https://doi.org/10.1590/0034-737X2024710045
ISSN: 2177-3491

Trait prediction through computational intelligence and machine learning applied to the improvement of white oat (*Avena sativa* L)

Antônio Carlos da Silva Júnior[1*], *Isabela Castro Sant'Anna*[2], *Michele Jorge da Silva*[1], *Leonardo Lopes Bhering*[3], *Moysés Nascimento*[3], *Ivan Ricardo Carvalho*[4], *José Antônio Gonzalez da Silva*[4], *Cosme Damião Cruz*[3]

[1] Universidade Federal de Viçosa, Departamento de Biologia Geral, Viçosa, MG, Brazil. michele-jorgesilva@gmail.com, leonardo.bhering@ufv.br, cdcruz@ufv.br

[2] Instituto Agronômico (IAC), Centro de Seringueira e Sistemas Agroflorestais, Votuporanga, SP, Brazil. isabelacsantanna@gmail.com

[3] Universidade Federal de Viçosa, Departamento de Estatística, Viçosa, MG, Brazil. moysesnascim@gmail.com

[4] Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Ijuí, RS, Brazil. ivan.carvalho@unijui.edu.br, jose.gonzales@unijui.edu.br

*Corresponding author: antonio.silva.c.junior@gmail.com

Editors:
Teogenes Senna de Oliveira

Submitted: May 17th, 2023.
Accepted: August 30th, 2024.

ABSTRACT

The prediction of traits allows the breeder to guide strategies to select and accelerate the progress of genetic improvement. The objective of this work was to determine the best prediction approach and establish a network with better predictive power for white oat using methodologies based on artificial intelligence, and machine learning. Seventy-eight white oat genotypes were evaluated. The design was randomized blocks with three replications. The models were evaluated with and without fungicide, and prediction models were established using four sets of experiments. The grain yield was used as a response trait the others as explanatory traits. The coefficient of determination was considered to evaluate the proposed methodologies. The importance of the traits was assessed through the impact of destructuring or disturbing the information of a given input on the estimation of R^2. For machine learning, decision trees, bagging, random forest, and boosting were used. The traits indicated to assist in decision-making are plant height, leaf rust severity, and lodging percentage. The R^2 ranged from 30.14% - 96.45% and 10.57% - 94.61% for computational intelligence and machine learning, respectively. A high estimate of the coefficient of determination, which was larger than the other estimates, was obtained using the bagging technique.

Keywords: *Avena sativa* L.; multiple regression; decision trees; Artificial neural networks.

INTRODUCTION

White oats (*Avena sativa* L.) are of great agricultural importance worldwide. Brazil is the fifth-largest producer globally and has experienced a substantial increase in areas cultivated with white oats in the last ten years (Conab, 2022). This crop can be used to produce grain, forage, and straw in a no-tillage system (Corazza *et al.*, 2021). Oat forage is preferred over other annual forage crops because of its high palatability and dry matter content (McCartney *et al.,* 2008; Kim *et al.,* 2014; Sharma *et al.,* 2022).

Estimating the importance of predictor traits in breeding programs allows for faster progress and selecting and predicting traits with low heritability and/or measurement difficulty (Silva Junior *et al.*, 2021 and 2023). Although the simultaneous assessment of traits provides a wide variety of information, identifying which predictor trait is more critical is challenging for breeders (Parmley *et al.*, 2019). The estimation of the importance of traits can be performed using artificial neural networks (ANNs) with algorithms such as that proposed by Goh (2005), who modified the Garson (1991) algorithm (Goh, 2005), which consists of partitioning the neural network connection weights to determine the relative importance of each input trait in the network.

Regression, artificial intelligence, and machine learning-based methodologies have been successfully used in studies of predicting phenotypic traits. Parmley *et al.* (2019) evaluated high-dimensional phenotypic traits in soybeans through a machine learning approach to predict seed yield for the prescriptive development of cultivars for agricultural practices. Silva Junior *et al.* (2023) used these methodologies to predict grain yield, grain length-width ratio, and panicle length in flood-irrigated rice. Silva Junior *et al.* (2021) evaluated the importance of auxiliary traits of the main trait based on phenotypic information and previously known genetic structure information using computational intelligence and machine learning to develop good predictive tools in breeding programs. However, there are no studies in the literature related to yield prediction and verification of the importance of traits for grain yield in white oat cultures.

Given the above, this research aims to (1) compare different methods to predict grain yield in white oat and evaluate the model's performance (2) evaluate the relationship between predictor and grain yield traits in white oat and (3) identify more relevant predictors based on different prediction approaches. For that, we employed three predictive models (regression, artificial intelligence, and machine learning) and some plant trait related to grain yield such as plant height, leaf rust severity, and lodging percentage.

MATERIALS AND METHODS

Experimental data

The field experiment was carried out in the experimental area of the Instituto Regional de Desenvolvimento Rural (IRDeR) at the Universidade Regional do Noroeste do Estado do Rio Grande do Sul (UNIJUÍ) located in the municipality of Augusto Pestana, Rio Grande do Sul, Brazil, at coordinates 28° 26' 30'' S and 54° 00' 58'' W and an altitude of 280 m. The soil is classified as typical Dystric Rhodic Ferralsol (World Reference Base, 2015). According to the Köeppen climate characterization, the region's climate is of the Cfa type (humid subtropical), with four distinct seasons. The average annual temperature is 19.9 °C, and the average annual rainfall is 1774 mm. The collection of plant material is in accordance with relevant institutional, national and international guidelines and legislation.

Seventy-eight white oat genotypes were evaluated in 2018 and 2019. Each year, they were assessed with and without a fungicide to establish in four sets of experiments (E1, E2, E3, and E4). The design was randomized blocks with three replications.

Grain yield (GY, kg ha^{-1}) was used as the response trait. Other traits were used as explanatory traits (inputs), mass of one thousand grains (MTG, grams); hectoliter weight (HW, kg ha^{-1}); days between emergence and maturation (DEM, day); lodging percentage (LP, percentage, where 1% means minimal bedding and 100% means complete bedding); days from emergence to flowering (DEF, day); days from flowering to maturity (DFM, day); plant height (PH, cm); leaf rust severity (LRS); stem rust severity (SRS); and leaf spots (LS). They were used to in artificial neural networks for white oat genotypes.

Methodologies for predicting and verifying the importance of traits

Multiple regression

Stepwise multiple regression is a trait selection method that aims to explain the relationship between a set of independent traits and a dependent trait (Ghani & Ahmad, 2010). The adopted model is represented by Equation:

$$y = \beta_0 + \sum_{k=1}^{n} \beta_k x_k + \varepsilon$$

where y is the response trait, x_1 a x_k are the explanatory traits, β_0 represents the intercept, β_1 e β_k are the linear coefficients associated with x_1 a x_k, and ε residual effect.

