompassed 1,000 iterations.

The output of the hidden layer is computed by taking a linear combination of the I input variables and then applying the logistic activation function, φ(.). Specifically, the output of the k^{th} neuron in the layer is determined as follows:

$$W_{1k} = \varphi\left(w_{0k} b_1 + \sum_{i=1}^{I} w_{ik} M_i\right), k = 1, \cdots, K,$$

where K is the number of neurons in the hidden layer, w_{ik} is the synaptic weight between the *i*th marker and the *k*th neuron, and w_{0k} is the synaptic weight of the bias b_1 for the *k*th neuron.

The ANN output is given as follows:

$$Y_N = w_0'' b_2 + \sum_{i=1}^{J} W_{1i} w_i'',$$

where w_i'' is the synaptic weight between the *i*th neuron of the hidden layer and the network output layer, and w_0'' is the synaptic weight of the bias b_2.

Cross-validation

The performance of ANN models was evaluated through a k-fold cross-validation process, dividing the population (1000 observations) into five groups, each comprising 200 observations. In each iteration, one group served as the validation population, while the remaining four constituted the training population. This process was repeated five times, ensuring all individuals participated in the validation population exactly once.

Two-step genomic prediction and comparison of marker selection approaches

Initially, the training sets (800 observations) underwent direct marker selection via MARS and indirect marker selection via Boosting in each fold. Subsequently, the selected markers were used as input variables for ANN. Accuracy and mean square error (MSE) were measured using the validation set (200 observations), except for time, which was recorded during the training phase. Accuracy was determined using the Pearson correlation between observed GEBVs and predicted GEBVs, while MSE was calculated based on the mean squared difference between observed and predicted GEBVs (Figure 1).

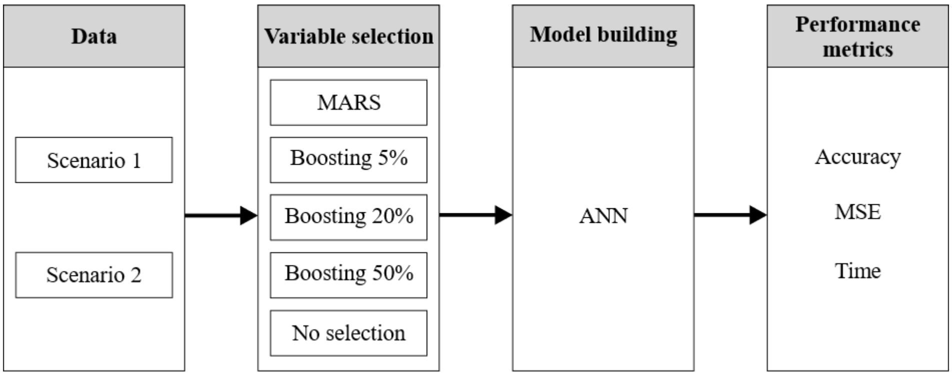

Figure 1. Two-step genomic prediction scheme.

The analyses were conducted using the R software (R Core Team, 2022).

For MARS adjustment, the "*earth*" function from the "*earth*" package was employed. For Boosting, the "*gbm*" function from the "*gbm*" package was used. Finally, the ANN analysis was performed using the "*mlp*" function from the "*RSNNS*."

Comparison of GWS methodologies

Following the k-fold process and obtaining performance metrics, the network topology with the highest accuracy value (i.e., a network model with the number of neurons ranging from 1 to 30 in the hidden layer) was selected to represent the network with the best predictive capacity within each k-fold iteration. Accuracy, MSE, and time values were summarized by calculating their mean and standard error (SE) to compare the performance of different two-step genomic prediction methods to ANNs used without prior marker selection.

Results

Figures 2, 3, and 4 illustrate the accuracy, MSE, and time required for accessing the results for both scenarios, SC1 and SC2. During the MARS fitting, a fixed number of ten markers was consistently selected, irrespective of the scenario or the training set used in the k-fold process. Consequently, only ten markers out of the total 2,000 were used for the subsequent training and validation of the ANN models.

