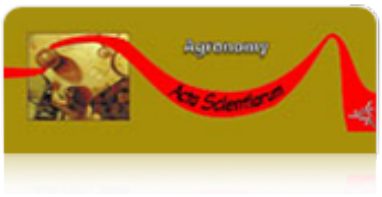

Acta Scientiarum. Agronomy
ISSN: 1679-9275
ISSN: 1807-8621
Editora da Universidade Estadual de Maringá - EDUEM

Silva, Antônio Carlos da; Sant'Anna, Isabela Castro; Silva, Gabi Nunes; Cruz, Cosme Damião; Nascimento, Moysés; Lopes, Leonardo Bhering; Soares, Plínio César

Computational intelligence to study the importance of characteristics in flood-irrigated rice

Acta Scientiarum. Agronomy, vol. 45, e57209, 2023
Editora da Universidade Estadual de Maringá - EDUEM

DOI: https://doi.org/10.4025/actasciagron.v45i1.57209

Available in: https://www.redalyc.org/articulo.oa?id=303075473020

Acta Scientiarum

http://www.periodicos.uem.br/ojs/
ISSN on-line: 1807-8621
Doi: 10.4025/actasciagron.v45i1.57209

Computational intelligence to study the importance of characteristics in flood-irrigated rice

Antônio Carlos da Silva Júnior[1*], Isabela Castro Sant'Anna[2], Gabi Nunes Silva[3], Cosme Damião Cruz[1], Moysés Nascimento[4], Leonardo Bhering Lopes[1] and Plínio César Soares[5]

[1]Departamento de Biologia Geral, Universidade Federal de Viçosa, Av. PH Rolfs, s/n, 36570-900, Viçosa, Minas Gerais, Brazil. [2]Centro de Seringueira e Sistemas Agroflorestais, Instituto Agronômico, Votuporanga, São Paulo, Brazil. [3]Departamento Acadêmico de Matemática e Estatística, Universidade Federal de Rondônia, Ji-Paraná, Rondônia, Brazil. [4]Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil. [5]Empresa de Pesquisa Agropecuária de Minas Gerais, Viçosa, Minas Gerais, Brazil. *Author for correspondence. E-mail: antonio.silva.c.junior@gmail.com

ABSTRACT. The study of traits in crops enables breeders to guide strategies for selecting and accelerating the progress of genetic breeding. Although the simultaneous evaluation of characteristics in the plant breeding programme provides large quantities of information, identifying which phenotypic characteristic is the most important is a challenge facing breeders. Thus, this work aims to quantify the best approaches for prediction and establish a network of better predictive power in flood-irrigated rice via methodologies based on regression, artificial intelligence, and machine learning. Multiple regression, computational intelligence, and machine learning were used to predict the importance of the characteristics. Computational intelligence and machine learning were notable for their ability to extract nonlinear information from model inputs. Predicting the relative contribution of auxiliary characteristics in rice through computational intelligence and machine learning proved to be efficient in determining the relative importance of variables in flood-irrigated rice. The characteristics indicated to assist in decision making are flowering, number of grains filled by panicles and length of panicles for this study. The network with only one hidden layer with 15 neurons was observed to be efficient in determining the relative importance of variables in flooded rice.

Keywords: *Oryza sativa* L.; multiple regression; computational intelligence; machine learning.

Received on December 23, 2020.
Accepted on April 16, 2021.

Introduction

Plant breeding is effective in increasing the productivity of crops. The primary objective of plant breeding is to increase the frequency of good alleles in plant populations such that superior crops are developed with high productivity, resistance to diseases and pests, tolerance to abiotic stresses, and superior adaptation to environments (Yu, Campbell, Zhang, Walia, & Morota, 2019).

In general, productivity prediction is performed using multiple linear regression. Although interesting, multiple regression models have some limitations, such as the size of the sample data. Specifically, when the observation number is less than the number of parameters, it is not possible to obtain the estimates using the usual estimation methods. Additionally, such models do not allow the adjustment of complex nonlinear relationships possibly existing in some data sets. Artificial neural networks (ANNs) provide an interesting alternative because they can capture nonlinear relationships between predictors and responses (Gianola, Okut, Weigel, & Rosa, 2011; Skawsang, Nagai, Nitin, & Soni, 2019) and ignore assumptions in the data sets.

The application of artificial intelligence, such as ANN, allows the capture of nonlinear effects among the data set and has been used in studies of prediction in plant breeding (Silva et al., 2014; Silva et al., 2017; Sant'anna et al., 2019). However, although ANNs are powerful predictive tools compared to conventional models, such as multiple linear regression (Paruelo & Tomasel, 1997; Olden & Jackson, 2002; Beck, 2018), they have the limitation of neglecting to quantify the importance of the variables.

Quantifying the importance of variables for prediction in breeding programmes allows for faster progress, selecting and predicting characteristics that have low heritability and/or measurement difficulty. Although simultaneous evaluation of characteristics provides a wide variety of information, identifying which predictor variable is most important is a challenge for breeders (Parmley, Higgins, Ganapathysubramanian, Sarkar, &

Singh, 2019). The quantification of the importance of variables can be performed by ANNs through algorithms such as Goh (1995), who proposed a modification in Garson's (1991) algorithm that consists of partitioning the neural network connection weights to determine the relative importance of each variable entering the network.

Other interesting alternatives for studies of the prediction and importance of variables are methodologies based on machine learning, such as decision trees (Beucher, Møller, & Greve, 2019; Parmley et al., 2019) and their refinements, such as *bagging* (Degenhardt, Seifert, & Szymczak, 2019), *random forest*, and *boosting* (Degenhardt et al., 2019). Such methodologies allow good predictions and the importance of the characteristics to be obtained through measures based, for example, in the index of Gini and Entropy (Hastie, Tibshirani, & Friedman, 2009). These methodologies enable the quantification of the impact of the disruption or disturbance of the input information on the estimate of the determination coefficient.

Methodologies based on regression, artificial intelligence, and machine learning have been used successfully in a prediction study. Parmley et al. (2019) evaluated the phenotypic characteristics of high dimensionality soybeans through a machine learning approach to predict seed yield regarding the prescriptive development of cultivars for agricultural practices. Skawsang et al. (2019) applied such methodologies to predict the population of insect pests using climatic and phenological factors of the host plant. However, there are no studies in the literature related to yield prediction and verification of the importance of variables for grain yield in rice culture. Unlike the methods of regression, artificial intelligence and machine learning do not make any prior assumptions about the data structure, in which it captures linear and nonlinear dependencies between the predictor and the response variables, making it a suitable tool for the researcher.

Given the above, this work aims to i) predict grain yield, grain length and width ratio, and panicle length in flood-irrigated rice through regression, artificial intelligence, and machine learning methodologies; ii) quantify the best approaches to prediction; and iii) establish a network of better predictive power in flood-irrigated rice.