The estimate of the coefficient of determination (R^2) was used to verify how much of the independent variable is explained by the total variation of the dependent trait. The description of R^2 is found in Equation:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}$$

where y is the observed values, and $\hat{y}$ is the predicted.

Pearson's correlation analysis was used to evaluate the relationship between GY and other traits (Mukaka, 2012).

Computational intelligence for the importance of traits

Multilayer perceptron - MLP

The importance of predictors in an MLP network was quantified using two techniques. The first technique, based on Garson's (1991) algorithm modified by Goh (2005) consists of partitioning neural network connection weights to determine the relative importance of each input trait within the network.

The equation of the relative importance of traits is given as

$$RI = VW$$

The matricial model is shown as follows:

$$RI = \begin{pmatrix} RI_1 \\ RI_2 \\ \vdots \\ RI_k \end{pmatrix} = \left(W^1_{N_1 E}\right)' \left(W^2_{N_2 N_1}\right)' \ldots \left(W^c_{N_{c-1} y}\right)'$$

where W_x^c represents the matrix of weights of the layer c neuron, considering N_j neurons and N_{j-1} inputs; E is the first neuron that starts from inputs; y refers to the desired output layer and RI is the relative importance of the trait.

After the network is established, the importance of traits (inputs) can also be obtained by considering the impact of destructuring or disturbing the information of a given input on the estimation of the coefficient of determination.

The relative importance of the trait by the permutation of R^2 is described in the following equation:

$$VR_{x_i} = R^2_{\text{obs}} - \bar{R}^2_{\text{perm},x_i}$$

where R^2_{obs} is the R^2 of the RNA model adjusted to the observed predictor and response traits; $\bar{R}^2_{perm,xi}$ is the R^2 of the ANN model fitted to the modified dataset where x_i is permuted; and $\bar{R}^2_{perm,xi}$ is the average value of $R^2_{perm,xi}$ after the mth permutation of the datasets.

After some criteria were used on the best topology, the following MLP network structures were adopted: (a) topology 1: 10-11-1: ten inputs with 11 hidden neurons in the middle layer and one neuron in the output layer; (b) topology 2: 10-11-11-1: ten inputs and two hidden layers with 11 neurons in the middle layers and one neuron in the output layer; (c) topology 3: 10-11-11-11-1: ten inputs and three hidden layers with 11 neurons in the middle layers and one neuron in the output layer; (d) topology 4: 10-3-4-11-1: ten inputs and three hidden layers with 3, 4 and 11 neurons in the middle layers and one neuron in the output layer.

Radial basis function – RBF

The prediction efficiency is measured by the coefficient of determination and the relative importance of each input estimated by the technique of destructuring the information of each explanatory trait, as already described for the MLP.

Machine learning for the importance of traits

To quantify the importance of traits through a machine learning approach, the decision tree and its refinements, random forest, bagging, and boosting were used Silva Junior *et al.* (2023) and Costa *et al.* (2021).

The importance of trait IV is described in the following equation:

$$IV_{x_i} = MSE_{\text{perm},x_i} - MSE_{\text{nperm}}$$

where $MSE_{perm,xi}$ is the permutation of the values of each trait in the dataset where x_i is swapped; MSE_{nperm} represents the original nonpermuted trait data.

Importance of traits in reduced models for grain yield prediction

The biometric technique that led to the best GY prediction results and information regarding the importance of predictors was considered. The mean estimate of the relative contributions of the explanatory traits to prediction, after correcting for auxiliary traits for minor relative contributions. The choice of this technique was based on the estimation of the foreign certificate and on the traits of minor auxiliary functions.

Training and validation sets

The dataset was divided into two parts: a training and validation set. The training set included the same individuals for modeling using all methodologies and was composed of 67% of the individuals, which corresponds to 2/3 of the randomly selected individuals. The remaining 33% (1/3) of the individuals constituted the validation set. In previous studies, 60% to 90% of individuals constituted the training set (González-Camacho *et al.*, 2012). The analyses were performed with the aid of R software using the NeuralNetTools package (Beck, 2018) and Genes (Cruz, 2016).

RESULTS AND DISCUSSION

Prediction of grain yield using different approaches

The ML approach presented the best grain yield prediction performance was bagging (R^2 of 93.44%), followed by boosting (R^2 of 83.96), while the random forest approach showed the worst performance (R^2 of 39.79) (Table 1). The estimate of the coefficient of determination for all methodologies using the ten defining agronomic traits for the prediction of white oat grain yield (GY) is shown in Table 1.

Based on Table 1, it is possible to compare the approach that is more efficient for the prediction of GY. Higher values of R^2 indicate that the prediction target trait has a better fit considering the ten explanatory traits used as predictors in this analysis (Silva Junior *et al.*, 2023). Among the methodologies used in this study, it was found that multiple regression presented a lower estimate of R^2, indicating the existence of nonlinear associations between the explanatory traits not considered in the model. Artificial intelligence and machine learning methodologies stood out for their ability to extract nonlinear information from model inputs (Parmly *et al.*, 2019; Skawsang *et al.*, 2019), as seen in Table 1. Other authors have already highlighted the abilities of neural networks (Silva *et al.*, 2014; Sant'Anna *et al.*, 2015) and machine learning approaches (Sousa *et al.*, 2020; Silva Junior *et al.*, 2021) to better capture nonlinear relationships when compared to conventional methodologies.

The results obtained by different approaches show that there was a discrepancy between the maximum estimate of R^2 for the predictive trait in the same environments (Table 1). This discrepancy in the estimate of R^2 was also reported by (Silva Junior *et al.*, 2023). It is noteworthy that the differences in results obtained in these analyses are indicative that the environment influences the estimate of R^2 and, consequently, the choice of the best prediction model for the response trait.

The machine learning approach proved to be more efficient than the other approaches (Table 1). There was a low estimate of the maximum R^2 when using the random forest procedure in all environments. On the other hand, this procedure was superior to the multiple regression approach for the same environment, except the environment without fungicide (E3), where a value of 10.57% was obtained. The low estimate of the maximum R^2 in the random forest procedure was also demonstrated for flood-irrigated rice Silva Junior *et al.* (2023) and simulated data with differ-

Table 1: Mean of the maximum estimate of the coefficient of determination for the training set, in four environments corresponding to the data set of experiments no and with fungicide in two agricultural years, to predict the grain yield in white oat (*Avena sativa* L.)

