Acta Scientiarum

http://www.periodicos.uem.br/ojs/
ISSN on-line: 1807-8621
Doi: 10.4025/actasciagron.v47i1.72552

Marker pre-selection as a strategy to enhance genomic prediction with machine learning: Exploring the influence of trait-specific genomic structures

Wagner Faria Barbosa[1*], Antônio Carlos da Silva Júnior[2], Ithalo Coelho de Sousa[3], Francyse Edite de Oliveira Chagas de Moraes[4], Michele Jorge Silva Siqueira[5], Leonardo Lopes Bhering[2], Moysés Nascimento[1] and Cosme Damião Cruz[2]

[1]Departamento de Estatística, Universidade Federal de Viçosa, Av. PH Rolfs, s/n, 36570-900, Viçosa, Minas Gerais, Brazil. [2]Departamento de Biologia Geral, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil. [3]Departamento Acadêmico de Matemática e Estatística, Universidade Federal de Rondônia, Campus de Ji-Paraná, Ji-Paraná, Rondônia, Brazil. [4]Escola Estadual Effie Rolfs, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil. [5]Departamento de Genética, Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, São Paulo, Brazil. *Author for correspondence. E-mail: barbosawf@gmail.com

ABSTRACT. This study focused on incorporating dimensionality reduction based on marker significance to better harness the potential of machine learning for genomic prediction in different trait-genomic structures. The aim was to show that outcomes achieved with reduced data would improve predictive accuracy (R^2) and precision (root-mean-square error: RMSE) while reducing computational time. Distinct subsets of markers, in simulated data, were chosen by prioritizing importance via the Bagging technique. Predictive modelling was subsequently conducted using both Bagging and the diverse architectures of a Multilayer Perceptron (MLP) neural network. This study was carried out with six traits of an F_2 simulated population (derived from contrasting homozygotes) with 1,000 individuals. Three traits had three different heritabilities (0.4, 0.6, and 0.8) and were controlled by a set of 40 quantitative trait loci (QTLs). Additionally, four QTLs with more pronounced heritability effects (set at unity) were introduced in three other traits while preserving the same genetic control structure as the earlier traits. In our investigation, as the number of markers increased, both techniques gradually increased training time; however, the time needed for computation notably extended beyond the threshold of 100 markers for Bagging. In comparison to the MLP model, the Bagging model generally obtained better accuracy (higher R^2) and precision (lower RMSE) values regardless of heritability and added QTLs. Most importantly, results highlight that for traits subject to robust genetic control of additional QTLs, MLP networks experienced a decline in prediction performance from a few markers (~10). In contrast, Bagging kept constant or subtly improved predication performance. Finally, the dimensionality reduction procedure effectively improves genomic prediction, and Bagging captures complex genetic control structures for prediction better than MLP networks.

Keywords: bagging; multilayer perceptron; artificial neural networks; genomic wide selection.

Received on June 14, 2024.
Accepted on November 5, 2024.

Introduction

Genomic-wide selection (GWS), initially proposed by Meuwissen et al. (2001), has been an indispensable tool for breeders by associating molecular marker information with phenotypic traits. Through GWS, the estimation of individual genomic estimated breeding values (GEBVs) becomes possible without the need for phenotyping, thus amplifying genetic gains while reducing time and resources (Sant'Anna et al., 2020). This capacity to accelerate genetic progress through DNA insights is valuable as it enhances selective accuracy even without prior selection-associated mapping knowledge (Alkimim et al., 2020; Barbosa et al., 2021).

However, GWS faces substantial challenges in high-dimensional situations, where the number of markers exceeds the number of genotypes and phenotypes (Akdemir et al., 2017; Azevedo et al., 2014; Crossa et al., 2017). Notably, the abundance of markers across the genome juxtaposed with fewer individuals often leads to model overparameterization (Long et al., 2010). In this context, artificial neural networks (ANNs) and machine learning exhibit significant promise by capturing nonlinear marker relationships directly from the data (i.e., without a prior model definition); a feat often beyond the ability of conventional GWS models (Howard et al., 2014; Long et al., 2011a; Sant'Anna et al., 2020).

Different methods have been employed to address this dimensionality concern, including RR-BLUP (ridge-regression best linear unbiased prediction) (Endelman, 2011), Bayesian methodologies (Meuwissen et al., 2001), computational intelligence-based approaches utilizing radial basis function networks (Sant'Anna et al., 2020) and multilayer perceptron (MLP) neural networks (Barbosa et al., 2021). Machine learning methodologies, such as Boosting, Random Forests, Bagging, and their refined variants (Barbosa et al., 2021; Sousa et al., 2020), have also been utilized.