Material and methods

Description of the experiment

The experiments were carried out in the State of Minas Gerais, Brazil, in the experimental fields of the Agricultural Research Corporation of Minas Gerais (EPAMIG), in the city of Leopoldina (21°31'48.01'' S, 42°38'24'' W), Lambari (21°58'11.24'' S, 45°20'59.6'' W), and Janaúba (15°48'77'' S, 43°17'59.09'' W). Seventy-five genotypes of flood-irrigated rice belonging to the flood-irrigated rice breeding programme were evaluated in the agricultural year 2012/2013. The design was randomized blocks with three replications.

The evaluated characteristics were grain yield (GY, kg ha^{-1}), panicle length (LP, cm), and grain length-to-width ratio (LGW), which were used as response variables and the others as explanatory variables (inputs), that is, plant height (HP, cm), flowering (FL, days), lodging (LO), number of full grains per panicle (GP), percentage of full grains (FG, %), tillering (TI), length (GL, mm), width (GW, mm) and thickness (GT, mm) of grains, and weight of 100 grains (WG, g). They were used to compose artificial neural networks of genotypes of flood-irrigated rice in the State of Minas Gerais.

Methodologies for predicting and verifying the importance of characteristics

Multiple regression

Multiple regression, through the stepwise strategy (Ghani & Ahmad, 2010), was used to predict the variable responses to grain yield, panicle length, and grain length-to-width ratio as a function of the other measured variables and was considered to be explanatory.

The adopted model is represented by Equation 1:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon, \quad (1)$$

where y is the response variable (grain yield, panicle length or grain length-to-width ratio), x_1 a x_k are the explanatory variables, 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 variable.

The description of R^2 is found in Equation 2:

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

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

Artificial intelligence

For better network efficiency, before training and validation, the data were normalized in the range between -1 and 1. The training data set, in each location, was established by 2/3 of the phenotypic information, using the strategy of aggregating information from two of the three repetitions for training and the information from the other repetition used as a validation set. In this cross-validation strategy, individuals from each repetition participated at least once in the validation data set in cross-validation (k-fold) k = 3 partitions.

Perceptron Multilayer - PMC

The maximum number of training seasons was set at 5,000; the mean square error (MSE), as a criterion to stop processing the network, was defined as 1.0×10^{-3}. All trained networks had a neuron in the output layer and a single hidden layer, with 15 neurons. The sigmoid tangent activation function was used in the hidden layer, and the training algorithm was *Bayesian regulation backpropagation*. To quantify the efficiency of the prediction R^{2*}.

Importance of variables

To quantify the importance of variables through the PMC network, two techniques were used. The first is based on the Garson (1991) algorithm modified by Goh (1995) (AG), which consists of partitioning the neural network connection weights to determine the relative importance of each input variable within the network. This algorithm describes the relative magnitude of the importance of the descriptors (predictor) in their connection with outcome variables through the dissection of synaptic weights from the neural network. In the second technique, the importance of variables (inputs) is assessed through the impact of the disruption or disturbance of the information of a given input on the estimation of the determination coefficient. Thus, this importance is estimated by exchanging information or by making constant the phenotypic values shown for each variable and verifying changes in the estimates of the R^2. When we disturb the values of a variable and R^2 decreases, there is an indication that the input variable is important about the others for purposes of prediction with the network already established.

Radial Base Function network – RBF

The radial base function network is characterized by having only one hidden layer and making use of the Gaussian activation function (Cruz & Nascimento, 2018). The structure of the RBF to better predict grain yield, panicle length, and grain length-to-width ratio was established with 10 to 30 neurons (increased by 2, with each processing), and the radius established between 5 and 15 increased by 0.5. The efficiency of the prediction was measured by the R^2, and the relative importance of each entry was measured by the technique of destroying the information of each explanatory variable, as already described for PMC.

Machine learning

To predict grain yield, panicle length, and grain length-to-width ratio and quantify the importance of variables through a machine learning approach, a decision tree and its refinements were used, *random forest*, *bagging,* and *boosting*. The R^2 measured the quality of the predictive model fit, and information from the minimum quadratic error (MSE) was used to quantify the importance of variables in flood-irrigated rice crops. The minimum square error was estimated as described in Equation 3 below:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad (3)$$

where y_i and $\hat{y}_i$ correspond to the observed and predicted values of observation in genotype i, respectively, and n is the total number of observations (variable, depending on the environment analysed).

In these techniques, the importance of the explanatory variable is the quantification of the mean decrease in the prediction precision, which consists of the estimate of the percentage of increment of minimum square error (IMSE), which is constructed when we exchange the values of each variable of the data set and are

compared with the prediction of the original unchanged data set for the variable. Analogous to the regression analysis, it is the average increase of the squares of the residuals of the data set when the variable is exchanged (Li & Zhan, 2019). Higher values of IMSE represent the importance of the highest variable. For better efficiency of the prediction estimate of the importance of variables, 5,000 trees were generated.

The analyses were performed with the aid of R software using the NeuralNetTools (Beck, 2018) and Genes (Cruz, 2016) packages, which use an interface with MATLAB software (Matlab, 2016).

Results and discussion

Prediction by different approaches

The estimate of R^2 for all methodologies using the explanatory variables to predict grain yield (GY), panicle length (PL), and grain length and width ratio (LGW) in flood-irrigated rice is shown in Figure 1. Based on Figure 1, it is possible to compare and define the variables that proved to be most efficient for the prediction of GY, PL, and LGW. Higher values of this estimate indicate that the target prediction variable has a better adjustment than the other explanatory variables (Roy & Roy, 2008; Hassanzadeh, Ghavami, & Kompany-Zareh, 2015). Among the methodologies used in this study, it was found that multiple regression showed a lower estimate of R^2 (Figure 1) for the same variable, indicating the existence of nonlinear associations between the explanatory variables not considered in the model. Artificial intelligence and machine learning methodologies, in turn, stood out for their ability to extract nonlinear information from model inputs (Parmley et al., 2019; Skawsang et al., 2019), as seen in Figure 1. Other authors have already highlighted the abilities of neural networks to better capture nonlinear relationships when compared to conventional methodologies (Silva et al., 2014; Sant'anna et al., 2016).