Approach	Technique	E1	E2	E3	E4	Average of performance
ML	BO	92.29	86.69	81.23	79.23	83.96
	DT	85.37	76.39	61.78	64.65	70.52
	BA	**94.61**	**93.89**	**92.70**	**92.98**	93.44
	RF	64.91	55.09	10.57	24.48	39.79
AI	PMC-1	73.25	71.42	30.14	59.84	65.63
	PMC-2	**96.45**	**90.12**	56.72	57.94	74.03
	PMC-3	86.13	88.58	61.45	68.62	77.38
	PMC-4	75.16	85.32	**87.34**	58.77	80.24
	RBF	90.12	73.76	80.72	**76.44**	78.58
Conventional	RM	61.02	46.07	20.67	32.72	39.40

AI: Artificial Intelligence; ML: Machine Learning; RM: Multiple Regression; PMC: Multilayer Perceptron; PMC: Multilayer Perceptron; PMC-1: Multilayer Perceptron with (10-11-1); PMC-2: Multilayer Perceptron (10-11-11-1); PMC-3: Multilayer Perceptron (10-11-11-11-1); PMC-4: Multilayer Perceptron (10-3-4-11-1); RBR: Radial Base Network; DT: Decision Tree; RF: Random Forest; BA: Bagging; BO: boosting. E: environments. E1 and E3: no fungicide; E2 and E4: with fungicide.

ent heritability Silva Junior *et al.* (2021). This procedure involves randomly resampling the set of explanatory traits and building several decision trees that constitute a random forest, allowing the prediction and estimation of scores that will lead to the evaluation of the importance of predictors in a process repeated several times.

Regarding the environments and the bagging procedure, the estimates of R^2 were higher than 92.70%, making bagging the best approach for the analyzed datasets. High estimates (with reference to values of approximately 80%) of R^2 were also obtained using machine learning methodologies with boosting and bagging procedures on all prediction datasets (Table 1). Silva Junior *et al.* (2021) showed that the machine learning approaches with bagging and boosting procedures were more consistent in obtaining a higher overall mean estimate of R^2 of predictive traits. The decision tree (DT) and random forest methodologies did not stand out from other machine learning procedures (Table 1).

Thus, machine learning is actually more efficient for selecting phenotypic traits because it can handle reduced or redundant information about phenotypic traits (Sousa *et al*., 2020). Costa *et al.* (2021) evaluated the importance of variables using bagging, random forest, boosting, decision tree, MLP and RBF and reported that MLP and RBF achieved better results. Silva Junior *et al.* (2023) verified that the computational intelligence and machine learning methodologies in prediction allowed the identification of explanatory phenotypic traits that should be prioritized and established as auxiliary traits for indirect selection.

Artificial intelligence approaches based on RBF yielded estimates with R^2 greater than 70% in all environments (Table 1). In this procedure, the maximum R^2 was 90.12% (± 5.79), and the minimum was 73.75% (± 1.67), corresponding to environments E1 and E2, respectively. Silva Júnior *et al.* (2023) found a maximum R^2 ranging from 48% to 99% in different environments for flood-irrigated rice crops. For simulated data with different genetic structures, the maximum estimate of R^2 ranges from 44% to 54% (Silva Junior *et al.,* 2021), and (Sant'Anna *et al.*, 2020) obtained consistent results of R^2 for different genetic structures. Rosado *et al.* (2020) evaluated bean cultivars and obtained an estimate of R^2 for the trait days to first flower and flowering days of 94.10% and 94.40%, respectively. This procedure has a good ability to handle complex interactions compared to semiparametric and linear regression approaches (Sant'Anna *et al.*, 2019 and 2020). Generally, the data used as training information is quickly learned in RBF, providing a unique solution compared to perceptron ANNs (Sant'Anna *et al.*, 2020; González-Camacho *et al.*, 2012).

Sant'Anna *et al.* (2020) applied the RBF in studies using simulated traits with 30% and 60% heredity for trait selection. The authors found that greater efficiency in the selection could be obtained using the RBF when the scenario involved epistatic interactions in the gene control of the studied traits. González-Camacho *et al.* (2012) observed that it is possible to improve prediction in nonparametric models when the selection includes markers that are not directly related to the traits of interest. Silva Junior *et al.* (2023) applied the RBF to predict grain yield, grain length-width ratio, and panicle length in flood-irrigated rice. The authors argued that the RBF has good performance in predicting the importance of traits. Silva Junior *et al.* (2021) evaluated the importance of auxiliary traits of the main trait based on phenotypic information and previously known genetic structure information using the RBF and demonstrated the efficiency of this network to quantify the importance of traits.

Regarding MLP-1 (10-11-1), the highest estimate of the maximum R^2 was observed in E1 (73.25%) and the lowest, with an estimate of 30.14%, was observed in E3. Both environments correspond to those without fungicide. In the procedures MLP-2 (10-11-11-1) and MLP-3 (10-11-11-11-11-1), the highest estimates were observed in E1 and E2 and the smallest in E3 and E4, respectively. MLP-3 (10-11-11-11-1) and MLP-4 (10-3-4-11-1) have the same number of hidden layers. We observed lower estimates of the maximum R^2 for the MLP-4 procedure, except in the E3 environment. This shows that the number of neurons in the layer influences the estimation of the maximum R^2. Silva Junior *et al.* (2021) argued that the number of neurons influences the estimation of the coefficient of determination.

The MLP network is widely used in the predictive process (Silva Junior *et al.*, 2021; Sousa *et al.*, 2020) since the success of this network has already been demonstrated by several research groups, who have shown mathematically that, with only a single hidden layer, this network works very well with different numbers of neurons in the hidden layer (Sousa *et al.*, 2020).

The efficiency of ANNs in prediction problems, given their ability to extract relevant information from large datasets and generalize relatively inaccurate information (Sant'Anna *et al.*, 2020), was very well expressed in the

results obtained (Table 1). The same can be seen for methodologies based on machine learning, which are capable of dealing with more reduced or redundant information in the input traits (Silva Junior *et al.*, 2023). However, another study that is as important as prediction and that is often not carried out is the identification of more important predictive traits, which is an important factor in the decision-making process (Beucher *et al.*, 2019). Thus, after the prediction analyses, analyses were carried out to quantify the importance of traits through artificial intelligence and machine learning methodologies to identify, among the set of explanatory traits, those that should be prioritized and identified as auxiliary traits in indirect responses to selection.