Despite the undeniable promise held by ANNs and machine learning (ML) methodologies for genomic analysis, their effectiveness can be hindered by an abundance of markers, which triggers substantial computational burdens and makes the learning process complex. This challenge arises because a considerable proportion of the possible explanatory resources, namely, markers, contribute either minimally or not at all to genomic prediction (Ehret et al., 2015; Long et al., 2011b). Only markers showing linkage disequilibrium (LD) with quantitative trait loci (QTLs) are relevant in the prediction of GEBVs. In addition, these markers can explain the genetic variation underlying the traits of interest (Resende et al., 2012).

Nevertheless, a prudent approach involving prior selection of a subset of markers potentially linked with the target traits can be adopted as a strategy to alleviate the complications arising from high dimensionality and enhance the predictive capabilities of the models (Long et al., 2010).

Given these considerations, this study aimed to: i) select distinct subsets of markers by prioritizing their importance through the Bagging technique; ii) assess the effectiveness of dimensionality reduction in easing computational timing costs and increasing predictive accuracy through the reimplementation of the Bagging method and the exploration of diverse architectures of MLP neural networks; and iii) evaluate the impact of trait-specific genomic structures on the predictive capacity of the employed approaches.

Material and methods

Data simulation

A simulated dataset was generated for an F_2 population through a controlled cross between two distinct homozygous parental lines. This dataset included 1,000 individuals and was intended to help compare different approaches for predicting genetic values. The simulation was carried out using Genes software (Cruz, 2013).

The genome of this simulated population consisted of ten linkage groups (LGs), which followed a diploid species (2n = 20). Each of these LGs spanned a length of 100 cM and encompassed approximately 200 SNP-like codominant molecular markers. These markers were strategically positioned at equal intervals along the LGs, resulting in a total of 2,010 markers across the genome.

Six distinct traits (X_1, X_2, X_3, X_4, X_5, and X_6) were introduced within this simulated population. These traits were designed to have broad-sense heritabilities (h^2) of 0.4, 0.6, and 0.8 for different pairs (X_1-X_2, X_3-X_4, X_5-X_6, respectively). Furthermore, the genetic control of all six traits involved 40 QTLs with minor effects. Therefore, the phenotypic values of a given trait, under the influence of minor-effect QTLs, were obtained using the following model:

$$Y_i = \mu + \sum_{j=1}^{40} p_j \alpha_j + \sum_{j=1}^{40} p_j \alpha_j \, \alpha_{j+1} + \varepsilon_i$$

where: Y_i is the phenotypic value of the ith observation; μ is the overall mean; α_j is the effect of the favourable allele in the jth locus and assumes genotypic values of $u + a_j$, $u + d_j$, and $u - a_j$, corresponding to the genotypic classes AA, Aa, and aa, which were encoded as 1, 0, and -1, respectively; in addition, α_j represents the combined additive and dominant effects (i.e., $a_j = a_i + d_i$), where d_i/a_i averages 1; $\alpha_j \alpha_{j+1}$ accounts for interactions between favourable alleles at different loci; ε_i follows a normal distribution $N(0, \sigma^2)$, where σ^2 is obtained by $(1 - h^2)\sigma_g^2/h^2$, with σ_g^2 representing the genotypic variance.

Traits X_2, X_4, and X_6 possessed an additional layer of complexity by inserting four major-effect QTLs, which remained invulnerable to environmental influences (i.e., $h^2 = 1$) which resulted in a more pronounced genetic impact on these traits. These additional QTLs simulated the effect of recessive genes without a cumulative effect. Therefore, the dominant pattern (A-B-C-D-) caused Y_i to increase to the maximum increment determined by polygenic action. In contrast, other patterns (i.e., when at least one recessive pair is present) increased to the minimum polygenic increment. These increment values were arbitrary and established consistently with the magnitude and scale of the analysed variable.

Notably, the simulated data structure exhibited specific characteristics due to the predefined simulation parameters and intended purposes. Therefore, (a) all traits were highly correlated since they predominantly shared the same QTLs; (b) each controlling QTL was accompanied by a corresponding marker, resulting in 40 markers directly linked to the QTLs of traits X_1, X_3, and X_5 and 44 markers linked to the QTLs of traits X_2, X_4, and X_6; (c) 40 common markers were distributed across the initial eight LGs; (d) the four major-effect markers were confined to the first four LGs (Table 1); and (e) the final LGs lacked markers associated with the traits, which rendered 402 markers redundant and unnecessary for the prediction process.

Table 1. Positioning of the five markers controlling all six simulated traits (X_1, X_2, X_3, X_4, X_5, and X_6) within their respective linkage groups (LGs). The markers exerting significant effects on traits X_2, X_4, and X_6 are distinctly highlighted in bold.