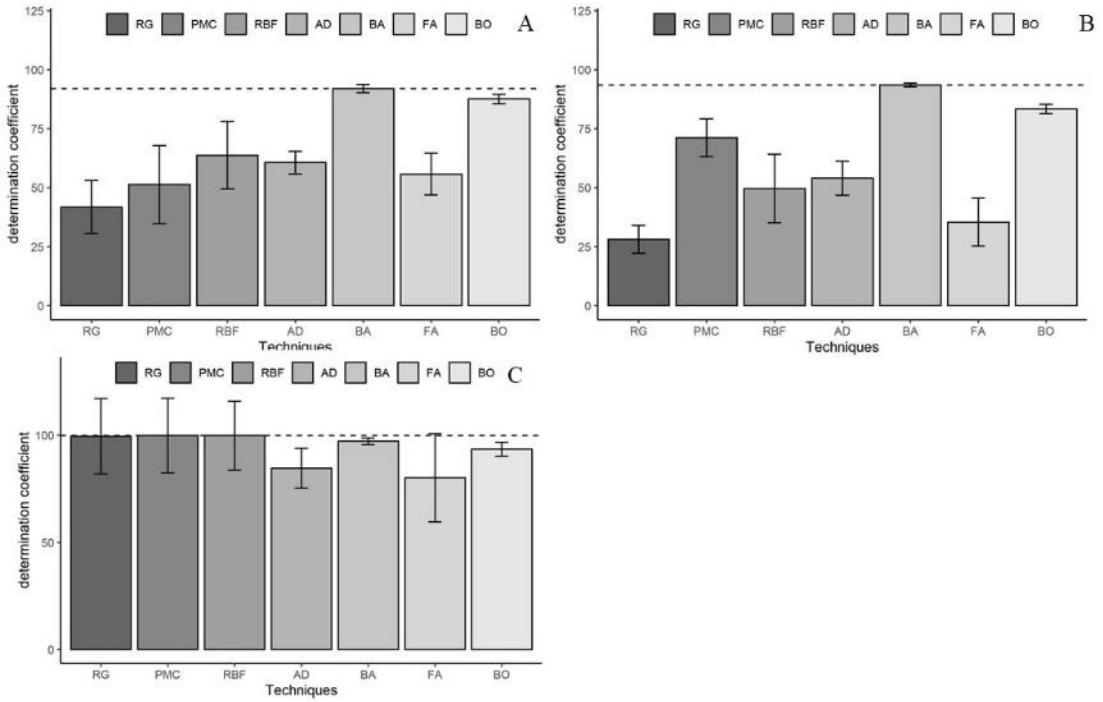

Figure 1. Maximum estimate of the coefficient of determination in three environments to predict grain yield (GY), panicle length (PL), and grain length and width ratio in flood-irrigated rice (LGW). A: panicle length; B: grain yield; C: grain length-to-width ratio; RG: multiple regression; PMC: multilayer perceptron; RBR: radial base network; AD: decision tree; FA: *random forest*; BA: *bagging*; BO: *boosting*.

The results obtained by different approaches show that there was a discrepancy between the maximum estimate of R^2 for all predictive variables in the same environments (Figure 1). The artificial intelligence approach in the Leopoldina environment provided a higher estimate for the predictive variables PL and GY in the RBF procedure, 83.44 and 78.90%, respectively. The GY response variable had the best estimate of R^2 in the Lambari and Janaúba environments in the PMC network with only one neuron in the output layer and a single hidden layer (Figure 1). In the Leopoldina and Lambari environments, for the LGW response variable, a maximum estimate of R^2 was approximately 100% by multiple regression and artificial intelligence approaches.

On the other hand, it is variable in Janaúba, with a maximum estimate of 62%. The differences in the results obtained in these analyses indicate that the environment influences the estimation of R^2 and, consequently, the cause and effect relationships between the response variable and the set of explanatory variables.

Machine learning approaches proved to be more efficient than other approaches (Figure 1). There was a low estimate of R^2 for the predictive variable GY in the Janaúba environment in the *random forest* procedure, which corresponds to 18.57%. This result is inferior to all the approaches used in this study. In this same environment, but for bagging procedures, the estimate of R^2 was 94.76%. High estimates of R^2 above 80%) were obtained using machine learning methodologies by the procedures *bagging* and *boosting* for all predictive variables (Figure 1). The decision tree (AD) and *random forest* methodologies did not stand out from the other machine learning procedures (Figure 1). Sousa et al. (2020) emphasized that the AD's low predictive accuracy can be improved using ensemble methods such as *bagging*, *random forest*, and *boosting*. These strategies combine multiple AD to reduce the variability.

Random forests and *bagging* these methods have good predictive performances in practice; they work well for high-dimensional problems and can be used with multiclass output, categorical predictors, and imbalanced problems (Gregorutti, Michel, & Saint-Pierre, 2017). This author had satisfactory result variable selection with the *random forests* algorithm in the presence of correlated predictors.

When the variables are correlated, the simple correlation coefficient produces incomplete information. This is because a high correlation between two variables may have resulted from a third or a group of variables over another variable. Traditional methods, as well as path analysis, decompose into direct and indirect effects on the main variable, and logistic regression becomes unstable in the presence of high correlations. Multicollinearity is caused by the high correlation between the variables, which provides a problem of lack of adjustment of the model that affects the estimates of the parameters. In the literature, the ability of RNA to circumvent the problem of multicollinearity has already been highlighted (Cruz & Nascimento, 2018). These authors presented an application in which a response variable is predicted through five explanatory variables. By including a sixth explanatory variable, which would assume the same values as the fifth variable, it did not affect the accuracy of the ANN - Adaline in any way. However, they reinforce that in the classic multiple linear regression approach, there would be no solution, since there would be two columns, in the prediction matrix X, linearly dependent, so that the established multicollinearity would lead to an X'X matrix without a common inverse.

The efficiency of ANNs in prediction problems, given their ability to extract relevant information from large data sets and generalize relatively inaccurate information (Porwal, Carranza, & Hale, 2003), was very well expressed by the results obtained (Figure 1). The same can be seen for methodologies based on machine learning, which are capable of handling more reduced or redundant information in the input variables (Quinlan, 1996). However, another study as important as the prediction and which is often not carried out is the identification, among the explanatory variables, those of greater importance despite constituting important information in the process of understanding the adjusted model and decision making about dimensionality reduction in future studies (Beucher et al., 2019). Thus, after the prediction analysis, the quantification of the importance of variables was performed using artificial intelligence and machine learning methods to identify, among the set of explanatory variables, those that should be prioritized and identified as auxiliary characteristics in indirect responses to selection.

Importance of variables in prediction by the artificial intelligence approach

For ease of interpretation, we will denote R^2 the quality of prediction of the methodology and R^{2*} this same quality of adjustment after the disturbance in the explanatory variable.

Multilayer Perceptron (PMC)

Neural networks tend to perform well when compared to other predictive algorithms based on machine learning (Santos, Dean, Weaver, & Hovanski, 2018). These algorithms are capable of learning from linear and nonlinear relationships in the data (Somers & Casal, 2009; Haddouche, Chetate, & Said Boumedine, 2018). It can also measure and incorporate direct effects and effects of interaction between variables in predictive models (Tsang, Cheng, & Liu, 2017).