Linear relationship between predictor and grain yield traits in white oat

The greatest linear associations with GY may be a preliminary indication that individual traits are important in its prediction. In multivariate prediction models, a predictor trait with high correlation with the response trait may lose its importance due to its redundancy, considering that, in the model, it may be represented by another association. Thus, in addition to quantifying the linear relationships between the predictor and response, it is important to quantify and appreciate the linear relationships, expressed by linear correlation coefficients, between all predictors in the search for redundancies. In this work, these associations were represented in a correlation network that contains red and green lines that represent negative and positive correlations, respectively, and their width is proportional to the magnitude of the correlations (Figure 1). Regarding the phenotypic correlation network, the structure of correlated groups was obtained to predict GY. In this network, the similarity between the phenotypic traits and the phenotypic correlation patterns was highlighted.

The traits that presented groups with GY in E1 were MTG, HW and PH, which showed positive correlations but varied in magnitude, and LRS, which showed a negative correlation. To E2, the positively correlated traits consisted of PH and MTG and the negatively correlated traits consisted of LS and DFM. For E3, which represents a case with no fungicide, the trait that was negatively correlated was SRS. Environment 4, the positively correlated group consisted of HW and DEF and the negative correlated group consisted of DEM (Figure 1).

Importance of trait in prediction using an artificial intelligence approach

Multilayer perceptron (MLP)

Estimates of the coefficient of determination of grain yield prediction with MLP attribute perturbation of the

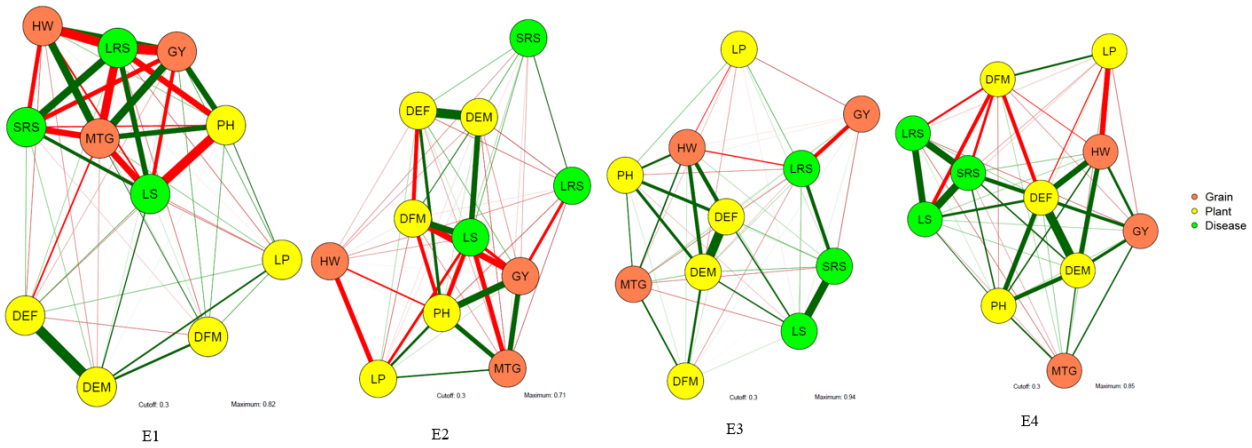

Figure 1: Phenotypic correlation network for the three distinct groups in four environments corresponding to without and with fungicide in two agricultural years, to predict grain yield in white oat (*Avena sativa* L.). The line width is proportional to the strength of the correlation. E1 and E3; E2 and E4 represent the environments without and with fungicide, respectively. The orange color represents the grain characteristics; The yellow color represents the plant characteristics and the green disease severity. MTG = Thousand Grain Mass in grams; HW = Hectoliter Weight; DEM = Days between Emergency and Maturation; PH= percentage of lodging; GY = Grain yield in kilograms per hectare; DEF = Days from Emergence to Flowering; DFM= Days from Flowering to Maturation; PH= Plant Height; LRS= Leaf Rust Severity; SRS=Stem Rust Severity and LS= Leaf Spots.

genotypic information are shown in Figure 2. These results show large discrepancies in R^2 when comparing the environments with each other, which makes interpretation difficult. In environments E1 and E4, which correspond to environments without fungicide, the traits LP, PH, LRS were efficient in quantifying the response trait GY due to the reduction in the estimate of R^2 as a function of the attribute perturbation of the phenotypic information.

Regardless of the number of neurons in the output layer and a single hidden layer, the most important traits were determined to predict GY (Figure 3). This result shows that these traits are important in predicting GY, as the perturbation of their values led to a considerable reduction in the quality of the fit. In the E2 environment, MTG was the most important trait in predicting GY.

There was a difference in the number of neurons in the output layer and hidden layer, indicating that the most important traits in E4 correspond to the fungicide environment. With only one neuron in the output layer and a single hidden layer, DEF and SRS were the most important traits due to the reduction in the estimate of R^2. With two neurons in the middle layer and a single hidden layer, LRS and LS were the most important. With one neuron in the input layer and three hidden layers with 11 neurons in the intermediate layer and one neuron in the output layer, the traits that proved to be the most important were HW and SRS. On the other hand, with three hidden layers with 3, 4 and 11 neurons in the intermediate layer, the important traits in predicting the GY were LRS, DEF and HW. Given the significant decreases in the estimated values of R^2 observed when the variables were disturbed, Silva Junior *et al.* 2023 reported that the most important traits were grain width and length in irrigated rice when using only one neuron in the output layer and a single hidden layer.

The importance of the traits was quantified by assigning destructuring to the genotypic information referring to each trait to observe the changes in the values of R^2. It is important to note that reductions in the estimative of R^2 after attribute disruption of the genotypic information referring to each trait are indicative that this trait is important in relation to the others for purposes of prediction with the already established network.