In SC1, a reduction in accuracy was observed as the number of selected markers increased. In this scenario, the highest accuracy (0.96 ± 0.00) was achieved with markers directly selected via MARS (10 markers), representing a substantial 30% increase compared to the accuracy (0.73 ± 0.01) obtained with no marker selection (Figure 2).

Similarly, in SC2, a pattern analogous to SC1 was observed. However, the highest accuracy (0.98 ± 0.00) was attained when the top 5% most important markers (100 markers) were selected by Boosting, leading to a notable 15% increase in accuracy relative to the accuracy (0.85 ± 0.01) obtained with no marker selection (Figure 2).

In both scenarios, marker selection resulted in a sparse set of markers covering a significant portion of the genome. Supplementary Figures S1 and S2 present the positions of markers selected by the best methodology for SC1 and SC2, along with the corresponding QTLs. The supplementary material is available for download at https://www.licae.ufv.br/pesquisa/.

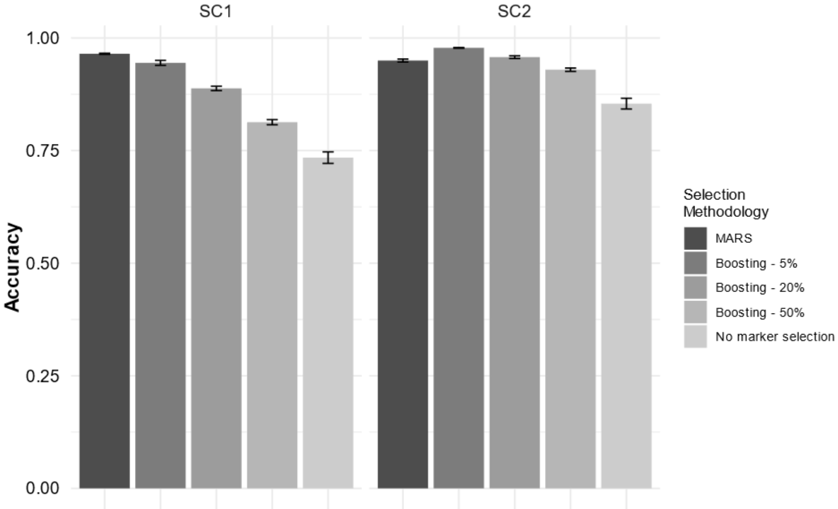

Figure 2. Accuracy (± SE) obtained in the validation by the ANN models adjusted with different sets of markers, either directly selected via MARS or indirectly via Boosting using the top 5, 20, and 50% most important markers. An ANN model adjusted with all 2,000 markers (no marker selection) is included for comparison. SC1: 1 QTL on each of the ten chromosomes; SC2: 10 QTL on each of the ten chromosomes.

As expected, predictions obtained with no marker selection exhibited the highest MSE values in both scenarios (Figure 3). However, in SC1, the ANN predictions achieved the lowest MSE when MARS selected the markers (5.33 ± 0.34), representing a substantial 90% reduction in MSE compared to no marker selection. In SC2, the top 5% most important markers identified by Boosting yielded the lowest MSE (11.65 ± 1.29), resulting in an approximately 83% decrease in MSE compared to no marker selection (Figure 3).