The PMC network is widely used in the predictive process (Gedeon, Wong, & Harris, 1995; Santos et al., 2018) since the success of this network has already been shown in several research groups that have shown mathematically that, with only a single hidden layer, this network works very well with different numbers of neurons in the hidden layer (De Oña & Garrido, 2014; Santos et al., 2018).

The importance of the variables was quantified by assigning a zero value to the phenotypic information related to each variable to observe what changes would occur in the values of the R^{2*}. The results of the PMC network are shown in Table 1. It is important to note that, in this table, reductions in the values of R^{2*} after assigning zero value to the phenotypic information referring to each variable, they are indicative that this variable is important about the others for purposes of prediction with the network already established.

Table 1. Estimates of the coefficient of determination, provided by the use of the PMC, to predict grain yield, panicle length and grain length and width after disturbance (zero value assignment) in the explanatory variable values.

PL				GY				LGW			
Input	E1	E2	E3	Input	E1	E2	E3	Input	E1	E2	E3
LO	48.65	51.12	36.67	LO	8.08	24.54	5.33	LO	98.47	99.98	63.03
HP	9.07	47.62	47.92	HP	0.04	48.22	7.97	HP	99.98	99.98	62.70
GL	36.37	41.86	6.77	GL	0.52	16.01	7.54	GL	37.58	46.56	20.09
GT	37.37	47.66	28.78	PL	6.43	34.78	15.29	PL	99.97	99.97	63.00
FL	46.40	46.01	32.09	GT	7.42	22.20	13.83	GT	99.95	99.97	61.22
GW	37.38	51.31	18.46	FL	12.03	7.34	12.96	FL	99.86	99.96	58.88
GP	46.53	51.07	32.97	GW	16.30	21.61	2.12	GW	39.64	36.55	43.32
WG	47.95	51.14	20.28	GP	10.72	57.00	5.27	GP	99.98	99.98	62.96
TI	47.90	50.02	24.57	WG	17.64	17.01	9.68	WG	99.98	99.97	62.76
FG	45.27	50.26	3.73	TI	1.70	39.68	4.35	TI	99.99	99.97	62.30
GY	41.68	49.80	21.73	FG	10.81	27.50	11.17	FG	99.98	99.97	63.01
LGW	40.83	17.43	31.29	LGW	21.32	13.31	26.62	GY	99.97	99.97	63.00

LO: lodging, HP: height (cm), GL: grain length (mm), PL: panicle length (cm), GT: grain thickness (mm), FL: flowering (days), GW: grain width (mm), GP: number of filled grains per panicle, WG: weight of 100 grains (g), TI: tillering, FG: percentages of filled grains and LGW: length-to-width ratio of grains, GY: grain yield, Environment, E1: Leopoldina, E2: Lambari, E3: Janaúba.

The results in Table 1 show great discrepancies in the R^{2*} when comparing the environments with each other, which makes interpretation difficult. For the response variable LGW, it was efficient to quantify grain length and width due to the reduction in the estimate of R^{2*} as a result of the strategy of assigning a zero value to phenotypic information. It should be remembered that such changes must be seen concerning the values of the R^2 of prediction, which was approximately 100% in the environments of Leopoldina and Lambari, and Janaúba was 63% (Figure 1). For Leopoldina, when zeroing the variables, for example, HP, GL, and TI, the R^{2*} values of these variables were 0.04, 0.52, and 1.70, respectively (Figure 1). This result shows that these variables are important in predicting GY because the disturbance in their values has led to a considerable reduction in the quality of the adjustment. In Lambari, the variable that presented the highest contribution was FL. Independent of the predictive variable in PMC, with only one neuron in the output layer and a single hidden layer, they agreed to point out that the most important variables were grain width and length, given the significant falls in the values of the estimate of R^{2*} observed when zeroing the variable.

To overcome the difficulties faced when adopting PMC networks to study the importance of variables, an alternative is to use the AG algorithm, which takes into account the partitioning of the RNA connection weights to determine the relative importance of each input variable within the network. The weights that connect neurons in an ANN are partially analogous to the coefficients in a generalized linear model (Beck, 2018) so that the combined effects of weights in the model's predictions represent the relative importance of predictors in their associations with the variable of the predictor. The large number of adjustable weights in an artificial neural network makes it very flexible in modelling nonlinear effects but imposes challenges for its interpretation. In this algorithm, the numbers of neurons were used to obtain the maximum estimate of R^{2*} for a better estimate of the relative contribution of variables.

The percentages of the relative contribution estimated by the GA method are described in Table 2. In this table, for the GY response variable, the results were consistent in pointing plant height (HP), flowering (FL), and the number of full grains per panicle (GP) in terms of relative contribution. For the variable response PL, the variable with the greatest relative contribution was grain yield (GY) in the environments of Leopoldina and Lambari; however, in Janaúba, the variable that stood out was the length and width of grains. Regarding the explanatory variable LGW, the percentages of the relative contribution revealed that the variables grain length and grain width had the largest relative contribution. This result was expected since the length and width of grain variables are determinants of LGW. The results indicate that the GA approaches are efficient in quantifying the importance of variables in studies involving PMC neural networks.

Table 2. Percentages of the relative contribution estimated by the method of Garson (1991) modified by Goh (1995) of 12 variables to predict grain yield, panicle length, and grain length and width ratio in flood-irrigated rice in three environments in the State of Minas Gerais.

	GY				PL				LGW		
VP	E1	E2	E3	VP	E1	E2	E3	VP	E1	E2	E3
LO	6.50	6.00	5.57	LO	8.48	6.94	7.63	LO	6.94	7.74	7.97
HP	11.00	16.0	15.12	HP	8.30	8.99	8.19	HP	7.81	8.16	7.14
GL	8.90	6.10	5.96	GL	6.89	8.00	7.60	GL	9.84	9.26	9.23
PL	8.26	8.51	8.12	GT	8.52	8.40	8.15	PL	8.65	8.42	8.54
GT	6.80	6.02	6.00	FL	7.58	8.22	8.20	GT	8.77	9.19	10.48
FL	13.00	12.8	13.55	GW	8.30	8.64	8.20	FL	7.49	8.50	8.90
GW	6.60	5.67	6.79	GP	8.15	8.44	8.30	GW	8.00	8.60	9.29
GP	11.20	14.10	13.02	WG	8.10	8.00	8.80	GP	8.71	8.70	8.06
WG	7.42	6.60	6.80	TI	8.04	8.30	7.89	WG	8.27	7.99	8.68
TI	6.82	5.50	5.85	FG	8.75	7.89	9.48	TI	8.92	8.12	5.14
FG	7.00	6.50	6.55	GY	10.00	9.23	8.50	FG	7.99	7.92	7.62
LGW	6.50	6.20	6.67	LGW	8.89	8.95	9.06	GY	8.61	7.40	8.95

PV: predictive variable, LO: lodging, HP: height (cm), GL: grain length (mm), PL: panicle length (cm), GT: grain thickness (mm), FL: flowering (days), GW: grain width (mm), GP: number of filled grains per panicle, WG: weight of 100 grains (g), TI: tillering, FG: percentages of filled grains and LGW: length-to-width ratio of grains, GY: grain yield. Environment E1: Leopoldina, E2: Lambari, E3: Janaúba.