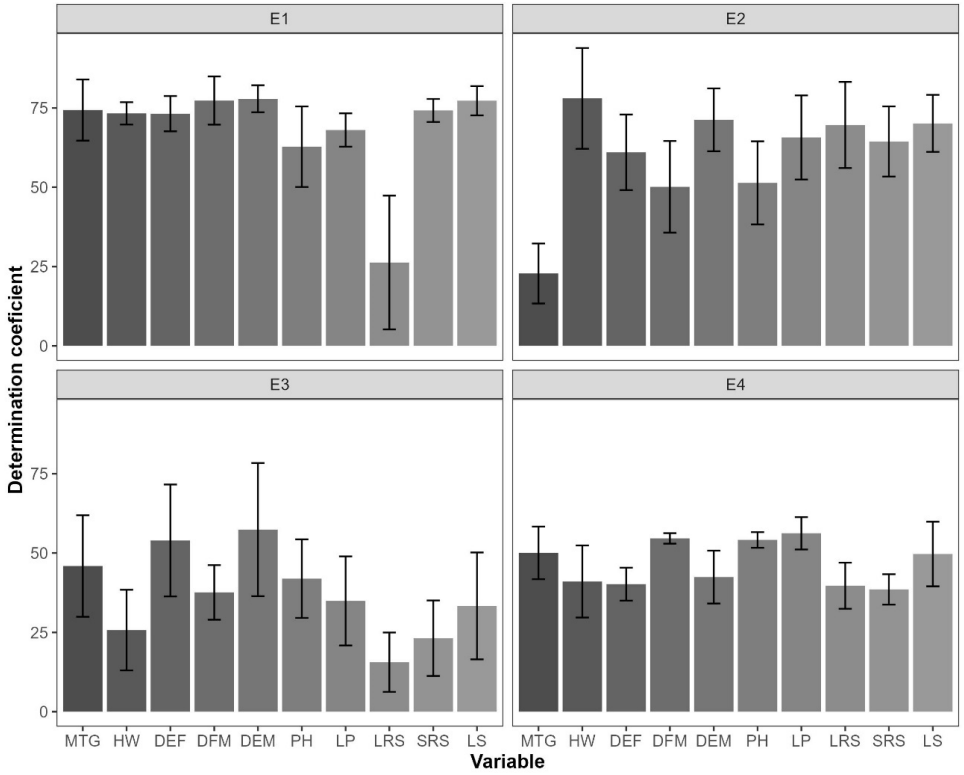

Figure 2: Estimates of the coefficient of determination of grain yield prediction in white oat (*Avena sativa* L.), using PMC attributing perturbation to genotypic information. MTG = Thousand Grain Mass in grams; HW = Hectoliter Weight; DEM = Days between Emergency and Maturation; PH= percentage of lodging; GY = Grain yield in kilograms per hectare; DEF = Days from Emergence to Flowering; DFM= Days from Flowering to Maturation; PH= Plant Height; LRS= Leaf Rust Severity; SRS=Stem Rust Severity and LS= Leaf Spots; E: environments. E1 and E3: no fungicide; E2 and E4: with fungicide.

Radial basis function (RBF)

The estimation of the importance of traits in white oat based on attribute disturbance of the information of an input trait after the RBF has been established is described in Figure 4. In this table, the relative importance of each input is estimated by the technique of destructuring the information of each explanatory trait. When using this strategy, drastic reductions in the values of R^2 were observed for the most important traits and LRS for the predictive variable GY in the E1 and E4 environments. In practice, the intensity of this trait reduces genetic progress to increase grain yield. In the E2 environment, the trait that suffered the greatest reduction in R^2 was DMF, with an estimate of 44.47%. This feature increased grain yield, as more photoassimilates were produced and translocated to grains. However, late cycle cultivars tend to be more productive in relation to the initial cycle, and an increase in the amount of photoassimilates that are translocated to the grains are obtained (Silva Junior *et al.*, 2023).

The results show that the most important trait using the RBF was MTG in the E2, E3 and E4 environments, with estimates of 58.97%, 47.98% and 40.97%, respectively. In practice, MTG influences the grain yield in white oats, since the higher MTG is, the higher the GY. This justifies the results of this study for white oats in the prediction of GY.

The results obtained support the expectation about the RBF with respect to quantifying and revealing the importance of the traits using the strategy of causing disturbances from the permutations or fixation of the phenotypic values of the input traits. Our study demonstrated the ability of RNA to quantify the importance of phenotypic traits in white oats. Techniques that show the impact of interruption or disturbance in the information of a given input in the estimation of the coefficient of determination and partition of the connection weights of the ANN were presented. These techniques were effective in estimating the true importance of phenotypic traits. Therefore, there is a certain agreement between the results found by the two computational intelligence methodologies of MLP networks and RBF networks.

Importance of traits in predicting by machine learning

Table 2 shows the means of the relative contributions of the explanatory traits for grain yield prediction by esti-

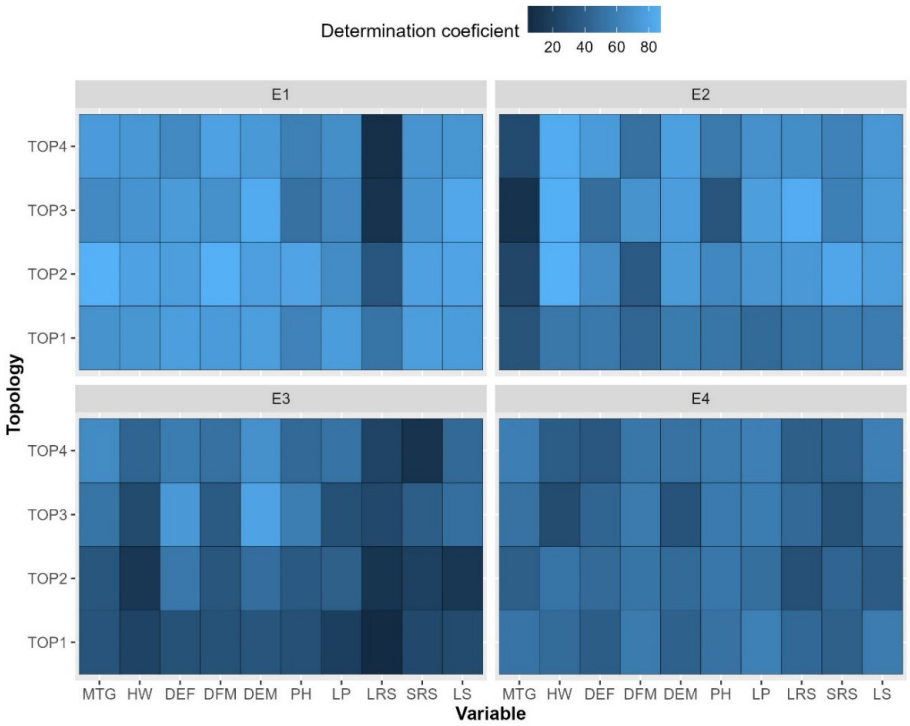

Figure 3: Estimates of the coefficient of determination in different topologies for predicting grain yield in white oat (*Avena sativa* L.), using PMC attributing disturbance to genotypic information. MTG = Thousand Grain Mass in grams; HW = Hectoliter Weight; DEM = Days between Emergency and Maturation; PH= percentage of lodging; GY = Grain yield in kilograms per hectare; DEF = Days from Emergence to Flowering; DFM= Days from Flowering to Maturation; PH= Plant Height; LRS= Leaf Rust Severity; SRS=Stem Rust Severity and LS= Leaf Spots; Topology- TOP1: Multilayer Perceptron with (10-11-1); TOP2: Multilayer Perceptron (10-11-11-1); TOP3: Multilayer Perceptron (10-11-11-11-1); TOP4: Multilayer Perceptron (10-3-4-11-1).

mating the minimum squared error increment percentage (SEIP), which is constructed by swapping the values of each trait in the dataset and comparing the results with the predictions using the original nonpermuted dataset of the traits. In this case, unlike the strategy used for the computational intelligence methodologies of the MLP and RBF networks, for which a lower value R^2 indicated a greater importance of a given trait for the model, in the machine learning approach, the importance of the explanatory trait is related to the estimation of the average decrease in the precision of the model through the SEIP. Thus, the higher this estimate is, the greater the importance of the trait.