Linkage groups	1st	2nd	Major-effect QTL	3rd	4th	5th
LG1	11	56	80	101	146	191
LG2	212	257	280	302	347	392
LG3	413	458	480	503	548	593
LG4	614	659	680	704	749	794
LG5	815	860	-	905	950	995
LG6	1016	1061	-	1106	1151	1196
LG7	1217	1262	-	1307	1352	1397
LG8	1418	1463	-	1508	1553	1598

Computational approaches

Bagging

Bagging was chosen for marker selection and genomic prediction in this work. This technique is a powerful machine learning approach that enhances predictive performance by combining the outcomes of multiple decision trees, each constructed from resampled dataset observations (bootstrapping). This combination effectively reduces prediction variance, minimizes error rates, and mitigates overfitting issues (Breiman, 1996, Breiman, 2001; Prasad et al., 2006). As a result, a set of B models, denoted as $\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x)$, is generated. These models are aggregated to create an average model: $\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x)$.

For the maker selection phase, the Bagging technique was applied across the entire dataset, encompassing 1,000 observations and 2,010 markers for each trait. The Bagging procedure was performed using the randomForest() function from the randomForest package (Liaw & Wiener, 2014) available for the R environment (R Core Team, 2022). To configure the function for running Bagging rather than Random Forest, the parameter controlling the random sampling of predictors at each split (referred to as "mtry") was set to the maximum number of available markers, i.e., 2010. This adjustment ensured that every available marker was considered a potential candidate during the split decision process. Moreover, the number of trees was set to 1,000 for the parameter referred to as "ntree", while the assessment of predictor importance was enabled by specifying the "importance" parameter as TRUE. Other function parameters were left at their default settings.

After training, markers were sorted in descending order by importance score according to the percent increase in mean squared error (%IncMSE) for each phenotypic trait. Subsets consisting of 1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 100, 500, 1,000, and 2,010 markers were employed and analysed by another round of Bagging or were introduced as inputs to the MLP networks. These processes were conducted within the framework of a cross-validation method to ensure a more rigorous evaluation.

Multilayer Perceptron (MLP) neural network

In this study, the MLP neural network, initially proposed by Rosenblatt (1958), was used for genomic prediction. This neural network was executed using the mlp() function from the RSNNS package (Bergmeir & Benítez, 2012), which is available within the R environment.

The neurons, denoted as W_m, are generated by linear combinations of M_j (markers) input variables. The ultimate output variable Y_k is determined as a function of linear combinations of neurons W_m, depicted by the following equations:

$$W_m = \sigma(\alpha_{om} + \alpha_m^T X), \qquad m = 1, 2, \ldots, M$$

$$T_k = \beta_{om} + \beta_k^T W, \qquad k = 1, 2, \ldots, K$$

$Y_k = g_k(T), \quad k = 1, 2, \ldots, K$

$W = (W_1, W_2, \cdots, W_m), T = T_1, T_2, \ldots, T_k.$

The weights $(\alpha_{om}, \alpha_m;\ m = 1, 2, \ldots, M)$ and $(\beta_{0k}, \beta_k;\ k = 1,2, \ldots, K)$ are unknown network parameters whose optimization aligned the ANN model with the training set. During this phase, the backpropagation of error, defined as "Std_Backpropagation" in the "learnFunc" parameter of the mlp() function, was employed to adjust the neuron weights iteratively. The adjustment measure was cross-entropy and computed as $R(\Theta) = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} log f_k(x_i)$. The evaluation encompassed MLP networks with one or two hidden layers, each having a range spanning from one to twenty neurons (1, 2, 5, 10, 15, and 20). As a result, 42 unique neural architectures were assessed, ensuring a comprehensive analysis. These architectures were conveniently defined through a numeric vector assigned to the "size" parameter within the mlp() function. The upper limit for learning iterations was set to 10,000 for the "maxit" parameter. The logistic function was employed as the activation function for the hidden layers, the identity function was assigned to the output layer, and both had predefined settings within the mlp() function. Finally, the learning rate, set to 0.005, was specified in the "learnFuncParams" parameter of the mlp() function.

Genomic prediction with cross-validation

A k-fold cross-validation process was implemented to ensure a rigorous evaluation. The first population of 1,000 observations was initially randomized, then systematically partitioned into five distinct groups, each forming 200 observations. Four of these groups were sequentially merged to train the model (either Bagging or MLP techniques), while the remaining group was used to test the model. This procedure was iterated five times, with alternating combinations of the groups to ensure that each observation took part in the validation phase exactly once.

Throughout the cross-validation routine for genomic prediction, the parameters employed in the randomForest() and mlp() functions remained consistent with those previously detailed. The subset of predictors (i.e., markers) varied within each training and validation phase according to its pre-selection setup through the Bagging technique. Moreover, the value chosen for the "mtry" parameter within the randomForest() function corresponded to the maximum number of markers used in every cross-validation round.

Prediction efficiency and computer effort

The prediction efficiency was evaluated through selective accuracy (R^2) and root-mean-square error (RMSE) metrics employed within the cross-validation routine. The accuracy metric measures how much the estimated values ($\hat{y}$) are related to the observed values (y). In the context of quantitative genetics, R^2 mirrors the trait's heritability (h^2) and is calculated as $[correlation(y, \hat{y})]^2$ (Cruz et al., 2012). Conversely, RMSE quantifies the disparity between estimated and observed values and is calculated as $\sqrt{\sum(\hat{y} - y)^2/n}$, where n is the number of individuals. Additionally, the computer effort was assessed by recording the training time (in seconds) for each model (Bagging or MLP techniques) across the rounds.