Radial Base Network (RBF)

The quantification of the importance of flood-irrigated rice characters by assigning a zero value to the information of an input variable after the RBF was established was performed and is described in Table 3. In this table, the values are used after causing disturbances in the input variables with the action of assigning zero value of the variable in each explanatory variable. When using this strategy of zeroing the value of the variable, drastic reductions in the values of R^{2*} were observed for the most important length (GL) and grain width (GW) variables when the target prediction variable was LGW. For other response variables, this result was very discrepant in quantifying the true importance of variables. When the explanatory variable was GY, in Janaúba, the variables that suffered the greatest reduction in R^{2*} were flowering - R^{2*} = 23.80 and weight of 100 grains (WG) - R^{2*} = 19.91; in Leopoldina, plant height variables were observed (HP) - R^{2*} = 21.26, grain width (GW) - R^{2*}= 24.83 and weight of 100 grains (WG) = 24.25; and in Lambari, the most important variable using this approach was flowering (FL) - R^{2*}= 28.43.

For the variable response PL, we observed changes in the values of R^{2*} in Leopoldina and Lambari for the variable flowering (FL) - R^{2*}= 47.77 and R^{2*}= 46.76, respectively. In Leopoldina, the percentages of full grains (FG) - R^{2*}= 25.51 also showed a drastic reduction in R^{2*}. In Lambari, lower estimates of R^{2*} were obtained for the variable weight of 100 grains (WG) - R^{2*}= 45.60. For Janaúba, the results show that the most important variables using the RBF were grain width (GW) - R^{2*}= 19.76 and weight of 100 grains (WG) - R^{2*}= 23.11.

Therefore, there is a certain agreement between the results found by the two computational intelligence methodologies of PMC networks and RBF networks.

Table 3. Estimates of the coefficient for determining the grain yield prediction, panicle length, and grain length-to-width ratio using the RBF assigning zero value to the genotype information.

	PL				GY				LGW		
VP	1	2	3	VP	1	2	3	VP	1	2	3
LO	83.20	62.80	43.61	LO	21.26	45.58	41.26	LO	99.91	99.84	65.72
HP	53.72	61.65	45.73	HP	70.88	47.39	49.52	HP	99.91	99.92	64.65
GL	70.57	60.72	44.23	GL	29.42	46.46	43.32	GL	40.93	43.11	29.84
GT	61.09	61.93	27.36	PL	38.12	40.79	36.92	PL	99.91	99.90	63.07
FL	47.77	46.76	41.69	GT	42.21	45.22	36.63	GT	99.83	99.89	62.68
GW	71.73	50.95	19.76	FL	30.67	28.43	23.80	FL	99.73	99.53	65.56
GP	59.41	62.16	29.75	GW	24.83	44.75	34.95	GW	44.98	43.19	46.46
WG	64.21	45.60	23.11	GP	40.96	44.99	34.98	GP	99.88	99.89	63.33
TI	67.71	63.58	26.15	WG	24.25	45.43	19.91	WG	99.89	99.74	62.52
FG	25.51	59.63	27.70	TI	25.58	45.95	35.04	TI	99.89	99.83	63.96
GY	54.38	56.28	31.91	FG	26.60	46.19	35.61	FG	99.84	99.78	63.69
LGW	71.29	60.59	44.11	LGW	31.42	44.27	29.73	GY	99.89	99.82	62.63

PV: predictive variable, LO: lodging, HP: height (cm), GL: grain length (mm), PL: panicle length (cm), GT: grain thickness (mm), FL: flowering (days), GW: grain width (mm), GP: number of filled grains per panicle, WG: weight of 100 grains (g), TI: tillering, FG: percentages of filled grains and LGW: length-to-width ratio of grains, GY: grain yield. Environment E1: Leopoldina, E2: Lambari, E3: Janaúba.

Importance of variables in prediction by the machine learning approach

Table 4 shows the averages of the relative contributions of the explanatory variables for predicting grain yield, panicle length, and grain length-to-width ratio by estimating the percentage of minimum square error increment (IMSE), which is constructed by exchanging the values of each variable in the data set and comparing it with the prediction of the original unix exchange data set for the variable. In this case, unlike the strategy used for the computational intelligence methodologies of PMC and RBF networks, for which lower values of R^{2*} indicated greater importance of that variable for the model, in the machine learning approach, the importance of the explanatory variable is related to the estimate of the average decrease in the accuracy of the model through IMSE so that the higher this estimate the greater the importance of the variable.

Table 4. The average estimate of the relative contributions of the explanatory variables for predicting grain yield, panicle length, and grain length-to-grain ratio in flood-irrigated rice continues using a machine learning approach in three environments in Minas Gerais.