Based on Table 2, the traits that obtained the highest SEIP estimate in all machine learning methodologies in relation to environments without fungicides were LRS, HW, PH, and MTG in E1 and DEF, SRS, and LRS in E3. The trait that was more efficient in these environments was LRS. This justifies that this trait can be used in the indirect selection process when the target prediction variable is GY. For environments with fungicides, the most important traits were MTG, DFM, PH, and LRS in E2 and DEF, DFM, DEM, and LRS in E4. For the environment with fungicide, the traits DFM and LRS proved to be efficient in estimating the prediction of grain yield in white oat.

The random forest and bagging methodologies were coincident in quantifying the same explanatory traits. A similar result was reported by (Silva Junior *et al.*, 2021). Regarding the boosting procedure, there were discrepancies in the results. On the other hand, this procedure was more consistent in terms of trait prediction. In this procedure, to estimate the importance of a trait using GY as the predictive target, the traits MTG, HW, PH, and LRS in E1 and MTG, DEF and LRS in E3 stood out in the environments without fungicides. For the fungicide environments, the important traits were MTG, DFM, PH, LRS, DEF, DFM, DEM, and LRS. When using the boosting procedure, the trait that stood out in all environments was LRS. This justifies that this trait can be used to predict GY in white oats.

The bagging technique involves generating several distinct training sets from the original dataset. The final predictions are calculated by averaging all generated predictions. This is useful for decision tree and artificial neural network techniques that are sensitive to small changes in training data (Song *et al.*, 2021).

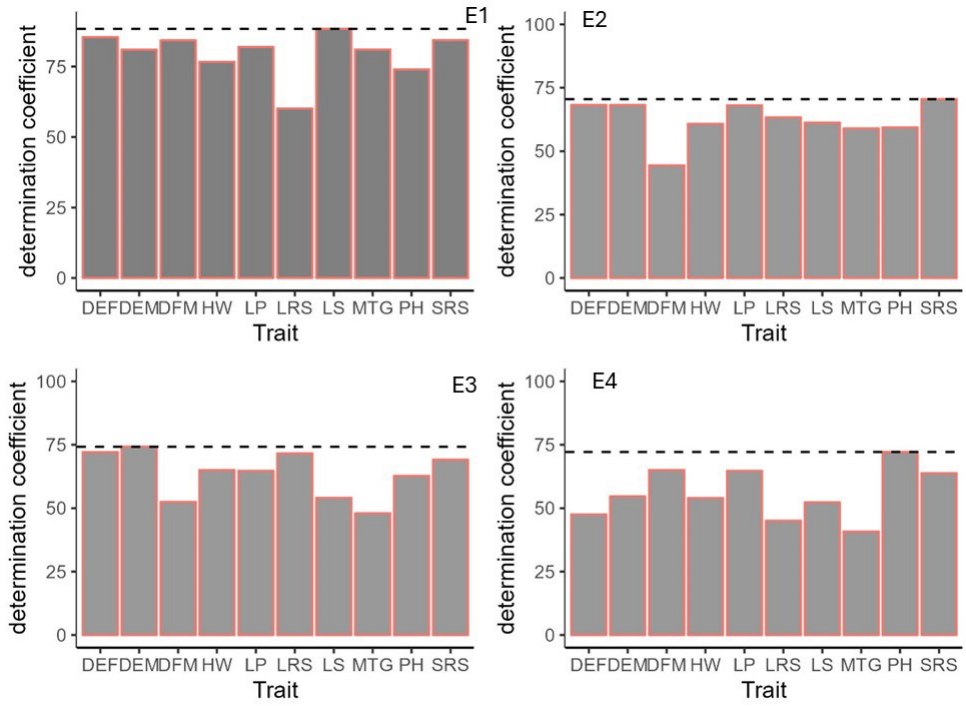

Figure 4: Coefficient estimates for determining grain yield prediction in white oat (*Avena sativa* L.) using the RBF attributing perturbation to genotypic information. MTG = Thousand Grain Mass in grams; HW = Hectoliter Weight; DEM = Days between Emergency and Maturation; PH= percentage of lodging; GY = Grain yield in kilograms per hectare; DEF = Days from Emergence to Flowering; DFM= Days from Flowering to Maturation; PH= Plant Height; LRS= Leaf Rust Severity; SRS=Stem Rust Severity and LS= Leaf Spots; E: environments. E1 and E3: no fungicide; E2 and E4: with fungicide.

Importance of traits in reduced models for predictions using the ML approach

Machine learning

The bagging biometric technique, which led to the best GY prediction results and provided information regarding the importance of predictors, is considered here. The average estimate of the relative contributions of the explanatory traits for grain yield prediction in white oat using the bagging technique after eliminating auxiliary traits of smaller relative contributions in four environments with and without fungicide application is shown in Table 3. The choice of the bagging technique was based on the estimate of the coefficient of determination (Table 1), which was greater than 90%, and the elimination of auxiliary traits of the smallest relative contributions, as shown in Table 2.

The importance of predictors through the elimination of auxiliary traits of smaller relative contributions was quantified in several ways. First, only one of the predictor traits (DFM, LP, PH, and LP) in E1, E2, E3, and E4, respectively, was eliminated. Then, the two traits with the least contribution were eliminated. Finally, the SRS and LS traits, which showed a lower estimate of the squared error increment percentage in all environments, were eliminated.

After eliminating auxiliary traits with smaller relative contributions, the maximum estimate of the coefficient of determination was similar when all auxiliary traits were used to predict GY (Tables 1 & 3).