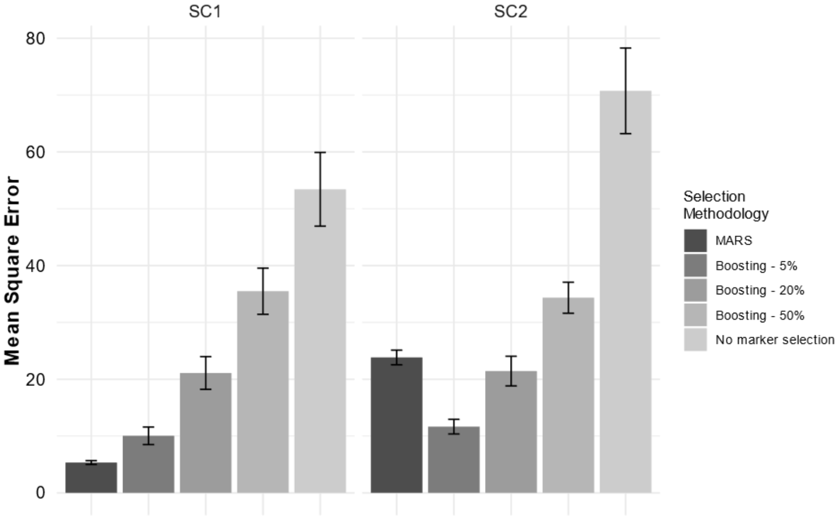

Figure 3. Mean square error (± SE) obtained in the validation phase by the ANN models adjusted with different sets of markers, either directly selected via MARS or indirectly via Boosting using the top 5, 20, and 50% most important markers. An ANN model adjusted with all 2,000 markers (no marker selection) is included for comparison. SC1: 1 QTL on each of the ten chromosomes; SC2: 10 QTL on each of the ten chromosomes.

In SC1, the computational time for training the ANN varied from 2.38s (with MARS-selected markers) to 27.54s (with no marker selection). In SC2, the range was 3.55s (with the top 5% most important markers selected via Boosting) to 38.59s (with no marker selection). As expected, computational time was longer without marker selection and was generally slightly higher in SC2 compared to SC1 (Figure 4).

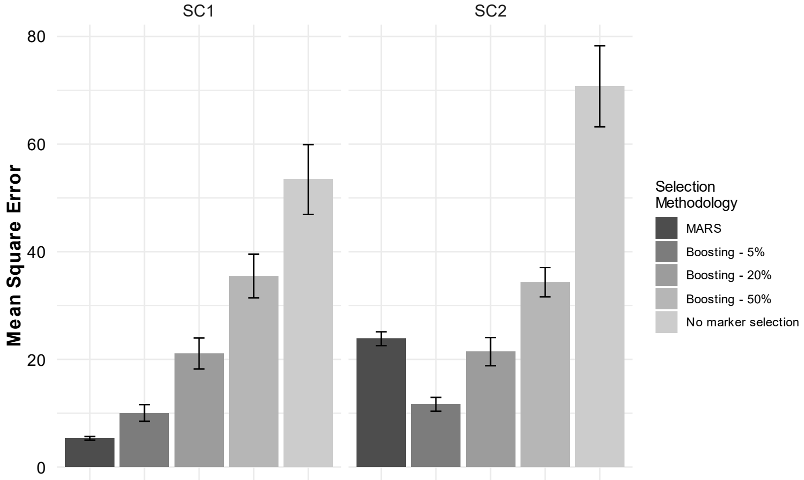

Figure 4. Mean square error (± SE) obtained in the training phase of the ANN models adjusted with different sets of markers, either directly selected via MARS or indirectly via Boosting using the top 5, 20, and 50% most important markers. An ANN model adjusted with all 2,000 markers (no marker selection) is included for comparison. SC1: 1 QTL on each of the ten chromosomes; SC2: 10 QTL on each of the ten chromosomes.

Discussion

This study proposed a two-step genomic prediction procedure that combines either MARS or Boosting with an ANN to enhance computational efficiency and predictive capacity. Initially, MARS and Boosting were employed for marker selection in two scenarios (SC1 and SC2). Scenario SC1 involved a trait with $h_a^2 = 0.4$ and ten trait-controlling loci, while scenario SC2 involved a trait with $h_a^2 = 0.2$ and 100 trait-controlling loci. Subsequently, the selected markers were used as input for the ANN to establish the genomic prediction model.

While ANNs are favored in genetic breeding for their flexibility and lack of *a priori* assumptions about the genotype-phenotype relationship (Glória et al., 2016), they are disadvantaged by their high computational cost compared to traditional genomic prediction methods. For instance, ANNs can be up to nine times slower than GBLASSO (Sousa et al., 2020).

A potential solution to overcome this limitation is to employ the proposed two-step genomic prediction approach, as ANNs alone do not perform marker selection. Marker selection is crucial to reduce the impact of non-informative markers and improve the accuracy of genomic prediction. Additionally, marker selection supports the assumption that not all markers in the genome are in linkage disequilibrium (LD) with the QTLs.