Outcomes were summarized into average R^2, RMSE, and their respective standard errors (SE), which were all derived from the five validation rounds. To gain a general understanding of the prediction effectiveness and computational effort of the neural networks, these metrics were further averaged across neural structures (n = 42) after being averaged across folds. These results are depicted in different graphical plots constructed using the ggplot2 package within the R software (Wickham, 2016). The metrics were aggregated based on the distinct number of marker subsets and later presented in tabular format. Supplementary figures are available at https://github.com/barbosawf/Optmizing-genomic-prediction-with-marker-selection.

Results

Figure 1 displays the training times for the Bagging and MLP techniques applied to the six simulated traits. Notably, discrepancies were observed in training times across the two approaches. Specifically, in the Bagging approach, there was a substantial increase in computational time across all six traits when the marker count surpassed 100; at this threshold, the average computational time was 24.20 ± 0.08 seconds. Without marker selection, the average time increased significantly to 530.81 ± 0.81 seconds (Figure 1A). In contrast, the MLP network showed a smoother increase in time as the markers were progressively incorporated into the models

(Figure 1A and B). Within the MLP framework, the average time across the traits ranged from 17.21 to 48.42 seconds, which covers the span of 1 to 2,010 markers (Figure 1B).

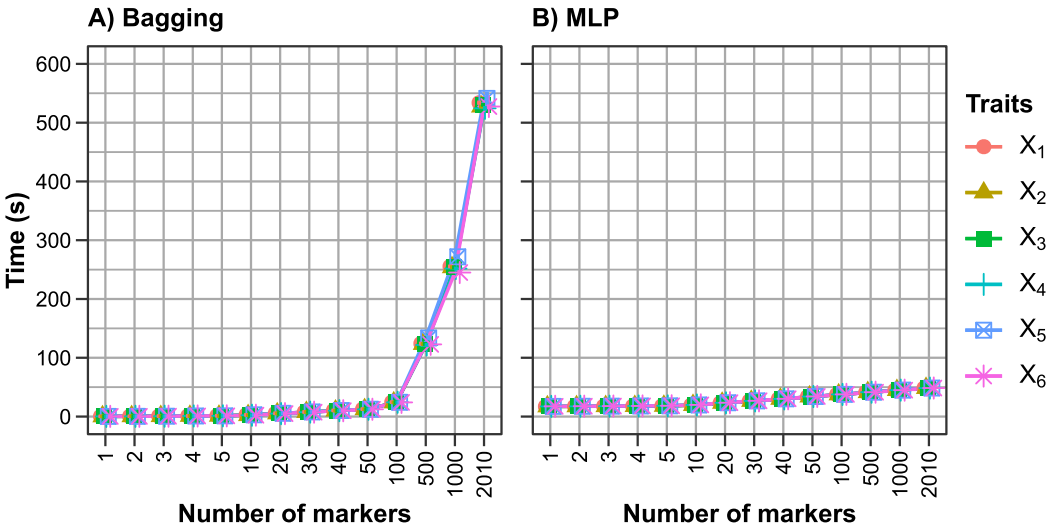

Figure 1. Time (in seconds) taken in the training of Bagging (A) and MLP (B) models with six simulated traits (X_1, X_2, X_3, X_4, X_5, and X_6) using different marker subsets. For (A), the values were averaged across the five folds of the validation procedure, while for (B), the mean of the folds was further averaged across all 42 neural architectures evaluated. Vertical lines indicate standard error.

Using a heatmap, further scrutiny of the computational effort of MLP networks revealed that training time distinctly increases as the number of markers increases, particularly in models featuring a higher density of neurons within the hidden layers (Figure 2). Conversely, the first hidden layer contributes more to the computational effort than the second layer (Figure 2). Furthermore, training durations for both Bagging and MLP network models showed no discernible association with trait heritability nor the presence of influential QTLs (Figure 2).