The literature has highlighted machine learning techniques as efficient tools in quantifying the relative importance of traits in view of their simplicity, the nonuse of assumptions about the distribution of explanatory traits, and their robustness in relation to quantity, redundancy and environmental influences (Tan *et al.*, 2014; Beucher *et al.*, 2019; Silva Junior *et al.,* 2021). Furthermore, such techniques do not require an inheritance specification model and can account for nonadditive effects without increasing the number of covariates in the model or the computation time (González-Recio *et al.*, 2011). The bagging technique shows good predictive performance in practice; it works well for multidimensional problems and can be used with output from multiple classes, categorical predictors, and unbalanced problems (Gregorutti *et al.*, 2017). Satisfactory results of trait selection using the bagging and random forest algorithms in the presence of correlated predictors were reported by (Ferreira *et al.*, 2017). Discriminatory power, redundancy, precision, and complexity can influence the indices or statistics used to quantify the importance of auxiliary traits in predicting a main trait.

Genetic improvement for desired traits in different crops has been a time-consuming, laborious and expensive process. Breeders study generations of plants and identify and modify desired genetic traits as they assess how traits are expressed in offspring (Ferreira *et al.*, 2017). The application of computational intelligence and machine learning to identify ideal sets of observable traits (phenotypes) can allow informed decisions and yield highly relevant results in breeding programs. In addition, these methodologies can help predict auxiliary traits with the best performance under different agricultural management practices.

Table 2: Average estimate of the relative contributions of the explanatory traits for grain yield prediction in white oat using a machine learning approach, in four environments corresponding to without and with fungicide application

VA	E1			E2			E3			E4		
	BA	**RF**	**BO**	**BA**	**RF**	**BO**	**BA**	**RF**	**BO**	**BA**	**RF**	**BO**
MTG	**7.58**	**7.94**	**12.37**	**10.47**	**9.84**	**10.89**	1.04	1.53	**4.49**	3.55	3.21	3.75
HW	**10.11**	**10.68**	**15.29**	2.19	2.23	6.57	2.2	1.75	3.51	3.83	4.44	3.93
DEF	3.29	2.42	7.55	6.85	5.79	6.73	**3.58**	**3.5**	**4.60**	**11.46**	**11.49**	**9.18**
DFM	1.59	2.21	2.97	**16.94**	**16.84**	**12.25**	0.8	-0.4	4.22	**5.57**	**4.68**	**4.82**
DEM	3.46	3.1	6.35	6.44	6.06	6.28	2.14	1.86	3.43	**6.12**	**5.95**	**5.17**
PH	**10.74**	**10.3**	**9.65**	**10.01**	**8.29**	**9.32**	-0.93	-0.24	2.72	0.8	-0.45	1.02
LP	2.83	2.49	5.94	1.36	1.08	2.79	3.04	3.04	2.96	0.36	-0.66	0.89
LRS	**20.87**	**20.1**	**29.59**	**9.27**	**9.29**	**16.05**	**9.91**	**10.91**	**12.29**	**4.19**	**4.58**	**7.02**
SRS	7.32	7.76	5.60	3.09	2.25	3.65	**3.52**	**3.97**	**3.71**	0.8	1.62	2.04
LS	3.11	3.67	4.69	3.62	2.91	3.74	3.22	2.95	3.30	3.99	3.49	3.59

MTG = Thousand Grain Mass in grams; HW = Hectoliter Weight; DEM = Days between Emergency and Maturation; PH= percentage of lodging; GY = Grain yield in kilograms per hectare; DEF = Days from Emergence to Flowering; DFM= Days from Flowering to Maturation; PH= Plant Height; LRS= Leaf Rust Severity; SRS=Stem Rust Severity and LS= Leaf Spots; FA: random forest; BA: Bagging; BO: Boosting; VA: auxiliary variable; E: environments. E1 and E3: no fungicide; E2 and E4: with fungicide.

Table 3: Estimate of the coefficient of determination for the training set, in four environments corresponding to the data set of experiments without and with fungicide in two agricultural years, to predict the grain yield in white oat (Avena sativa L.) utilizing the bagging technique

Predictors	E1	E2	E3	E4
R^2 (T=10)	**94.61**	**93.89**	**92.70**	**92.98**
Deleted	DFM	LP	PH	LP
R^2 (T=9)	94.85	94.34	92.83	93.05
Deleted	DFM, LP	LP, HW	PH, DFM	LP, PH
R^2 (T=8)	94.26	93.50	92.03	93.11
Deleted	SRS, LS	SRS, LS	SRS, LS	SRS, LS
R^2 (T=8)	94.95	94.40	91.74	92.84

HW = Hectoliter Weight; LP= percentage of lodging; DEF = Days from Emergence to Flowering; DFM= Days from Flowering to Maturation; PH= Plant Height; SRS=Stem Rust Severity and LS= Leaf Spots; E: environments. E1 and E3: no fungicide; E2 and E4: with fungicide; R^2: coefficient of determination; T: traits.

We compared different approaches to selecting or discarding traits that have been recently proposed to identify relevant predictive variables within a regression problem. Furthermore, we included in our comparison a traditional method that aims to find a small subset of important traits with optimal predictive performance in the white oat crop. It is noteworthy that the traits used in this study are difficult to obtain, and their evaluation can be costly if there is a greater number of genotypes to be evaluated. In this context, the study of the most important traits in the prediction becomes necessary since it is possible to reduce physical efforts, costs, use of labor, and time in the experimentation (Ferreira *et al.*, 2017).

Therefore, our study presents the performance of some methodologies to assess the relative contributions of each variable through computational intelligence and machine learning in white oat cultures. Thus, the approach to estimate the effect of explanatory traits on genetic improvement has successfully identified the true importance of each trait, including those that exhibit strong and weak correlations with the main trait, which in our case is grain yield.

Methodologies based on machine learning and computational intelligence do not depend on stochastic information and tend to be more efficient, while conventional methodologies depend on the normal distribution of phenotypic traits. Furthermore, in machine learning and computational intelligence methodologies, no assumptions about the model are made, and complex factors in predictive models can be captured. In machine learning, a priori knowledge of prediction is not needed if the data produce these effects, and no assumptions are made about the distribution of phenotypic values (Sousa *et al.*, 2020). Machine learning algorithms have the advantage of modeling data nonlinearly and nonparametrically (Osco *et al.*, 2020). Unlike many traditional statistical methods, these algorithms are built with the advantage of handling noisy, complex, and heterogeneous data (Osco *et al.*, 2020). Researchers now have the ability to identify the individual and interactive contributions of predictor traits to the white oat crop using artificial intelligence and machine learning.

CONCLUSION

Computational intelligence and machine learning methodologies were used to quantify the importance of explanatory traits in predicting white oat grain yield. The model with only one hidden layer was efficient in determining the relative importance of variables in white oat. The traits indicated to assist in decision-making are plant height, leaf rust severity, and lodging percentage. The R^2 ranged from 30.14%-96.45% and 10.57%-94.61% for computational intelligence and machine learning, respectively.