MARS was selected for marker selection in this study due to its nonparametric regression capabilities, which have not been extensively utilized in genetic breeding studies. Nevertheless, MARS offers several advantages, including flexibility in model adjustment and the ability to detect important variables in high-dimensional scenarios (Zabihi, Pourghasemi, Motevalli, & Zakeri, 2019; Nayana, Kumar, & Chesneau, 2022). This potential has also made MARS highly promising for predicting and selecting genomic markers, outperforming conventional statistical models such as GBLUP (Costa et al., 2022). Furthermore, Boosting, an ensemble learning method, has demonstrated success in genetic breeding studies (Silveira, Lima, Nascimento, Nascimento, & Silva, 2020; Sousa et al., 2020; Westhues et al., 2021). It is particularly indicated for regression problems due to its high predictive capabilities, surpassing methods like Random Forest and Support Vector Machine. Additionally, it incorporates automatic indirect marker selection (Ogutu, Piepho, & Schulz-Streeck, 2011).

This study demonstrates the efficiency of marker selection for genomic prediction in both evaluated scenarios (SC1 and SC2). Compared to using ANN without marker selection, a significant improvement of over 30% in predictive capacity was observed when using markers selected with MARS in SC1. Similarly, in SC2, approximately a 15% improvement in predictive capacity was achieved when the top 5% most important markers were indirectly selected via Boosting. It is noteworthy that MARS and Boosting (5%) selected ten markers in SC1 and 100 markers in SC2, which align with the QTL numbers in these scenarios. These results suggest that MARS is an intriguing method for selecting markers in oligogenic genetic architectures. Conversely, the Boosting approach performs better in the SC2 scenario, which comprises 100 QTLs.

The markers selected by MARS and Boosting, as displayed in Supplementary Figures S1 and S2, underscore the effectiveness of using only a few markers to predict GEBVs. Markers with high LD among themselves represent redundant information and may be discarded. Therefore, incorporating information into the model requires only one marker from those with high LD. This implies that high-density panels may not necessarily improve predictive capacity (Song & Hu, 2022).

Sousa et al. (2022) found consistent results with this study when assessing the effect of marker selection on genomic prediction in *Coffea canephora*. Li et al. (2018) employed machine learning methods to select 3000 markers, achieving a predictive capacity of 0.41 using Random Forest for marker selection and 0.46 using Gradient Boosting Machine (GBM), compared to 0.43 using the entire available panel of SPN markers (38,082). These researchers observed that the predictive capacity increased as the number of markers decreased, reaching optimal values with the use of 3.4% to 6.9% of the total number of available markers.

Sant'Anna et al. (2020a) demonstrated that by using simulated data and selecting markers via stepwise regression, the radial basis function network had improved reliability in predicting and reduced the ANN fitting time by up to 20 times. Additionally, the efficiency of MARS for selecting variables to serve as input in ANNs has been demonstrated by Kao and Chiu (2020). These researchers found that a network adjusted with prior selected variables can be up to three times faster than a network without variable selection while producing more accurate results.

Additionally, it was shown that marker selection significantly reduced the computational time required for fitting the ANN. The time to adjust the ANN for prediction was 11.57 times smaller for the two-step prediction using marker selection by MARS in SC1. While in SC2, the time was up to 10.87 times faster for ANN with the selection of the top 5% most important markers using Boosting. Moreover, an increased time required to fit the ANN model was observed as more markers were included as input variables. For instance, with the top 50% most important markers via Boosting, the time spent to fit the ANN model was only 1.03 times faster than the ANN fitting without marker selection in SC1 and 1.79 times faster than the ANN fitting without marker selection in SC2.

This study supports the hypothesis that marker selection efficiently reduces computational costs while increasing predictive accuracy. The explanation lies in the reduction of the marker pool, which, in turn, reduces the search space for the ANN, contributing to the improved predictive power of the model (Sant'Anna, Silva, Nascimento, & Cruz, 2020b).