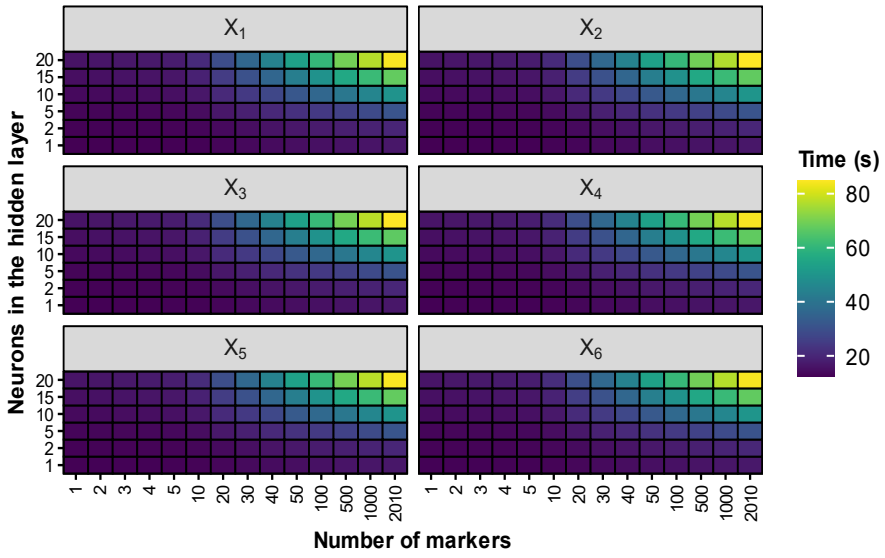

Figure 2. Heatmap illustrating the time (in seconds) invested in training multilayer perceptron (MLP) neural network models for six simulated traits (X_1, X_2, X_3, X_4, X_5, and X_6), varying marker quantities, and diverse neural architectures. The color gradient indicates the averaged value obtained from the k-fold training process.

In addition to training efforts, the R^2 and RMSE estimates, analysed by MLP neural network and Bagging approaches, offer a comprehensive assessment of the overall predictive capabilities of the models across the spectrum of the six simulated traits. Upon examination of Table 2 and Figures 3 and 4, an explicit correlation appears between the R^2 estimate, the RMSE estimate, the heritability of the simulated traits, and the presence of additional QTLs. Notably, higher heritability (h^2) values were directly associated with elevated R^2 estimates and lower RMSE values (Table 2; Figure 3). The traits featuring an additional four QTLs presented high R^2

estimates (Table 2, Figure 3), which coincided with elevated RMSE values (Table 2; Figure 4). When the performance of the techniques was examined, the Bagging approach consistently yielded the highest R^2 values and the lowest RMSE scores across all simulated traits (Table 2; Figures 3 and 4). Furthermore, following a rapid increase in R^2 and a decrease in RMSE with four markers, the prediction capacity of the Bagging models slightly improved as the number of markers increased (Figures 3A and 4A). Conversely, the MLP models experienced a clear decline in their predictive capacity as the number of markers increased. In traits with four additional QTLs, this trend occurred after the MLP models reached their peak performance between 4 and 20 markers (Figures 3B and 4B). MLP model accuracy (R^2), however, was not impacted by increasing the number of markers for the traits without the additional QTLs (i.e., X_1, X_3, X_5), whereas precision (RMSE) was impacted, especially for the trait with the lowest heritability (i.e., X_1).

Table 2. The overall mean of accuracy (R^2) and precision (root-mean-square error: RMSE) in the validation of either perceptron multilayer (MLP) neural networks or Bagging models trained with six different simulated traits (X_1, X_2, X_3, X_4, X_5, and X_6).

h^2	Additional QTLs?	Traits	R^2		RMSE	
			Bagging [a]	MLP [b]	Bagging [a]	MLP [b]
0.2	No	X_1	0.14 ± 0.01	0.09 ± 0.01	14.20 ± 0.14	16.43 ± 0.51
	Yes	X_2	0.47 ± 0.02	0.25 ± 0.02	16.04 ± 0.27	20.61 ± 0.65
0.4	No	X_3	0.22 ± 0.01	0.15 ± 0.01	11.07 ± 0.10	12.63 ± 0.27
	Yes	X_4	0.58 ± 0.02	0.33 ± 0.03	13.21 ± 0.34	17.67 ± 0.50
0.6	No	X_5	0.26 ± 0.02	0.17 ± 0.01	9.63 ± 0.13	10.94 ± 0.17
	Yes	X_6	0.62 ± 0.03	0.37 ± 0.03	11.89 ± 0.37	16.31 ± 0.53

[a]Values were sequentially averaged by folds (5) and subsets of markers (14). [b] Values were sequentially averaged by folds (5), neural structures (42), and subsets of markers (14).