	PL								
	BA			FA			BO		
VP	E1	E2	E3	E1	E2	E3	E1	E2	E3
LO	0	-1.13	0	0	-1.34	0	0	2.31	0
HP	10.89	7.65	-0.05	11.37	7.43	-1.27	13.24	7.96	6.31
GL	2.84	0.28	-0.04	2.97	0.62	0.5	8.33	8.20	11.51
GT	2.37	1.02	0.97	1.46	1.74	1.41	9.14	7.93	8.94
FL	-0.96	6.69	-1.13	-2.01	5.59	-2.4	3.83	4.66	1.25
GW	3.32	12.01	1.76	2.99	10.38	1.78	7.57	18.62	12.30
GP	7.68	0.8	-0.04	5.82	0.67	0.29	11.37	9.28	15.98
WG	1.71	0.87	1.23	0.61	1.74	2.35	5.07	9.82	9.49
TI	-0.16	-0.16	-0.45	0.35	1.17	-0.84	1.59	0	3.65
FG	4.12	-1.75	0.35	3.58	-1.27	-0.42	8.52	6.41	9.09
GY	-0.38	5.21	2.39	0.74	5.08	3.20	13.55	13.76	10.68
LGW	5.56	7.24	-1.37	5.85	7.15	-0.53	15.88	10.79	11.73
	GY								
	BA			FA			BO		
VP	E1	E2	E3	E1	E2	E3	E1	E2	E3
HP	2.52	2.63	-0.68	2.98	2.02	-1.27	9.93	7.98	11.2
LO	0	-0.87	0	0	-0.19	0	0	5.21	0
GL	2.39	-0.92	1.19	3.68	0.1	0.14	11.97	12.54	14.05
PL	1.07	10.33	-0.27	-0.22	10.32	-1.23	8.2	14.8	13.52
GT	0.43	-2.42	-0.47	1.49	-1.19	-2.61	7.52	6.21	7.91
FL	-2.5	5.51	-0.72	-1.72	6.67	-1.73	5.6	5.39	1.71
GW	-0.9	1.46	-3.84	-0.63	1.16	-2.25	10.91	4.59	8.28
GP	2.34	-0.41	-2.72	1.76	2.43	-3.34	13.4	15.09	10.39
WG	-0.25	1.87	1.24	2.15	1.36	1.64	10.95	9.14	11.59
TI	-0.33	0.5	-0.96	-1.98	0.6	-1.06	2.26	0.76	2.23
FG	0.64	-0.32	0.41	-1.14	1.02	1.25	5.22	7.57	8.3
LGW	0.37	1.53	-0.31	0.39	2.52	0.78	9.61	7.41	9.15
	LGW								
	BA			FA			BO		
VP	E1	E2	E3	E1	E2	E3	E1	E2	E3
LO	0	1.62	0	0	-0.7	0	0	1.75	0
HP	-0.13	-0.16	1.06	0.3	0.85	0.92	3.76	4.37	6.92
GL	18.99	18.32	18.37	19.3	17.73	20.26	25.51	27.33	22.62
PL	8.56	11.82	-0.89	8.53	12.01	-1.68	8.79	14.57	10.29
GT	1.87	2.43	-0.36	1.14	2.62	0.98	4.67	3.77	6.05
FL	3.95	3.67	1.84	1.34	3.58	0.93	2.81	1.45	0.83
GW	19.65	17.23	11.32	19.28	18.28	9.88	20.41	20.37	16.46
GP	1.58	0.46	0.81	3.59	1.82	-1.16	10.36	7.26	9.44
WG	9.52	0.21	-0.14	7.61	0.85	-0.76	9.34	6.51	8.42
TI	-0.94	-1.07	-0.72	-1.33	-0.82	1.01	1.66	0	2.81
FG	-0.22	2.22	0.28	-1.28	0.16	2.83	3.83	4.54	5.22
LGW	-1.35	3.01	1.37	-0.72	2.67	0.67	7.56	6.18	10.52

PV: predictive variable, LO: lodging, HP: height (cm), GL: grain length (mm), PL: panicle length (cm), GT: grain thickness (mm), FL: flowering (days), GW: grain width (mm), GP: number of filled grains per panicle, WG: weight of 100 grains (g), TI: tillering, FG: percentages of filled grains and LGW: length and width ratio of grains, GY: grain yield. Environment E1: Leopoldina, E2: Lambari, E3: Janaúba, FA: random forest, BA: *bagging*, BO: *boosting*.

Based on Table 4, the variables that obtained the highest estimate in all machine learning methodologies were length (GL) and grain width when the prediction target variable was grain length and width ratio (LGW) in all environments. For this same response variable, another variable that had a high IMSE estimate was panicle length (PL) in Leopoldina and Lambari, and Janaúba did not consider this variable to be the most important due to the low estimate of the IMSE percentage. On the other hand, the weight variables of 100 grains (WG) and the number of full grains per panicle (GP) proved to be efficient in quantifying the prediction of LGW by *boosting*. This procedure proved to be more consistent in predicting variables compared to the others.

The variable that obtained the highest IMSE estimate when PL was the target prediction variable was plant height (HP) for Leopoldina and Lambari. On the other hand, this variable in Janaúba was not highlighted in predicting PL. In Leopoldina, another variable that stood out in predicting PL was the number of grains filled per panicle (GP) for all machine learning approaches. When using the explanatory variable PL, the variable GY presented the highest IMSE in Janaúba for procedure *bagging*. Regarding the procedure *boosting* and about the same predictive variable, the results show discrepancies. On the other hand, this procedure was more consistent in predicting the variable. In this procedure, to quantify the importance of a variable using PL as a predictive target, the variables GP, GY, and LGW stood out in Leopoldina. In Lambari, other variables showed better performance in predicting PL, for example, GW, GY, and LGW, and in Janaúba, they were PL, GW, GP, GY, and LGW.

When the target prediction variable was GY, in Leopoldina, the variables that obtained an estimate of the high IMSE percentage were plant height (HP) and grain length (GL) in all machine learning procedures. On the other hand, in Lambari, the variable that stood out was panicle length (PL). In this environment, another variable that showed better predictive performance when GY was used as the main variable was flowering (FL) in *bagging* and *random forest*. In the *boosting* procedure, the variables that stood out were HP, GL, PL, GP, WG, and LGW in all environments.

The literature has highlighted machine learning techniques as efficient tools in quantifying the relative importance of variables, in view of simplicity, the nonuse of assumptions about the distribution of explanatory variables, and their robustness to quantity, redundancy, and environmental influences (Tan et al., 2014; Beucher et al., 2019). On the other hand, we verify this premise for the regression method. *Random forests* and *bagging* these methods have good predictive performances in practice; they work well for high-dimensional problems and can be used with multiclass output, categorical predictors, and imbalanced problems (Gregorutti et al., 2017). This author had satisfactory result variable selection with the *random forests* algorithm in the presence of correlated predictors.

Grain yield is a trait controlled by several genes and is therefore a quantitative inheritance (Freitas et al., 2007). Therefore, grain yield depends on the interaction of several yield components, for example, numbers of spikelets and grains per panicle, mass of a thousand grains, spike fertility index and panicle length, which are controlled by genetic factors, and environmental factors. The length of the panicle, the number of spikelets per panicle, the fertility of the spikelets, and the mass of a thousand grains directly affect grain yield (Evans & Bhatt, 1977). Thus, knowledge of these relationships can help breeders select new cultivars, which can increase the productivity and quality of grains and decrease the cost of production and the environmental impact.

The longer the flowering period in the rice culture, the more photoassimilates are produced and translocated to the grains, and consequently, an increase in grain yield. However, late-cycle cultivars tend to be more productive about the early cycle since they obtain an increase in the amount of photoassimilates that are translocated to the grains. According to Ntanos and Koutroubas (2002), productivity in rice has been justified by differences in the dynamics of the distribution of assimilates between organs during plant growth and development. From the results of these studies, it was found that the production of dry matter and the translocation of photoassimilates contributed significantly to the development of grains in different cultivars and, consequently, a direct relationship with grain yield.

Grain dimensions are the main determinants of grain weight and one of the three components (number of panicles per plant, number of grains per panicle, and weight of grains) of grain yield; therefore, they are important characteristics that affect yield in rice. In plant breeding applications, grain size is generally assessed by the weight of the grain, which is positively correlated with various characteristics, including the length, width, and thickness of the grain (Fan et al., 2006). These characteristics also influence acceptability for consumers, and therefore, the size/shape of the rice grain is an important preferential target characteristic for breeders (Huang et al., 2012; Anacleto et al., 2015). Cultivars of the short and long types are highly

preferred by many consumers in Japan, South Korea, and North China, while consumers in India, the USA, and other countries in South and Southeast Asia prefer long and medium grains (Misra et al., 2017).