Conclusion

Boosting and MARS proved to be efficient methodologies for marker selection. The proposed two-step prediction model emerged as an effective strategy for mitigating computational costs while enhancing the prediction accuracy of an ANN.

Acknowledgements

The authors gratefully acknowledge the *Fundação de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG), *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES), and *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq).

References

Abdulelah Al-Sudani, Z., Salih, S. Q., Sharafati, A., & Yaseen, Z. M. (2019). Development of multivariate adaptive regression spline integrated with differential evolution model for streamflow simulation. *Journal of Hydrology*, *573*, 1-12. DOI: https://doi.org/10.1016/J.JHYDROL.2019.03.004

Aono, A. H., Francisco, F. R., Souza, L. M., Gonçalves, P. S., Scaloppi Junior, E. J., Le Guen, V., ... Souza, A. P. (2022). A divide-and-conquer approach for genomic prediction in rubber tree using machine learning. *Scientific Reports*, *12*, 1-14. DOI: https://doi.org/10.1038/s41598-022-20416-z

Azevedo, C. F., Resende, M. D. V., Silva, F. F., Lopes, P. S., & Guimarães, S. E. F. (2013). Regressão via componentes independentes aplicada à seleção genômica para características de carcaça em suínos. *Pesquisa Agropecuária Brasileira*, *48*(6), 619-626. DOI: https://doi.org/10.1590/S0100-204X2013000600007

Costa, J. A., Azevedo, C. F., Nascimento, M., Silva, F. F., Resende, M. D. V., & Nascimento, A. C. C. (2020). Genomic prediction with the additive-dominant model by dimensionality reduction methods. *Pesquisa Agropecuária Brasileira*, *55*, 1-11. DOI: https://doi.org/10.1590/S1678-3921.pab2020.v55.01713

Costa, W. G., Celeri, M. O., Barbosa, I. P., Silva, G. N., Azevedo, C. F., Borém, A., ... Cruz, C. D. (2022). Genomic prediction through machine learning and neural networks for traits with epistasis.

Computational and Structural Biotechnology Journal, *20*, 5490–5499. DOI: https://doi.org/10.1016/j.csbj.2022.09.029

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., ... Varshney, R. K. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in Plant Science*, *22*(11), 961-975. DOI: https://doi.org/10.1016/j.tplants.2017.08.011

Cruz, C. D. (2013). Genes: a software package for analysis in experimental statistics and quantitative genetics. *Acta Scientiarum. Agronomy*, *35*(3), 271-276. DOI: https://doi.org/10.4025/actasciagron.v35i3.21251

Cruz, C. D., & Nascimento, M. (2018). *Inteligência computacional aplicada ao melhoramento genético* (1. ed.). Viçosa, MG: Editora UFV.

Ehret, A., Hochstuhl, D., Gianola, D., & Thaller, G. (2015). Application of neural networks with back-propagation to genome-enabled prediction of complex traits in Holstein-Friesian and German Fleckvieh cattle. *Genetics Selection Evolution*, *47*(1), 1-9. DOI: https://doi.org/10.1186/S12711-015-0097-5

Fialho, I. C., Azevedo, C. F., Nascimento, A. C. C., Teixeira, F. R. F., Resende, M. D. V., & Nascimento, M. (2023). Factor analysis applied in genomic prediction considering different density marker panels in rice. *Euphytica*, *219*(9), 88. DOI: https://doi.org/10.1007/s10681-023-03214-0

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189-1232. DOI: https://doi.org/10.1214/aos/1013203451

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*(1), 1-67. DOI: https://doi.org/10.1214/AOS/1176347963

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R.* New York, NY: Heidelberg Dordrecht; London, UK: Springer. DOI: https://doi.org/10.1007/978-1-4614-7138-7

Glória, L. S., Cruz, C. D., Vieira, R. A. M., Resende, M. D. V., Lopes, P. S., Siqueira, O. H. G. B. D., & Fonseca e Silva, F. (2016). Accessing marker effects and heritability estimates from genome prediction by Bayesian regularized neural networks. *Livestock Science*, *191*, 91-96. DOI: https://doi.org/10.1016/j.livsci.2016.07.015

Goddard, M. E., Hayes, B. J., & Meuwissen, T. H. E. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. *Journal of Animal Breeding and Genetics*, *128*(6), 409-421. DOI: https://doi.org/10.1111/j.1439-0388.2011.00964.x

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). *International Statistical Review*, *77*(3), 482. DOI: https://doi.org/10.1111/j.1751-5823.2009.00095_18.x

Ho, D. S. W., Schierding, W., Wake, M., Saffery, R., & O'Sullivan, J. (2019). Machine learning SNP based prediction for precision medicine. *Frontiers in Genetics*, *10*(267), 1-10. DOI: https://doi.org/10.3389/fgene.2019.00267

Howard, R., Carriquiry, A. L., & Beavis, W. D. (2014). Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3 Genes|Genomes|Genetics*, *4*(6), 1027-1046. DOI: https://doi.org/10.1534/g3.114.010298

Huang, H., Ji, X., Xia, F., Huang, S., Shang, X., Chen, H., ... Mei, K. (2020). Multivariate adaptive regression splines for estimating riverine constituent concentrations. *Hydrological Processes*, *34*(5), 1213-1227. DOI: https://doi.org/10.1002/HYP.13669

Kao, L. J., & Chiu, C. C. (2020). Application of integrated recurrent neural network with multivariate adaptive regression splines on SPC-EPC process. *Journal of Manufacturing Systems*, *57*, 109-118. DOI: https://doi.org/10.1016/j.jmsy.2020.07.020

Li, B., Zhang, N., Wang, Y. G., George, A. W., Reverter, A., & Li, Y. (2018). Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Frontiers in Genetics*, *9*(237), 1-20. DOI: https://doi.org/10.3389/fgene.2018.00237

Long, N., Gianola, D., Rosa, G. J. M., & Weigel, K. A. (2011). Marker-assisted prediction of non-additive genetic values. *Genetica*, *139*(7), 843-854. DOI: https://doi.org/10.1007/s10709-011-9588-7

Long, N., Gianola, D., Rosa, G. J. M., Weigel, K. A., Kranis, A., & Gonzlez-Recio, O. (2010). Radial basis function regression methods for predicting quantitative traits using SNP markers. *Genetics Research*, *92*(3), 209-225. DOI: https://doi.org/10.1017/S0016672310000157

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819-1829. DOI: https://doi.org/10.1093/genetics/157.4.1819

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). *Multivariate statistical machine learning methods for genomic prediction*. Cham, GE: Springer.

Nayana, B. M., Kumar, K. R., & Chesneau, C. (2022). Wheat yield prediction in India using principal component analysis-multivariate adaptive regression splines (PCA-MARS). *AgriEngineering*, *4*(2), 461-474. DOI: https://doi.org/10.3390/agriengineering4020030

Ogutu, J. O., Piepho, H. P., & Schulz-Streeck, T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proceedings*, *5*(Suppl. 3), 1-5. DOI: https://doi.org/10.1186/1753-6561-5-S3-S11

Paixão, P. T. M., Nascimento, A. C. C., Nascimento, M., Azevedo, C. F., Oliveira, G. F., Silva F. L., & Caixeta, E. T. (2022). Factor analysis applied in genomic selection studies in the breeding of *Coffea canephora*. *Euphytica*, *218*(42), 1-9. DOI: https://doi.org/10.1007/s10681-022-02998-x

Park, J., & Kim, J. (2018). Defining heatwave thresholds using an inductive machine learning approach. *PLoS ONE*, *13*(11), 1-11. DOI: https://doi.org/10.1371/journal.pone.0206872

R Core Team. (2022). *R: A language and environment for statistical computing.* Vienna, AT: R Foundation for Statistical Computing.

Resende, M. D. V., Silva, F. F., & Azevedo, C. F. (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*. Viçosa, MG: Editora UFV.

Rosado, R. D. S., Cruz, C. D., Barili, L. D., Souza Carneiro, J. E., Carneiro, P. C. S., Carneiro, V. Q., Silva, J.T., & Nascimento M. (2020). Artificial neural networks in the prediction of genetic merit to flowering traits in bean cultivars. *Agriculture*, *10*(12). DOI: https://doi.org/10.3390/agriculture10120638

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386-408. DOI: https://doi.org/10.1037/H0042519

Sant'Anna, I. C., Nascimento, M., Silva, G. N., Cruz, C. D., Azevedo, C. F., Gloria, L. S., & Silva, F. F. (2020a). Genome-enabled prediction of genetic values for using radial basis function neural networks. *Functional Plant Breeding Journal*, *1*(2), 1-8. DOI: https://doi.org/10.35418/2526-4117/v1n2a1

Sant'Anna, I. C., Silva, G. N., Nascimento, M., & Cruz, C. D. (2020b). Subset selection of markers for the genome-enabled prediction of genetic values using radial basis function neural networks. *Acta Scientiarum. Agronomy*, *43*(1), 1-10. DOI: https://doi.org/10.4025/actasciagron.v43i1.46307

Silva, G. N., Sant'Anna, I. C., Cruz, C. D., Nascimento, M., Azevedo, C. F., & Gloria, L. S. (2022). Neural networks and dimensionality reduction to increase predictive efficiency for complex traits. *Genetics and Molecular Research*, *21*(1), 1-13. DOI: https://doi.org/10.4238/gmr18982

Silveira, L. S., Lima, L. P., Nascimento, M., Nascimento, A. C. C., & Silva, F. F. (2020). Regression trees in genomic selection for carcass traits in pigs. *Genetics and Molecular Research*, *19*(1), 1-11. DOI: https://doi.org/10.4238/GMR18498

Song, H., & Hu, H. (2022). Strategies to improve the accuracy and reduce costs of genomic prediction in aquaculture species. *Evolutionary Applications*, *15*(4), 578-590. DOI: https://doi.org/10.1111/eva.13262

Sousa, I. C., Nascimento, M., Sant'anna, I. C., Caixeta, E. T., Azevedo, C. F., Cruz, C. D., ... Vergara Lopes Serão, N. (2022). Marker effects and heritability estimates using additive-dominance genomic architectures via artificial neural networks in *Coffea canephora*. *PLoS ONE*, *17*(1), 1-14. DOI: https://doi.org/10.1371/journal.pone.0262055

Sousa, I. C., Nascimento, M., Silva, G. N., Nascimento, A. C. C., Cruz, C. D., Almeida, D. P., ... Caixeta, E. T. (2020). Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. *Scientia Agricola*, *78*(4), 1-8. DOI: https://doi.org/10.1590/1678-992X-2020-0021

Voss-Fels, K. P., Cooper, M., & Hayes, B. J. (2019). Accelerating crop genetic gains with genomic selection. *Theoretical and Applied Genetics*, *132*, 669-686. DOI: https://doi.org/10.1007/s00122-018-3270-8

Westhues, C. C., Mahone, G. S., Silva, S., Thorwarth, P., Schmidt, M., Richter, J. C., ... Beissinger, T. M. (2021). Prediction of maize phenotypic traits with genomic and environmental predictors using gradient boosting frameworks. *Frontiers in Plant Science*, *12*(699589), 1-22. DOI: https://doi.org/10.3389/FPLS.2021.699589

Xu, Y., Zhang, X., Li, H., Zheng, H., Zhang, J., Olsen, M. S., … Qian, Q. (2022). Smart breeding driven by big data, artificial intelligence and integrated genomic-enviromic prediction. *Molecular Plant, 15*(11), 1664-1695. DOI: https://doi.org/10.1016/j.molp.2022.09.001

Zabihi, M., Pourghasemi, H. R., Motevalli, A., & Zakeri, M. A. (2019). Gully erosion modeling using gis-based data mining techniques in Northern Iran: A comparison between boosted regression tree and multivariate adaptive regression spline. In H. R. Pourghasemi, & M. Rossi (Eds.), *Natural hazards gis-based spatial modeling using data mining techniques* (p. 1-26). Cham, GE: Springer. DOI: https://doi.org/10.1007/978-3-319-73383-8_1