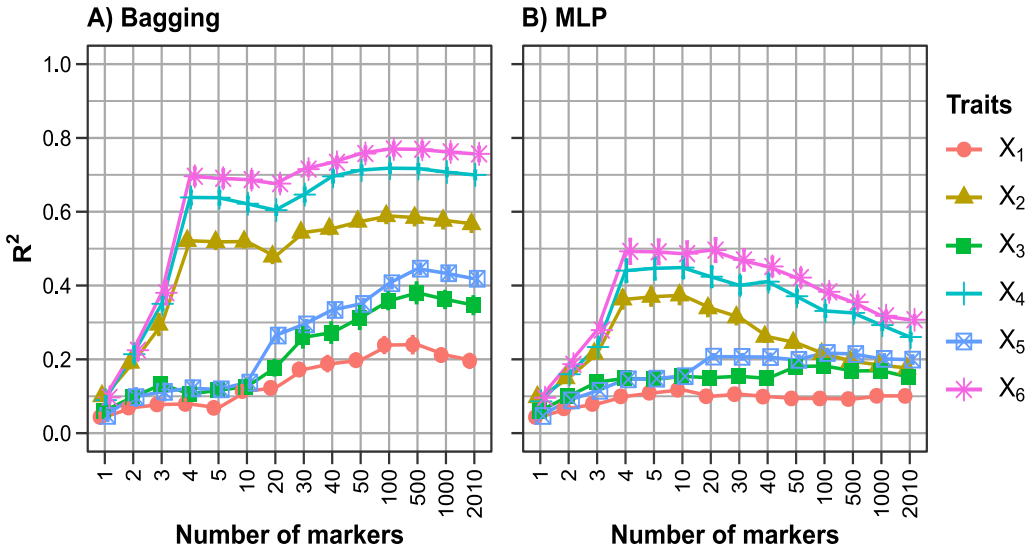

Figure 3. The validation procedure's accuracy (represented by R^2) obtained with different marker subsets for six simulated traits (X_1, X_2, X_3, X_4, X_5, and X_6). In the Bagging approach (A), the R^2 values are the averages derived from the five validation folds. In the MLP approach (B), the R^2 values were initially averaged across the five validation folds and then further aggregated across all examined neural architectures (n = 42). The vertical lines in both panels denote the standard errors associated with the accuracy measurements.

To assess the impact of various MLP network architectures on their predictive capabilities, heatmaps of the R^2 and RMSE values are presented in Figure 4. In this figure, the colour scale shows the mean estimate of the k-fold validation procedure for neural structures considering only a single intermediate layer. For traits with only small-effect QTLs (X_1, X_3, and X_5), the highest R^2 estimates (Figure 4A) and lowest RMSE scores (Figure 4B) were obtained by neural structures with only a few neurons. The most noteworthy R^2 (0.17 ± 0.02, 0.26 ± 0.03, and 0.33 ± 0.01) and RMSE (13.9 ± 0.31, 10.8 ± 0.24, and 9.13 ± 0.24) values were achieved using the most straightforward neural configurations with 1, 2, and 1 neuron(s) in the single hidden layer, respectively, for traits X_1, X_3, and X_5.

In contrast, traits X_2, X_4, and X_6 increased R^2 (Figure 4A) and decreased RMSE (Figure 4B) as the number of neurons increased. The most remarkable R^2 (0.52 ± 0.02, 0.62 ± 0.01, and 0.69 ± 0.02) and RMSE (15.5 ± 0.37, 12.9 ± 0.17, and 11.2 ± 0.29) scores were accomplished with a greater number of neurons (i.e., 20) in the single hidden layer for traits X_2, X_4, and X_6, respectively. Finally, introducing a second hidden layer into the

MLP neural networks did not seem to have a substantial impact on the R^2 and RMSE estimates, as they exhibited a colorimetric pattern similar to that observed when only a single hidden layer was utilized (Figure 4).