Methodologies based on machine learning and computational intelligence do not depend on stochastic information and tend to be more efficient. These methodologies make no assumptions about the model but capture complex factors such as epistasis and dominance in prediction models. It is not necessary to know if the data have these effects and do not require any assumptions about the distribution of phenotypic values (Sousa et al., 2020). Machine learning algorithms have the advantage of modelling data in a nonlinear and a nonparametric manner (Osco et al., 2020). Unlike many traditional statistical methods, these algorithms are built with the advantage of dealing with noisy, complex, and heterogeneous data (Osco et al., 2019).

In this study, we compare different approaches to quantifying the importance of variables to identify relevant predictive variables within a regression problem. Additionally, we included in our comparison a traditional method that aims to find a small subset of important variables with ideal forecasting performance in flood-irrigated rice.

It is noteworthy that the 13 characteristics used in this study are laborious to obtain, and their evaluation can be costly if there are a greater number of genotypes to be evaluated. In this context, the study of the most important characteristics in prediction is necessary, since it is possible to reduce physical effort, cost, labour, and time in experimentation (Ferreira et al., 2015).

Predicting the importance of flood-irrigated rice characteristics is of paramount importance for breeding programmes, as it directs genotype selection more practically, in addition to serving as a theoretical and practical framework in support of new recommendation cultivars. In practical terms, these results are consistent.

Therefore, our study presents the performance of some methodologies to evaluate the relative contributions of each variable through computational intelligence and machine learning in flood-irrigated rice culture. An approach to quantify the effect of explanatory variables on genetic improvement has successfully identified the true importance of each variable, including those that exhibit strong and weak correlations with the main variables, which in our case are grain yield, length of panicle and grain length-to-width ratio.

Researchers can now identify the individual and interactive contributions of the predictor variables to the rice crop using artificial intelligence and machine learning.

Conclusion

Computational intelligence and machine learning methodologies were able to quantify the importance of explanatory variables in the prediction of grain yield in rice, grain length and width ratio, and panicle length. In addition to artificial intelligence and machine learning, it is able to handle more reduced or redundant information in the input variables. The characteristics able to assist in decision making are flowering, number of grains filled by panicles, and panicle length. The network with only one hidden layer with 15 neurons was efficient in determining the relative importance of variables in flooded rice.

Acknowledgements

The authors would like to thank the Research Support Foundation of the State of Minas Gerais, the National Council for Scientific and Technological Development and the Coordination for the Improvement of Higher Education Personnel for the financial support and researcher of Embrapa Rice and Beans Dr. Orlando Peixoto de Morais (*in memory*). This study was financed in part by the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Financial Code 001. The authors gratefully acknowledge the *Fundação de Amparo à Pesquisa do Estado de São Paulo* (FAPESP) for researcher fellowship to ICS 2018/26408-0.

Reference

Anacleto, R. Cuevas, R. P., Jimenez, R., Llorente, C., Nissila, E., Henry, R., Sreenivasulu, N. (2015). Prospects of breeding high-quality rice using post-genomic tools. *Theoretical and Applied Genetics*, *128*(8), 1449-1466. DOI: https://doi.org/10.1007/s00122-015-2537-6

Beck, M. W. (2018). NeuralNetTools: Visualization and analysis tools for neural networks. *Journal of Statistical*, *85*(11), 1-20. DOI: http://dx.doi.org/10.18637 / jss.v085.i11

Beucher, A., Møller, A. B., & Greve, M. H. (2019). Artificial neural networks and decision tree classification for predicting soil drainage classes in Denmark. *Geoderma*, *352*, 351-359. DOI: https://doi.org/10.1016/j.geoderma.2017.11.004

Cruz, C. D. (2016). Genes Software – extended and integrated with the R, Matlab and Selegen. *Acta Scientiarum. Agronomy*, *38*(4), 547-552. DOI: http://dx.doi.org/10.4025/actasciagron.v38i4.32629

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

De Oña, J., & Garrido, C. (2014). Extracting the contribution of independent variables in neural network models: a new approach to handle instability. *Neural Computing and Applications*, *25*(3-4), 859-869. DOI: https://doi.org/10.1007/s00521-014-1573-5

Degenhardt, F., Seifert, S., & Szymczak, S. (2019). Evaluation of variable selection methods for random forests and omics data sets. *Briefings in Bioinformatics*, *20*(2), 492-503. DOI: https://doi.org/10.1093/bib/bbx124

Evans, L. E., & Bhatt, G. M. (1977). Influence of seed size, protein content and cultivar on early seedling vigor in rice. *Canadian Journal of Plant Science*, *57*(3), 929-935. DOI: https://doi.org/10.4141/cjps77-133

Fan, C., Xing, Y., Mao, H., Lu, T., Han, B., Xu, C., ... Zhang, Q. (2006). *GS3*, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. *Theoretical and Applied Genetics*, *112*(6), 1164-1171. DOI: https://doi.org/10.1007/s00122-006-0218-1

Ferreira, M. G., Azevedo, A. M., Siman, L. I., Silva, G. H., Carneiro, C. S., Alves, F. M., ... Nick, C. (2017). Automation in accession classification of Brazilian *Capsicum* germplasm through artificial neural networks. *Scientia Agricola*, *73*(3), 203-207. DOI: http://dx.doi.org/10.1590/1678-992X-2015-0451

Freitas, J. G., Cantarella, H., Salomon, M. V., Malovolta, V. M. A., Castro, L. H. S. M., Gallo, P. B., & Azzini, L. E. (2007). Produtividade de cultivares de arroz irrigado resultante da aplicação de doses de nitrogênio. *Bragantia*, *66*(2), 317-325. DOI: http://dx.doi.org/10.1590/S0006-87052007000200016

Garson, G. D. (1991). Interpreting neural network connection weights. *Artificial Intelligence Expert*, *6*, 46-51.

Gedeon, T. D., Wong, P. M., & Harris, D. (1995). *Balancing bias and variance: network topology and pattern set reduction techniques*. Berlin, Heidelberg, GE: Springer Berlin Heidelberg.