A high estimate of the coefficient of determination was obtained using the bagging technique, which was higher than that of the other approaches. Simpler models, excluding predictors, are as efficient as more complex models, indicating that quantifying the importance of predictors is important to minimize costs, ensuring the same level of efficiency as that of the predictive models.

ACKNOWLEDGMENT, FINANCIAL SUPPORT ABD FULL DISCLOSURE

The authors would like to thank the National Council for Scientific and Technological Development (CNPq) and Coordination for the Improvement of Higher Education

Personnel for the financial support. The present study was partly financed by the Coordination for the Improvement of Higher Education Personnel, Brazil (CAPES), Financial Code 001. The authors declare that they have no conflict of interest. The datasets used were analyzed during the current study available from the corresponding author on reasonable request.

REFERENCES

Beck MW (2018) NeuralNetTools: Visualization and analysis tools for neural networks. Journal of Statistical, 85:01-20.

Beucher A, Møller AB & Greve MH (2019) Artificial neural networks and decision tree classification for predicting soil drainage classes in Denmark. Geoderma, 352:351-359.

Conab - Companhia Nacional de Abastecimento. Available at: <https://www.conab.gov.br/>. Accessed on June 10th, 2022.

Corazza T, Carvalho IR, Silva JAG, Szareski VJ, Segatto TA, Port ED, Loro MV, Almeida HCF, Oliveira AC, Maia L & Souza VQ (2021) Genetic parameters and multi-trait selection of white oats for forage. Genetics and Molecular Research, 20:GMR18451.

Costa WGD, Barbosa IP, de Souza JE, Cruz CD, Nascimento M & de Oliveira ACB (2021) Machine learning and statistics to qualify environments through multi-traits in Coffea arabica. PLoS One, 16:e0245298.

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

Ferreira MG, Azevedo AM, Siman LI, Silva GH, Carneiro CS, Alves FM, Delazari FT, Silva DJH & Nick C (2017) Automation in accession classification of *Brazilian Capsicum* germplasm through artificial neural networks. Scientia Agricola, 74:203-207.

Goh ATC (2005) Back-propagation neural networks for modeling complex systems. Artificial Intelligence in Engineering, 9:143-151.

Ghani IMM & Ahmad S (2010) Stepwise Multiple Regression Method to Forecast Fish Landing. Procedia - Social and Behavioral Sciences, 8:549-554.

González-Camacho JM, Campos G, Pérez P, Gianola D, Cairns JE, Mahuku G, Babu R & Crossa J (2012) Genome-enabled prediction of genetic values using radial basis function neural networks. Theoretical and Applied Genetics, 125:759-771.

González-Recio O & Forni S (2011) Prediction across the genome of discrete traits using Bayesian regressions and machine learning. Genetics Selection Evolution, 43:47.

Gregorutti B, Michel B & Saint-Pierre P (2017) Correlation and variable importance in random forests. Statistics and Computing, 27:659-678.

Mukaka MM (2012) Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Medical Journal, 24:69-71.

Osco LP, Ramos APM, Pinheiro MMF, Moriya EAS, Imai NN, Estrabis N, Lanczyk F, Araujo FF, Liesenberg V, Jorge LAC, Li J, Ma L, Gonçalves WN, Junior JM & Creste JE (2020) A machine learning framework to predict nutrient content in valencia-orange leaf hyperspectral measurement. Remote Sensing, 12:906.

Parmley KA, Higgins RH, Ganapathysubramanian B, Sarkar S & Singh AK (2019) Machine learning approach for prescriptive plant breeding. Scientific Reports, 9:17132.

Rosado RDS, Cruz CD, Barili LD, de Souza Carneiro JE, Carneiro PCS, Carneiro VQ, da Silva JT & Nascimento M (2020) Artificial Neural Networks in the Prediction of Genetic Merit to Flowering Traits in Bean Cultivars. Agriculture, 10:638.

Sant'Anna IC, Ferreira RADC, Nascimento M, Carneiro VQ, Silva GN, Cruz CD, Oliveira MS & Chagas FEO (2019) Multigenerational prediction of genetic values using genome-enabled prediction. Plos One, 14:e0210531.

Sant'Anna IC, Tomaz RS, Silva GN, Nascimento M, Bhering LL & Cruz CD (2015) Superiority of artificial neural networks for a genetic classification procedure. Genetics and Molecular Research, 14:9898-9906.

Sant'Anna IC, Silva GN, Nascimento M & Cruz CD (2020) Subset selection of markers for the genome-enabled prediction of genetic values using radial basis function neural networks. Acta Scientiarum-Agronomy, 43:e46307.

Silva Junior AC, Sant'Anna IC, Silva GN, Cruz CD, Nascimento M, Lopes LB & Soares PC (2023) Computational intelligence and machine learning to study the importance of characteristics in flood-irrigated rice. Acta Scientiarum-Agronomy,45:e57209.

Silva Júnior AC, Silva MJ, Cruz CD, Sant'Anna IC, Silva GN, Nascimento M & Azevedo CF (2021) Prediction of the importance of auxiliary traits using computational intelligence and machine learning: A simulation study. Plos One, 21:a920715476.

Silva GN, Tomaz RS, Sant'anna IC, Nascimento M, Bhering LL & Cruz CD (2014) Neural networks for predicting breeding values and genetic gains. Scientia Agricola, 71:494-498.

Skawsang S, Nagai M, Nitin K & Soni P (2019) Predicting rice pest population occurrence with satellite-derived crop phenology, ground meteorological observation, and machine learning: A case study for the Central Plain of Thailand. Applied Sciences, 9:4846.

Song H, Liu A, Li G & Liu X (2021) Bayesian bootstrap aggregation for tourism demand forecasting. International Journal of Tourism Research, 01-14.

Sousa IC, Nascimento M, Silva GN, Nascimento ACC, Cruz CD, Fonseca F, Almeida DP, Pestana KN, Azevedo CF, Zambolim L & Caixeita ET (2020) Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. Scientia Agricola, 78:01-8.

Tan K, Li E, Du Q & Du P (2014) An efficient semi-supervised classification approach for hyperspectral imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 97:36-45.

Kim KS, Tinker NA & Newell MA (2014) Improvement of oat as a winter forage crop in the Southern United States. Crop Science, 54:1336-1346.

McCartney D, Fraser J & Ohama A (2008) Annual cool season crops for grazing by beef cattle. Canadian Journal of Animal Science, 88:517-533.

Sharma P, Leigh L, Chang J, Maimaitijiang M & Caffé M (2022) Above-Ground Biomass Estimation in Oats Using UAV Remote Sensing and Machine Learning. Sensors, 22:601.