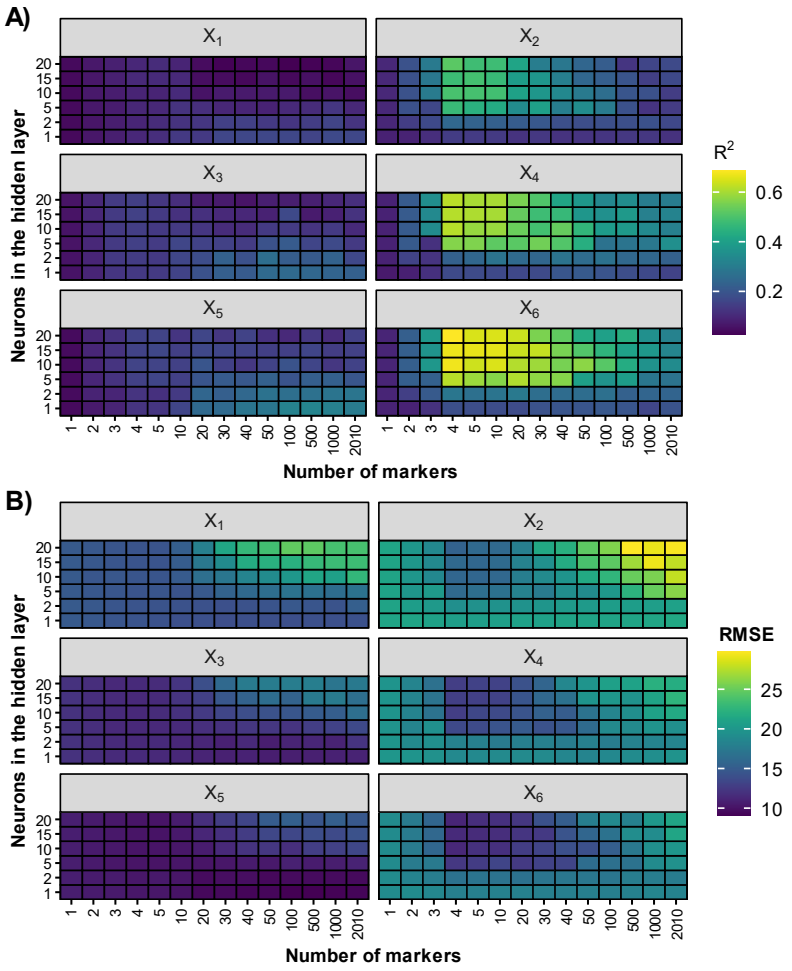

Figure 4. Heatmaps illustrating the validation procedure's accuracy (represented by R^2 in A) and precision (represented by the root mean square error [RMSE] in B) obtained with different marker subsets for six simulated traits (X_1, X_2, X_3, X_4, X_5, and X_6). Both performance metrics were obtained using different multilayer perceptron (MLP) neural network architectures. The color gradient indicates the averaged values of the R^2 (A) and RMSE (B) derived from the five validation folds.

Discussion

This study employed a decision tree-based machine learning approach, Bagging, to select unique subsets of markers based on prioritized importance to highlight the dual benefits of dimensionality reduction: enhancing predictive accuracy while reducing the computational effort. This improvement was analysed by the Bagging method (reiterated) and by exploring various MLP neural network architectures post-dimensionality reduction. In addition, the impact of different genomic trait structures on the techniques used was evaluated.

The foremost outcome to emphasize is the significant increase in computational time exhibited by Bagging after the number of markers surpassed 100. In contrast, MLP models consistently demonstrated steady growth in training time. A massive challenge for decision tree-based algorithms lies in scalability, which significantly extends the time required to process datasets with increased attributes and observations (Costa & Pedreira, 2023). This phenomenon occurs when the number of potential splits and the overall complexity of the trees experiences a notable increase.

ANNs facilitate model interactions through interconnections among neurons within their network layers (Rosenblatt, 1958). This process demands substantial computational resources, especially in high-dimensional scenarios where numerous associations lack biological significance prediction (Ehret et al., 2015; Long et al., 2011b). This situation is notably evident in a genome inundated with markers. However, under the conditions of this study, with a maximum of 2,010 markers, the training of MLP neural networks required

significantly less computational effort in comparison to the Bagging approach. Furthermore, the observed trend of smooth temporal growth across increasing neuron markers, particularly in the first layer, suggests that MLP neural networks do not excessively suffer from the scalability phenomenon.

In general, dimensionality reduction had a positive effect in the Bagging and MLP approaches. Nevertheless, the simulated heritability and additional QTLs strongly correlated with the R^2 or RMSE estimates. Thus, in both approaches, increased markers made accuracy more distinct between characteristics according to their simulated heritabilities. Furthermore, an increase in markers negatively affected accuracy and precision for MLP neural networks, but not for Bagging. The remarkable resilience of Bagging in consistently producing accurate estimates as the number of traits increases, especially when additional QTL effects are present, can be attributed to its emphasis on prioritizing the most predictive traits during the construction of multiple decision trees (Breiman, 1996; Breiman, 2001; Montesinos López et al., 2022; Prasad et al., 2006).

Conversely, MLP neural networks face a disadvantage due to the challenge of adapting neural weights under the burden of additional markers that lack biological associations with the traits (Ehret et al., 2015; Long et al., 2011b). The increase in the number of markers, arranged by importance (See Material and Methods), introduces noise that hinders optimal weight adjustments and adversely impacts predictions, particularly in traits with additional QTLs. Despite this disparity, the MLP models consistently upheld their accuracy when applied to traits unaffected by additional QTLs; albeit, at a level below that achieved by Bagging. This observation aligns with analogous studies on dimensionality reduction and reinforces the reliability of MLP models in scenarios where such genetic influences are absent (Sant'Anna et al., 2020; Silva et al., 2014; Silva et al., 2022).

Notably, introducing a second hidden layer with neurons in the MLP networks did not yield a significant enhancement in prediction across any of the traits. This observation suggests that a single hidden layer can effectively capture the nonlinearity inherent in quantitative traits, especially those with distinct genomic structures (as simulated in this study). However, our findings underscore that the pivotal factor influencing prediction accuracy in MLP networks is more closely related to the number of neurons within a single layer rather than the addition of extra layers.

Conclusion

Our results affirm the efficacy of Bagging as a proficient method for dimensionality reduction and underscore the effectiveness of employing decision-tree-based methods for such purposes, a correlation consistently highlighted in other studies (Arouisse et al., 2021; Walters et al., 2012). Moreover, ranking significant markers by importance during the prediction process can increase the accuracy and precision of the methods while concurrently lowering computational costs.

Data availability

We inform you that the data used in the research were made publicly available and can be accessed via the link https://github.com/barbosawf/Optmizing-genomic-prediction-with-marker-selection.