Ghani, I. M. M., & Ahmad, S. (2010). Stepwise multiple regression method to forecast fish landing. *Procedia - Social and Behavioral Sciences*, *8*, 549-554. DOI: https://doi.org/10.1016/j.sbspro.2010.12.076

Gianola, D., Okut, H., Weigel, K. A., & Rosa, G. J. M. (2011). Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genetics*, *12*(87), 1-14. DOI: https://doi.org/10.1186/1471-2156-12-87

Goh, A. T. C. (1995). Back-propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering*, *9*(3),143-151. DOI: https://doi.org/10.1016/0954-1810(94)00011-S

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, *27*, 659-678. DOI: https://doi.org/10.1007/s11222-016-9646-1

Haddouche, R., Chetate, B., & Said Boumedine, M. (2018). Neural network ARX model for gas conditioning tower. *International Journal of Modeling and Simulation*, *39*(3), 166-177. DOI: https://doi.org/10.1080/02286203.2018.1538848

Hassanzadeh, Z., Ghavami, R., & Kompany-Zareh, M. (2015). Radial basis function neural networks based on the projection pursuit and principal component analysis approaches: QSAR analysis of fullerene[C60]-based HIV-1 PR inhibitors. *Medicinal Chemistry Research*, *25*, 19-29. DOI: https://doi.org/10.1007/s00044-015-1466-x

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statiscal learning data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.

Huang, X., Zhao, Y. Wei, X., Li, C., Wang, A., Zhao, Q., ... Han, B. (2012a) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nature Genetics*, *44*, 32-39. DOI: https://doi.org/10.1038/ng.1018

Li, L., & Zha, Y. (2019). Estimating monthly average temperature by remote sensing in China. *Advances in Space Research*, *63*(8), 2345-2357. DOI: https://doi.org/10.1016/j.asr.2018.12.039

Matlab. (2016). *Software*. Natick, MA: The MathWorks Inc.

Misra, G., Badoni, S., Anacleto, R., Graner, A., Alexandrov, N., & Sreenivasulu, N. (2017). Whole genome sequencing-based association study to unravel genetic architecture of cooked grain width and length traits in rice. *Scientific Reports*, *7*(12478), 1-16. DOI: https://doi.org/10.1038/s41598-017-12778-6

Ntanos, D. A., & Koutroubas, S. D. (2002). Dry matter and Naccumulation and translocation for Indica and Japonica riceunder Mediterranean conditions. *Field Crops Research*, *74*(1), 93-101. DOI: https://doi.org/10.1016/S0378-4290(01)00203-9

Olden, J. D., & Jackson, D. A. (2002). Illuminating the "black box": a randomization approach for understanding variable contributions in artifical neural networks. *Ecological Modelling*, *154*(1–2), 135-150. DOI: https://doi.org/10.1016/s0304-3800(02)00064-9

Osco, L. P., Ramos, A. P. M., Moriya, E. A. S., Bavaresco, L. G., Lima, B. C., Estrabis, N., ... Araújo, F. F. (2019). Modeling hyperspectral response of water-stress induced lettuce plants using artificial neural networks. *Remote Sensing*, *11*(23), 1-15. DOI: https://doi.org/10.3390/rs11232797

Osco, L. P., Ramos, A. P. M., Pinheiro, M. M. F., Moriya, E. A. S., Imai, N. N., Estrabis, N., ... Creste, J. E. (2020). A machine learning framework to predict nutrient content in valencia-orange leaf hyperspectral measurement. *Remote Sensing*, *12*(6), 1-21. DOI: http://dx.doi.org/10.3390/rs12060906

Paliwal, M. & Kumar, U. A. (2011). Assessing the contribution of variables in feed forward neural network. *Applied Soft Computing*, *11*, 3690-3696.

Parmley, K. A., Higgins, R. H., Ganapathysubramanian, B., Sarkar, S., & Singh, A. K. (2019). Machine learning approach for prescriptive plant breeding. *Scientific Reports*, *9*(1), 1-12. DOI: https://doi.org/10.1038/s41598-019-53451-4

Paruelo, J. M., & Tomasel, F. (1997). Prediction of functional characteristics of ecosystems: a comparison of artificial neural networks and regression models. *Ecological Modelling*, *98*(2-3), 173-186. DOI: https://doi.org/10.1016/s0304-3800(96)01913-8

Porwal, A., Carranza, E. J. M., & Hale, M. (2003). Artificial neural networks for mineral potential mapping; a case study from Aravalli Province, Western India. *Natural Resources Research*, *12*(3), 155-171. DOI: https://doi.org/10.1023/A:1025171803637

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys*, *28*(1), 71-72. DOI: https://doi.org/10.1145/234313.234346

Roy, P. P., & Roy, K. (2008). On some aspects of variable selection for partial least squares regression models. *QSAR & Combinatorial Science*, *27*(3), 302-313. DOI: https://doi.org/10.1002/qsar.200710043

Sant'Anna, I. C., Ferreira, R. A. D. C., Nascimento, M., Carneiro, V. Q., Silva, G. N., Cruz, C. D., ... Chagas, F. E. O. (2019). Multigenerational prediction of genetic values using genome-enabled prediction. *PLoS ONE*, *14*(1), 1-14. DOI: https://doi.org/10.1371/journal.pone.0210531

Santos, R. P, Dean, D. L., Weaver, J. M., & Hovanski, Y. (2018). Identifying the relative importance of predictive variables in artificial neural networks based on data produced through a discrete event simulation of a manufacturing environment. *International Journal of Modelling and Simulation*, *39*(4), 234-245. DOI: https://doi.org/10.1080/02286203.2018.1558736

Silva, G. N., Nascimento, M., Sant'Anna, I. C., Cruz, C. D., Caixeta, E. T., Carneiro, P. C. S., ... Oliveira, M. S. (2017). Artificial neural networks compared with Bayesian generalized linear regression for leaf rust resistance prediction in Arabica coffee. *Pesquisa Agropecuária Brasileira*, *52*(3), 186-193. DOI: http://dx.doi.org/10.1590/s0100-204x2017000300009

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. DOI: http://dx.doi.org/10.1590/0103-9016-2014-0057

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*(22), 1-19. DOI: https://doi.org/10.3390/app9224846

Somers, M. J., & Casal, J.C. (2009). Using artificial neural networks to model nonlinearity: The case of the job satisfaction-job performance relationship. *Organizational Research Methods*, *12*(3), 403-417. DOI: https://doi.org/10.1177/1094428107309326

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

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. http://dx.doi.org/10.1016/j.isprsjprs.2014.08.003.

Tsang, M., Cheng, D., & Liu, Y. (2017). Detecting statistical interactions from neural network weights. In *6th International Conference on Learning Representations* (p. 1-21). Vancouver, CA: ICLR. DOI: https://doi.org/10.48550/arXiv.1705.04977

Yu, H., Campbell, M. T., Zhang, Q., Walia, H., & Morota, G. (2019). Genomic Bayesian confirmatory factor analysis and Bayesian network to characterize a wide spectrum of rice phenotypes. *G3: Genes, Genomes, Genetics*, *9*(6), 1975-1986. DOI: https://doi.org/10.1534/g3.119.400154