Acknowledgements

The authors extend their heartfelt gratitude to the *Fundação de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG), the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), and the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES, 001) for their invaluable support in funding this research.

References

Akdemir, D., Jannink, J.-L., & Isidro-Sánchez, J. (2017). Locally epistatic models for genome-wide prediction and association by importance sampling. *Genetics Selection Evolution*, *49*(74), 1-14. https://doi.org/10.1186/s12711-017-0348-8

Alkimim, E. R., Caixeta, E. T., Sousa, T. V., Resende, M. D. V., Silva, F. L., Sakiyama, N. S., & Zambolim, L. (2020). Selective efficiency of genome-wide selection in Coffea canephora breeding. *Tree Genetics & Genomes*, *16*(41). https://doi.org/10.1007/s11295-020-01433-3

Arouisse, B., Theeuwen, T. P. J. M., Van Eeuwijk, F. A., & Kruijer, W. (2021). Improving genomic prediction using high-dimensional secondary phenotypes. *Frontiers in Genetics*, *12*(667358), 1-12. https://doi.org/10.3389/fgene.2021.667358

Azevedo, C. F., Silva, F. F., Resende, M. D. V, Lopes, M. S., Duijvesteijn, N., Guimarães, S. E. F., Lopes, P. S., Kelly, M. J., Viana, J. M. S., & Knol, E. F. (2014). Supervised independent component analysis as an alternative method for genomic selection in pigs. *Journal of Animal Breeding and Genetics*, *131*(6), 452-461. https://doi.org/10.1111/jbg.12104

Barbosa, I. P., Silva, M. J., Costa, W. G., Castro Sant'Anna, I., Nascimento, M., & Cruz, C. D. (2021). Genome-enabled prediction through machine learning methods considering different levels of trait complexity. *Crop Science*, *61*(3), 1890-1902. https://doi.org/10.1002/csc2.2048

Bergmeir, C., & Benítez, J. M. (2012). Neural networks in R using the stuttgart neural network simulator: RSNNS. *Journal of Statistical Software*, *46*(7), 1-26. https://doi.org/10.18637/jss.v046.i07

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123-140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Costa, V. G., & Pedreira, C. E. (2023). Recent advances in decision trees: an updated survey. *Artificial Intelligence Review*, *56*(5), 4765-4800. https://doi.org/10.1007/s10462-022-10275-5

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., & Varshney, R. K. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science*, *22*(11), 961-975. https://doi.org/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. https://doi.org/10.4025/actasciagron.v35i3.21251

Cruz, C. D., Regazzi, A. J., & Carneiro, P. C. S. (2012). *Modelos biométricos aplicados ao melhoramento genético*. 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*(22), 1-9. https://doi.org/10.1186/S12711-015-0097-5

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome*, *4*(3), 250-255. https://doi.org/10.3835/plantgenome2011.08.0024

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. https://doi.org/10.1534/g3.114.010298

Liaw, A., & Wiener, M. (2014). Package "randomForest": Breiman and Cutler's random forests for classification and regression. *R Development Core Team*, *4*, 6-10.

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

Long, N., Gianola, D., Rosa, G. J. M., & Weigel, K. A. (2011a). Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. *Journal of Animal Breeding and Genetics*, *128*(4), 247-257. https://doi.org/10.1111/j.1439-0388.2011.00917.x

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

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

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Random Forest for Genomic Prediction. In O. A. Montesinos López, A. Montesinos López, & J. Crossa (Eds.), *Multivariate statistical machine learning methods for genomic prediction* (pp. 633–681). Springer International Publishing. https://doi.org/10.1007/978-3-030-89010-0_15

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, *9*(2), 181-199. https://doi.org/10.1007/s10021-005-0054-1

R Core Team (2022). *A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org

Resende, M. D. V., Silva, F. F., Lopes, P. S., & Azevedo, C. F. (2012). *Seleção genômica ampla (GWS) via modelos mistos (REML/BLUP), inferência bayesiana (MCMC), regressão aleatória multivariada e estatística espacial*. UFV.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386-408. https://doi.org/10.1037/h0042519

Sant'Anna, I. C., Silva, G. N., Nascimento, M., & Cruz, C. D. (2020). 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. https://doi.org/10.4025/actasciagron.v43i1.46307

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

Silva, G. N., Tomaz, R. S., Sant'Anna, I. C., Nascimento, M., Bhering, L. L., & Cruz, C. D. (2014). Neural networks for predicting breeding values and genetic gains. *Scientia Agricola*, *71*(6), 494-498. https://doi.org/10.1590/0103-9016-2014-0057

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

Walters, R., Laurin, C., & Lubke, G. H. (2012). An integrated approach to reduce the impact of minor allele frequency and linkage disequilibrium on variable importance measures for genome-wide data. *Bioinformatics*, *28*(20), 2615-2623. https://doi.org/10.1093/bioinformatics/bts483

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer.