DOI: 10.5433/1679-0359.2022v43n3p1017

Machine learning applied to the prediction of root architecture of soybean cultivars under two water availability conditions

Machine learning aplicado à predição da arquitetura radicular de cultivares de soja sob duas condições de disponibilidade hídrica

Anunciene Barbosa Duarte[1]*; Dalton de Oliveira Ferreira[2]; Lucas Borges Ferreira[3]; Felipe Lopes da Silva[4]

Highlights

Soybean plants under water deficit and control (no deficit) conditions were used.
Four machine learning models, single task learning and multitask learning were used.
SVM showed the best performance to predict root variables of 100 soybean cultivars.
Multitask learning provided similar results to single task learning.

Abstract

The objective of this study was to evaluate the performance of four machine learning models, as well as multitask learning, to predict soybean root variables from simpler variables, under two water availability conditions. In order to do so, 100 soybean cultivars were conducted in a greenhouse under a control condition and a stress condition. Aerial part and root variables were evaluated. The machine learning models used to predict complex root variables were artificial neural network (ANN), random forest (RF), extreme gradient boosting (EGBoost) and support vector machine (SVM). A linear model was used for comparison purposes. Multitask learning was employed for ANN and RF. In addition, feature importance was defined using RF and XGBoost algorithms. All the machine learning models performed better than the linear model. In general, SVM had the greatest potential for the prediction of most of the root variables, with better values of RMSE, MAE and R^2. Dry weight of the aerial part and root volume exhibited the greatest importance in the predictions. The models developed using multitask learning performed similarly to the ones conventionally developed. Finally, it is concluded that the machine learning models evaluated can be

[1] PhD Student in Plant Science, Department of Agronomy, Universidade Federal de Viçosa, UFV, Viçosa, MG, Brazil. E-mail: cieneduarte@live.com
[2] PhD Student in Genetics and Breeding Department of Biology, UFV, Viçosa, MG, Brazil. E-mail: daltonferreira.ufv@gmail.com;
[3] PhD Student in Agricultural Engineering, Department of Agricultural Engineering, UFV, Viçosa, MG, Brazil. E-mail: contato.lucasbf.gmail.com
[4] Prof. Dr. of Department of Agronomy, UFV, Viçosa, MG, Brazil. E-mail: felipe.silva@ufv.br
* Author for correspondence

Received: Apr. 08, 2021 - Approved: Jan. 14, 2022

used to predict root variables of soybean from easily measurable variables, such as dry weight of the aerial part and root volume.
Key words: *Glycine max* L. Multitask learning. Root morphology. Water deficit.

Resumo

O objetivo deste estudo foi avaliar o desempenho de quatro modelos de machine learning, bem como multitask learning, para predizer variáveis radiculares de soja a partir de variáveis simples, em duas condições de disponibilidade hídrica. Para isso,100 cultivares de soja foram conduzidas em casa de vegetação sob uma condição controle e uma condição estresse. Foram avaliadas as variáveis da parte aérea e da raiz. Os modelos machine learning usados para predizer variáveis complexas do sistema radicular foram rede neural artificial (RNA), random forest (RF), extreme gradient boosting (EGBoost) e support vector machine (SVM). O modelo linear foi usado para fins de comparação. O multitask learning foi empregado para RNA e RF. Além disso, a importância das variáveis foi definida usando algoritmos RF e XGBoost. Todos os modelos de machine learning apresentaram melhor desempenho do que o modelo linear. Em geral, SVM apresentou o maior potencial de predição da maioria das variáveis raiz, com melhores valores de RMSE, MAE e R^2. O peso seco da parte aérea e o volume da raiz exibiram as maiores importâncias nas predições. Os modelos desenvolvidos por meio do multitask learning apresentaram desempenhos semelhantes aos desenvolvidos convencionalmente. Por fim, conclui-se que os modelos de machine learning avaliados podem ser usados para predizer variáveis radiculares de soja a partir de variáveis facilmente mensuráveis, como massa seca da parte aérea e volume radicular.
Palavras-chave: Deficit hídrico. *Glycine max* L. Morfologia de raiz. Multitask learning.

Introduction

Soybean is of great importance to the world economy, given the versatility of its products and by-products. This crop has been extensively studied for several areas. However, despite many studies addressing this oilseed, the study of root morphology remains a major challenge for researchers, due to high labor costs and time spent to carry out the measurements (Falk et al., 2020). Roots are essential organs for the absorption of water and nutrients. Furthermore, they are the first organs to detect environmental stress generated by water deficit (Fenta et al., 2014).

Roots have high morphological plasticity in the soil (Ito, Tanakamaru, Morita, Abe, & Inanaga, 2006). Root architecture traits, such as length, diameter, surface area and volume, are associated with the performance of the plants under water deficit conditions (Sandhu et al., 2016). In addition, Dubey, Kumar, Abd-Allah, Hashem and Khan (2019) mentioned that soybean can overcome the impact of drought if there is a powerful and deep root system in the early stage of development. Thus, specific information about the root morphology of soybean cultivars may reflect potential drought tolerance.

To obtain detailed information on root morphology, software such as Winrhizo, which is based on image analyses, is usually used to measure variables such as total root length, projected area, surface area, average

diameter, length/volume, root tips, length per root diameter classes, and others (Wang & Zhang, 2009). However, software like this can have high costs, limiting its practical application. Furthermore, it is necessary to acquire images of roots, which requires intense work, given the need to distribute the roots in scanners in order to avoid overlapping. In the case of soybean, overlapping problems are particularly important when the root system is in an advanced stage of development. In this case, to be scanned, the roots may need to be partitioned, requiring even more time and manpower.

In contrast to complex root variables, aerial part variables (plant height, hypocotyl diameter, number of nodes, dry weight of aerial part) and some simpler root variables such as root length and volume can be easily obtained. The length of the root in depth, for example, can be measured with the aid of a tape measure, while the root volume can be measured manually by the volumetric method (Ratke, Santos & Souza, 2019). In this case, root volume is obtained by the displacement of water caused by the introduction of the roots in a graduated cylinder. Thus, such variables can be used to assist in the prediction of more complex root system variables. For this, a very promising strategy is the use of machine learning techniques to predict root architecture from simpler variables. This type of approach can save time and reduce costs.

Machine learning models have achieved high performance in different types of problems, including many applications in agriculture (Fan et al., 2018; L. Zhang et al., 2019; Rahmati et al., 2020; Ruiming & Shijie, 2020). So far, studies involving root architecture have benefited from machine learning techniques to determine root traits that best discriminate rice genotypes (Iyer-Pascuzzi et al., 2010); prediction of time series of soil water content at the depth of corn rooting (Karandish & Shahnazari, 2016); classification of corn roots (Zhong, Novais, Grift, Bohn, & Han, 2009); selection of phenotypic characteristics highly associated with the production of cassava roots (Santos Silva, Souza, & Oliveira, 2019). Recently, one study reported the use of machine learning in the development of a mobile platform for soybean root phenotyping (Falk et al., 2020). Unlike the studies available in the literature, our proposal is to estimate complex traits of the soybean root system from simpler variables, without requiring image acquisition.

When developing machine learning models, multitask learning can be used. For predicting unrelated variables, individual models are the standard recommendation. However, when the variables of interest have some relationship with each other, a single model may be able to predict all of them at the same time (multitask). Thus, multitask learning is an approach to learn various related tasks, extracting and sharing common information within the tasks (García-Laencina, Sancho-Gómez, & Figueiras-Vidal, 2013). So, in addition to being able to achieve better performances, only one model is needed to predict all the variables of interest, facilitating the development and use of the proposed solution. Many studies have demonstrated the success of multitask learning (Park, Kim, Park, & Kim, 2018; Singh, Sisodia, & Singh, 2020). Thus, according to our knowledge, this approach has not been used to predict complex variables of the root system.

Given the above, the objective of this study was to evaluate the performance of four machine learning models, as well as multitask learning, to predict root architecture of soybean cultivars from simpler variables, under two water availability conditions.

Materials and Methods

Database

Data were obtained from an a trial, carried out in a greenhouse, involving 390 plants from 100 commercial soybean cultivars, sown in 5 L vases containing clay soil. In this trial, to verify the effect of water deficit on soybean plants, two water availability conditions were evaluated: control condition and stress condition. In the control condition, the soil was maintained at field capacity, considering a soil water tension of 33 kPa. In the stress condition, a soil water tension of 900 kPa was adopted, as recommended by Krishnan, Singh, Verma, Joshi and Singh (2014). To apply water deficit, soil moisture was controlled by monitoring the weight of the vase-soil-plant system, maintaining its weight at the value corresponding to the moisture equivalent to the soil water tension of interest. The relationship between soil moisture and soil water tension was obtained through a soil water retention curve. The stress condition was started on vegetative stage V1 (first fully developed trefoil) and maintained for 20 days.

After the stress period, soybean plants were removed from the vases, separating the aerial part from the roots. Then, the following variables were measured in the aerial part: plant height (cm), hypocotyl diameter (mm) and number of nodes. In order to obtain dry weight of the aerial part (g), the plants were placed in an oven at 65 °C for 72 hours. Subsequently, the roots were carefully washed, and due to the easy measurement, length (cm) and root volume (mL) were manually evaluated. Root volume was measured as suggested by Laurett, Fernandes, Schmildt, Almeida and Pinto (2017).

To perform the evaluations related to root architecture, the roots were stored in 70% alcohol. For a better understanding of the root morphology of soybean cultivars, the roots were arranged in a tray (20 cm × 40 cm x 7 cm (width x length x height)), containing enough distilled water to cover the entire root, avoiding overlaps. Then, images were obtained with the EPSON EU88 scanner, with a resolution of 400 dpi. The images were processed using the WinRhizoPro 2009 software (Wang & Zhang, 2009). This software was chosen due to its good performance in root assessments (Sun et al., 2020; Puspasari et al., 2020).

With the aid of the WinRhizo software, the following root traits were measured: total root length (cm), surface area (cm^2), projected area (cm^2), length/volume ratio ($cm\ m^{-3}$), average diameter (mm) and root tips. Root lengths per root diameter classes were also measured using the mentioned software. Three classes were considered: length of roots with diameter less than 0.5 mm (L1), length of roots with diameter greater than 0.5 mm and less than 1.0 mm (L2), and length of roots with diameter greater than 1.0 and less than 1.5 mm (L3). An overview of the data collection process can be seen in Figure 1.

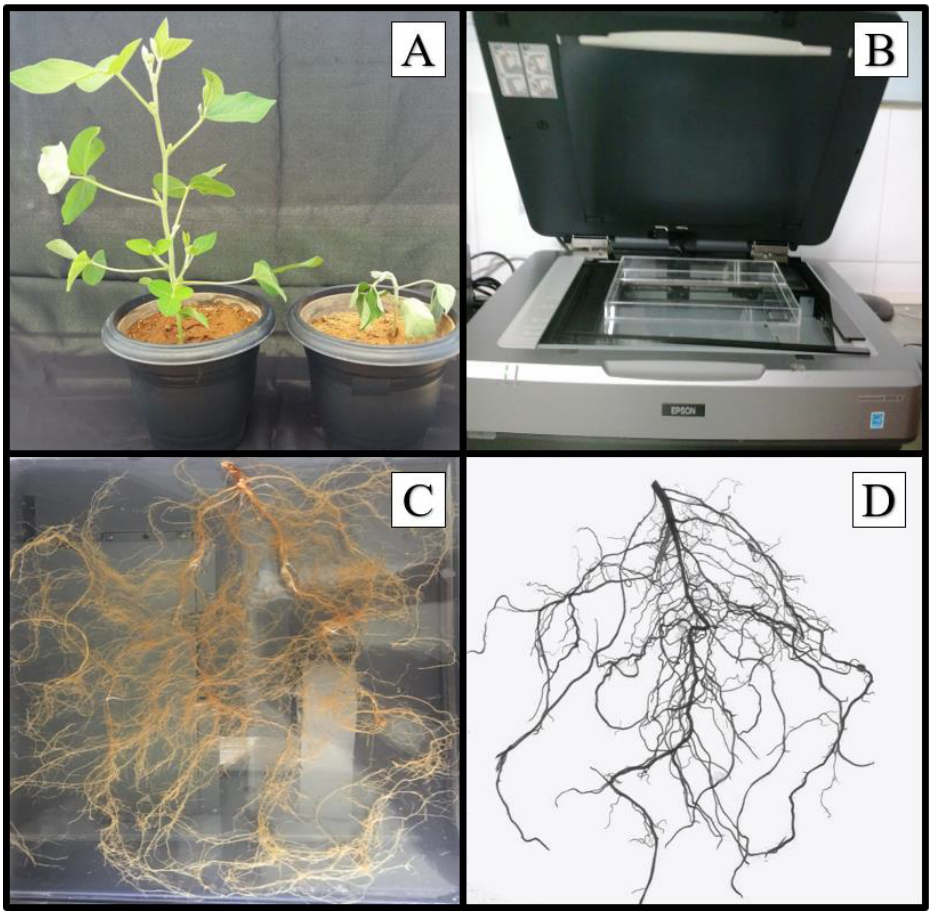

Figure 1. Demonstration of the data collection process: (A) soybean plant under the control condition (left) and the stress condition (right); (B) scanner; (C) roots to be scanned in a transparent tray; (D) scanned root image.

Variables used

The variables used in this study were divided into two groups. Group I consisted of the following variables: plant height (HEI), hypocotyl diameter (HD), number of nodes (NN), dry weight of the aerial part (DW), root length (depth) (LEN) and root volume (VOL). These variables were chosen because they are easy to acquire and, therefore, were used as input to the machine learning models. Therefore, these variables were used to predict more complex variables of the root system of soybean cultivars. On the other hand, group II was composed of the variables inherent to root architecture, measured by the WinRhizo software. Thus, these variables were considered in this study as variables of difficult acquisition. These variables were: total root length (TLEN) (obtained by summing all the lateral root lengths), projected area (PAR), surface area (SAR), average diameter (DIA), length/volume ratio (L/V), root tips (RTI) and length per root diameter classes (L1, L2 and L3). These variables composed the outputs of the models, that is, they were predicted from the input variables.

Machine learning models and modeling strategies

Four machine learning models and two modeling strategies were used to predict root variables of soybean. The modeling strategies were: (i) use of single task learning with four machine learning models (i.e., individual prediction of each output variable, developing a model for each variable); (ii) use of multitask learning, in which all the output variables are predicted at the same time, generating a single model. This latter approach was used only for artificial neural networks and random forest since multitask learning requires models capable of predicting multiple values at once.

The four machine learning models used were artificial neural networks (ANNs), random forest (RF), extreme gradient boosting (XGBoost) and support vector machine (SVM). For comparison purposes, linear models (LMs) were also used. LMs are simple and widely known in the literature, being able to capture linear relationships between input and output variables.

To implement the models, the Python programming language was used, with the following libraries: Scikit-learn, XGBoost and TensorFlow. To optimize hyperparameters and evaluate the performance of the models, the k-fold cross-validation was used, with k equal to 5. For this, the dataset was divided into 5 parts/folds. Each model was trained with data from 4 parts and evaluated in the remaining one. This procedure was repeated 5 times in order to ensure that all parts had been used both as training and as validation data. The final performance of the model was expressed as the average performance obtained in the 5 folds.

Hyperparameter optimization was done using grid-search, choosing the hyperparameter values that promoted the lowest mean prediction error obtained with k-fold. The use of k-fold cross-validation is important for a more accurate estimation of the performance of models, especially in small data sets, such as the one used in the present study. The dataset used was obtained from 390 soybean plants submitted to two water availability conditions (control condition and water stress/deficit condition), as previously presented. It promoted a wide variability in the data used in this study.

Artificial neural networks (ANNs)

ANNs are mathematical/computational models that resemble the architecture of the human brain (Hasson, Nastase, & Goldstein 2020). A multilayer perceptron is the best-known ANN architecture, which is composed of a series of layers, neurons and connections. The connections between artificial neurons receive synaptic weights, which are adjusted during the network training process. According to Tang, Chan, and Chan (2019), a neural network is typically composed of three layers: input layer, where the input variables are inserted; hidden layer, in which the data is processed; and output layer, where the results are produced. The ideal number of layers varies according to the objective, data used and complexity of the problem. More details on neural network can be found in Hasson et al. (2020) and Patil and Deka (2016).

In the present study, multilayer perceptron ANNs with one hidden layer were used. For the number of neurons in the hidden layer, the tested values ranged from 5 to 30,

with an interval of 5 neurons. For the number of training epochs, the following values were tested: 50, 100, 150, 200, 250 and 300. A sigmoid activation function was used in the hidden layer and a linear function was used in the output layer. The learning rate was set to 0.001 and the batch size was set to 32. The Adam training algorithm was used.

Random Forest (RF)

RF is a supervised learning algorithm, which is essentially based on a combination of decision trees (Bressan, Souza, Girelli, & Chemale, 2020). It is an efficient tool in studies involving classification, regression and selection/importance of variables (Sariyer, Tasar, & Cepe, 2019). This algorithm allows to verify the contribution of each input variable for the prediction process. Its main advantages are the ease of training, in addition to having low sensitivity to outliers, high computational efficiency and robustness against overfitting (Belgiu & Drãgut, 2016). More details on this method can be obtained in Izquierdo-Verdiguier and Zurita-Milla (2020).

RF usually requires less adjustments in hyperparameters during the training process. In this study, the hyperparameters adjusted were the number of trees (n_estimators), the number of features considered for splitting at each leaf node (max_features) and the minimum number of samples required to be at a leaf node (min_samples_leaf). For n_estimators, the following values were tested: 100, 200, 400 and 600; for max_features, the tested values varied from 1 to 5, with an interval of 1 unit; and for min_samples_leaf, the following values were tested: 1, 5, 10, 15.

Extreme Gradient Boosting (XGBoost)

XGBoost is a model similar to RF, also based on decision trees. This method combines the predictions of a set of trees, which are now added sequentially to maximize the predictive performance (Carmona, Climent, & Momparler, 2019). XGBoost is based on the principle of "boosting", combining all the predictions of a set of "weak" learners to develop a "strong" learner through additive training strategies (Fan et al., 2018). Therefore, its main advantage consists of improving performance and reducing overfitting. Furthermore, it can be used to determine the importance of input variables. More information on this algorithm is available in Chen and Guestrin (2016).

In this study, the following hyperparameters were optimized: number of trees (n_estimators), testing values 25, 50, 75, 100, 200 and 400; maximum tree depth (max_depth), which varied from 2 to 7, with an interval of 1 unit; learning rate (learning_rate), testing values 0.01, 0.1, 0.2 and 0.3; and subsample ratio of columns when constructing each tree (colsample_bytree), testing values 0.6, 0.8 and 1.0.

Support vector machine (SVM)

SVM is a supervised learning algorithm with great contributions in the last years as it has presented good results when applied to classification and regression problems (Raghavendra & Deka, 2014). The main advantage associated to this model is that it uses the kernel trick. For this, the polynomial function and the radial basis function (RBF) can be used. Thus, SVM consists of the application of a linear regression in a high-dimensional feature space obtained from the

input space by nonlinear mapping. In this way, the algorithm is able to create specialized knowledge about the problem, minimizing the prediction error. Further explanation on this model can be obtained in Saruta et al. (2013).

Regarding hyperparameter adjustment, the regularization parameter (C), the kernel coefficient gamma (gamma) and the epsilon coefficient (epsilon) were optimized. The tested values for C were 0.05, 0.1, 0.5, 1.0, 2.0 and 5.0. For gamma, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2 and 0.4 were tested. For epsilon, 0.05, 0.1 and 0.2 were tested.

Multitask learning

Multitask learning consists of sharing statistical information between related tasks, so that the overall performance of the tasks is improved (Goncalves et al., 2019). For this, it is assumed that all tasks, or at least a subset of them, are related to each other. In multitask learning, models with multiple outputs are developed. All output variables are predicted at the same time by a single model. It is expected that joint learning of tasks leads to performance improvements if compared to individual learning (single task learning), besides simplifying the predictive process, since only one model needs to be built. In this study, this approach was used only for the ANN and RF algorithms, which support prediction of multiple variables at the same time.

Importance and selection of variables

With the aid of RF and XGBoost models, input variable importances for the prediction of root variables of soybean cultivars were evaluated. After that, the most important variables were selected and, in order to validate the selection process, only the selected variables were used as input for the best performing model evaluated in this study.

Data normalization

All data used in this study (input and output variables) were normalized according to Equation 1. In each step of k-fold, the mean and standard deviation were obtained using only data from the training set, without including data from the validation set. This is done to avoid leakage of information from the validation set to the training set, ensuring greater robustness to the validation process.

Equation 1: $x_{ni} = \frac{x_i - \mu}{\sigma}$

Where x_{ni} is the normalized value, x_i is the observed value, μ is the mean and σ is the standard deviation.

Performance comparison criteria

The performance of the models was evaluated using the following statistical indicators: root mean square error (RMSE), coefficient of determination (R^2), mean bias error (MBE) and mean absolute error (MAE), according to the equations below.

$$RMSE = \sqrt{\frac{1}{n}\sum(P_i - O_i)^2}$$

$$MAE = \frac{1}{n}\sum|P_i - O_i|$$

$$MBE = \frac{1}{n}\sum(P_i - O_i)$$

$$R^2 = \left[\frac{\sum(P_i-\overline{P})(O_i-\overline{O})}{\sqrt{(\sum(P_i-\overline{P})^2)(\sum(O_i-\overline{O})^2)}}\right]^2$$

Where, P_i is the predicted value, O_i is the observed value, $\overline{P}$ is the mean of the predicted values, $\overline{O}$ is the mean of the observed values and n is the number of data pairs.

Results and Discussion

The correlation matrix containing the input and output variables considered in this study is shown in Figure 2. In general, the input variables (HD, HEI, NN, DW, LEN and VOL) have a good correlation with the output variables (TLEN, PAR, SAR, DIA, R / V, RTI, L1, L2 and L3). Among the input variables, LEN had the lowest correlations and DW and VOL had the highest correlations (above 70%) with the output variables. The output variables showed a high correlation between them, exceeding 90% in most cases.

Figure 2. Correlation matrix containing the input and output variables considered in the present study. HD: hypocotyl diameter; HEI: plant height; NN: number of nodes; DW: dry weight of the aerial part; LEN; root length; VOL: root volume; TLEN: total root length; PAR: projected area; SAR: surface area; DIA: average diameter; L/V: length/volume ratio; RTI: root tips. L1, L2 and L3: root length by root diameter (d) classes for d<0.5, 0.5<d<1, 1.0<d<1.5, respectively.

Single task learning

In general, all the machine learning models showed better results than those obtained with the linear model (Table 1). Although the linear model is preferred to solve regression problems due to its simplicity, computational efficiency and ease of interpretation, this method can detect only linear relationships between the input variable and the response variable. Thus, machine learning models have the potential to solve more complex problems, as seen in the present study.

Table 1
Statistical indices for four machine learning models used to predict root architecture of soybean cultivars

Variable	Model	RMSE	MAE	MBE	R^2
Total length	LM	619.50	422.83	1.88	0.71
	RF	588.06	418.02	2.75	0.74
	RNA	594.24	404.26	4.15	0.73
	SVM	592.62	389.77	-20.04	0.73
	XGBoost	608.71	430.14	18.19	0.72
Projected area	LM	29.52	19.74	0.14	0.73
	RF	27.86	19.81	0.24	0.76
	RNA	28.25	19.15	0.26	0.76
	SVM	28.45	19.22	-0.97	0.75
	XGBoost	29.11	20.05	1.33	0.74
Surface area	LM	92.73	62.00	0.43	0.73
	RF	87.52	62.23	0.76	0.76
	RNA	88.75	60.17	0.81	0.76
	SVM	89.37	60.37	-3.05	0.75
	XGBoost	91.47	62.98	4.17	0.74
Diameter	LM	0.03	0.03	0.00	0.10
	RF	0.03	0.03	0.00	0.10
	RNA	0.03	0.03	0.00	0.10
	SVM	0.03	0.03	0.00	0.07
	XGBoost	0.03	0.03	0.00	0.08
Length/volume	LM	619.50	422.83	1.88	0.71
	RF	588.06	418.02	2.75	0.74
	RNA	594.24	404.26	4.15	0.73
	SVM	592.62	389.77	-20.04	0.73
	XGBoost	608.71	430.14	18.19	0.72

continue...

continuation...

Root tips	LM	789.06	594.89	1.36	0.56
	RF	788.05	595.16	4.53	0.56
	RNA	782.85	587.31	-12.14	0.56
	SVM	792.20	582.85	-111.5	0.56
	XGBoost	804.40	607.53	19.66	0.54
L1	LM	473.14	331.91	1.00	0.68
	RF	451.38	327.84	1.93	0.71
	RNA	458.76	322.79	2.60	0.70
	SVM	455.70	309.57	-24.49	0.70
	XGBoost	465.97	335.48	13.48	0.69
L2	LM	128.79	87.50	0.48	0.71
	RF	121.66	83.97	1.32	0.74
	RNA	121.82	82.09	1.05	0.74
	SVM	122.48	81.23	0.06	0.74
	XGBoost	127.35	87.04	4.32	0.72
L3	LM	36.46	24.82	0.24	0.70
	RF	34.54	24.64	0.52	0.74
	RNA	34.64	24.06	0.62	0.73
	SVM	34.67	23.85	-1.77	0.73
	XGBoost	35.79	25.52	3.01	0.73

L1, L2 and L3: root length per root diameter (d) classes, for d<0.5, 0.5<d<1, 1<d<1.5, respectively.

In general, the ANN and SVM models, with a slight emphasis on the SVM model, exhibited the best performances among the models evaluated, obtaining lower RMSE and MAE values. The RF model also performed well, however, in relation to ANN and RF, it tended to have higher MAE values. In contrast, the XGBoost model presented the worst performance among the machine learning models, with higher RMSE and MAE values, and lower R^2 values. Thus, although XGBoost has been considered highly efficient and has achieved good results in several studies (Ni et al., 2020; Zhong, Johnson & Chen, 2020; Zhang, Wu, Zhong, Li, & Wang, 2020), this behavior was not observed in this work.

Observed and predicted values of the evaluated root variables, for the model with the best overall performance (SVM), are shown in Figure 3. In general, SVM had good ability to predict the variables. It is also observed that, for high values, there was a tendency of underestimation, in addition to greater scattering of predicted values in relation to the observed values.

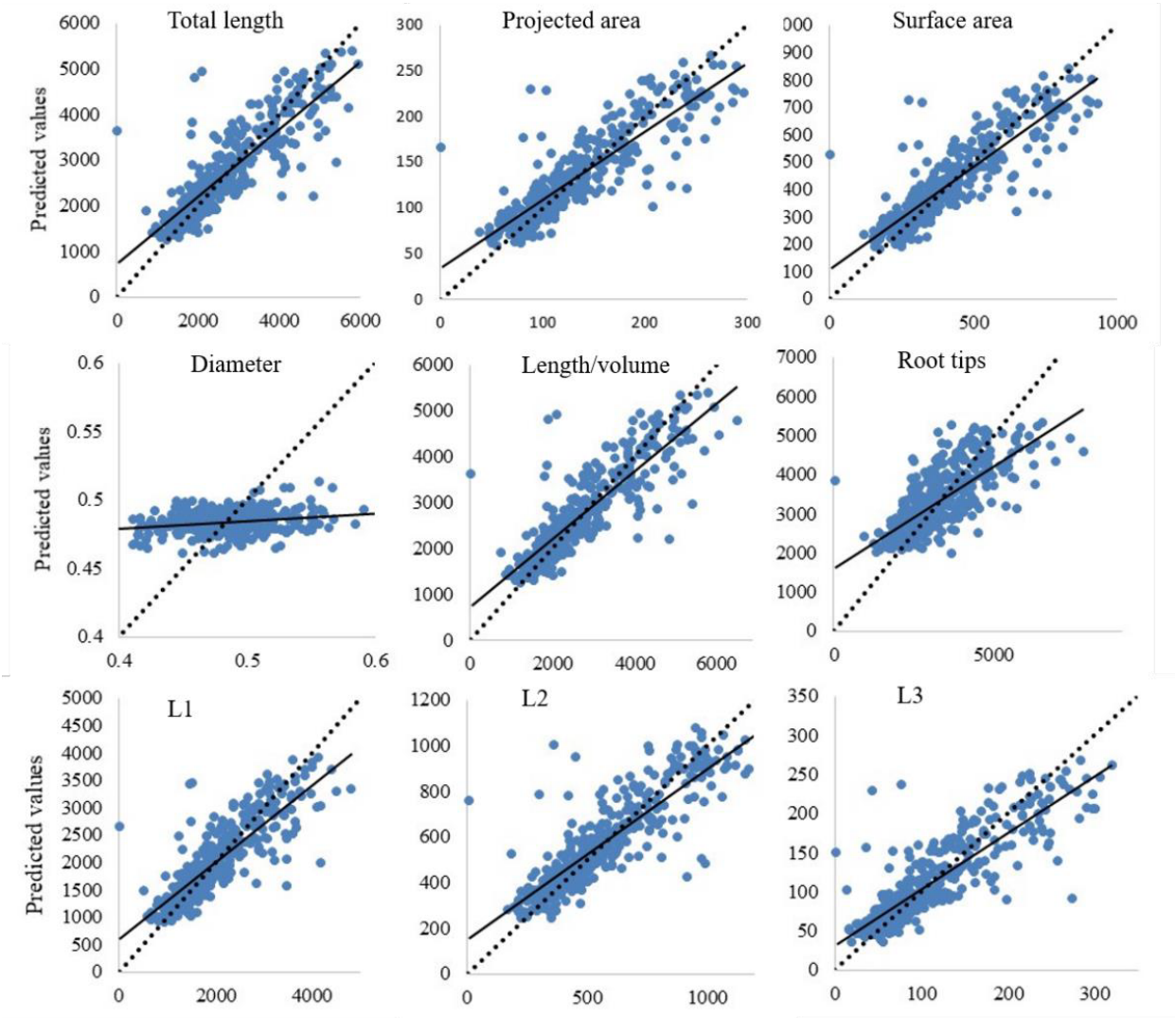

Figure 3. Observed and predicted values of root architecture variables of soybean cultivars under two water availability conditions using SVM. L1, L2 and L3 indicate root length per root diameter (d) classes, for d<0.5, 0.5<d<1, 1<d<1.5, respectively.

Although the machine learning models performed well for most of the variables studied, none of them was efficient in estimating the average diameter of soybean roots (Table 1 and Figure 3). This was possibly because the input variables were not able to explain the behavior of the mentioned variable. Furthermore, it is observed that the input variables have a low correlation with the average diameter of soybean roots (Figure 3).

Multitask learning

Among the four machine learning models tested in this study, only ANN and RF can be used with multitask learning, in which all output variables are predicted at the same time. The results obtained using this approach are shown in Table 2. The RF and ANN models showed similar performances, with a slight advantage for ANN, which generally

obtained lower MAE values. When using multitask learning, performances similar to those obtained with the individual prediction of root variables (single task learning) were obtained (Table 1). Although multitask learning presented very close results to those found in the first approach, the results obtained demonstrate that it is a potential tool in studies like this, since the use of a single model can facilitate the predictive process. In contrast, in the previous approach, which uses a specific model to predict each of the nine variables studied, a greater computational effort is required to train the models.

Table 2
Statistical indices for root architecture prediction of soybean cultivars using multitask learning with two machine learning models (RF and RNA)

Model	Model	RMSE	MAE	MBE	R^2
Total length	RF	590.87	420.78	3.04	0.74
	RNA	598.36	410.41	-2.90	0.73
Projected area	RF	28.35	20.20	0.16	0.75
	RNA	28.57	19.43	0.14	0.75
Surface area	RF	89.05	63.46	0.50	0.75
	RNA	89.92	60.80	0.47	0.75
Diameter	RF	0.03	0.03	0.00	0.06
	RNA	0.03	0.03	0.00	0.09
Length/volume	RF	590.87	420.78	3.04	0.74
	RNA	598.77	408.25	-3.27	0.73
Root tips	RF	783.31	591.24	6.61	0.56
	RNA	789.70	594.65	4.56	0.56
L1	RF	452.66	328.63	2.09	0.71
	RNA	458.13	324.94	-3.49	0.70
L2	RF	123.70	87.82	0.68	0.73
	RNA	122.94	82.79	0.41	0.73
L3	RF	34.81	24.98	0.22	0.74
	RNA	35.38	24.42	0.51	0.72

L1, L2 and L3 indicate root length per root diameter (d) classes, for $d<0.5$, $0.5<d<1$, $1.0<d<1.5$, respectively.

Feature importance

The XGBoost and RF models allowed to verify the importance of each input variable in the prediction of the analyzed variables (Figure 4). For the models developed using single task learning, the input variables with the greatest importances were root volume and dry weight of the aerial part.

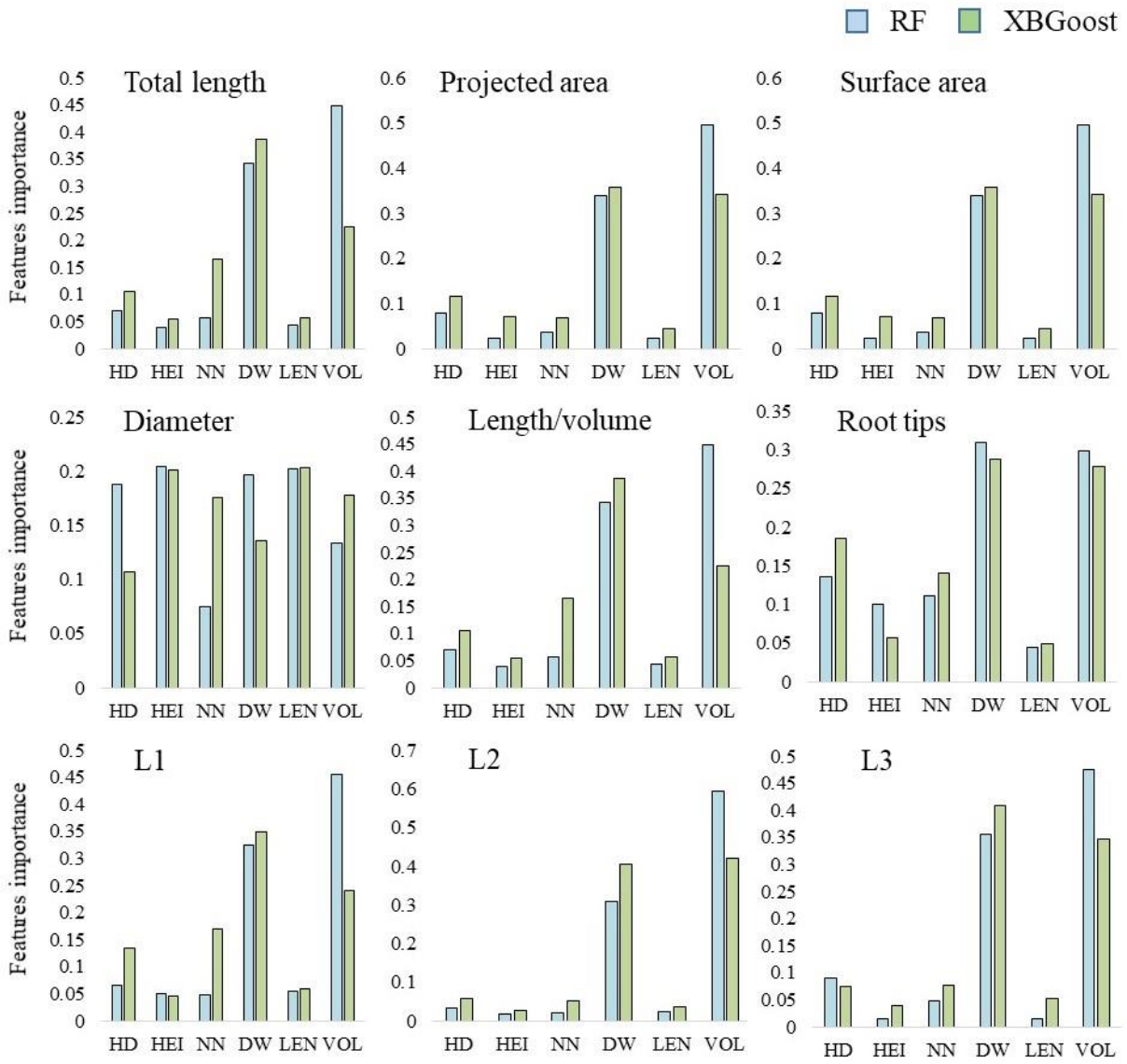

Figure 4. Feature importance for the prediction of root architecture of soybean cultivars using the Random Forest and XGBoost models. HD: hypocotyl diameter; HEI: plant height; NN: number of nodes; DW: dry weight of the aerial part; LEN; root length and VOL: root volume. L1, L2 and L3 indicate root length per root diameter (d) classes, for d<0.5, 0.5<d<1.0, 1.0<d<1.5, respectively.

For the RF model, root volume was the most important variable for all the predicted variables, except for diameter and root tips. For the XGBoost model, dry weight of the aerial part was the most important variable. Root growth in depth is strongly related to the growth of secondary roots (Strock, De La Riva, & Lynch, 2018). However, in this study, root length (manually measured) did not significantly contribute to the prediction of any of the root variables measured by WinRhizo. This is possibly due to the fact that root

length has a low correlation with the predicted variables (Figure 2). On the other hand, root volume and dry weight of the aerial part were the input variables that showed the greatest correlations with the output variables.

When using multitask learning, it was also observed the high importance of variables root volume and dry weight of the aerial part (Figure 5). Studies like this deserve attention, especially when there is a greater number of cultivars to be evaluated. In this study, 100 soybean cultivars were evaluated under two water availability conditions, with a total of 390 plants being evaluated, which required time and cost to carry out all the measurements. Through a feature importance analysis, it is possible to know which variables are most relevant. Thus, a smaller number of variables can be used to predict others.

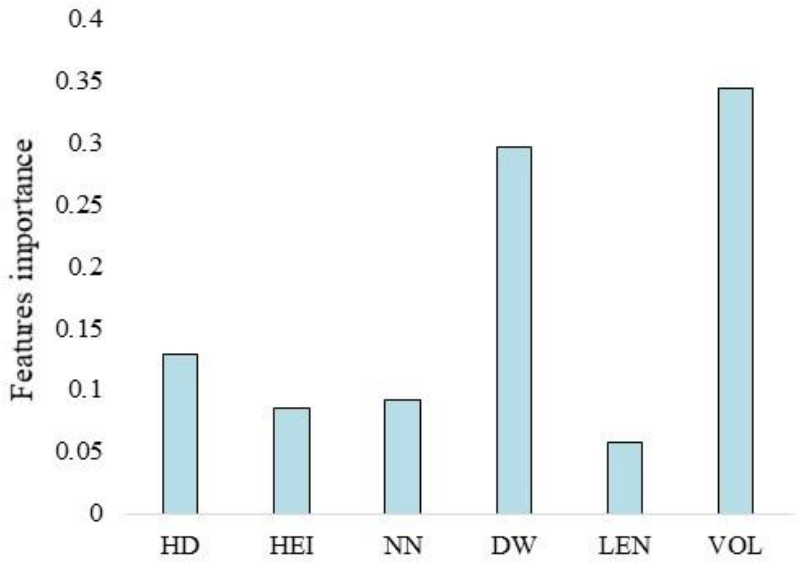

Figure 5. Feature importance in the prediction of root architecture of soybean cultivars under two water availability conditions using the random forest and multitask learning. HD: hypocotyl diameter; HEI: plant height; NN: number of nodes; DW: dry weight of the aerial part; LEN; root length and VOL: root volume.

Given the above, knowing which input variables are the most important is of great importance, especially for root variables. Based on such information, a researcher can, for example, exclude measurements of less relevant variables. In this sense, based on a feature importance analysis (Figure 5), the root variables of soybean cultivars were again predicted using only the two most important variables, suh as input and the SVM model, using the single task learning approach (Table 3). This model was chosen because it presented better performances in the prediction of most of the root variables studied.

Table 3
Statistical indices for the prediction of root architecture of soybean cultivars under two water availability conditions, using only two input variables and the SVM model

Variable	RMSE	MAE	MBE	R^2
Total length	613.63	416.82	3.57	0.72
Projected area	28.55	19.04	0.35	0.75
Surface area	89.69	59.82	1.11	0.75
Diameter	0.03	0.03	0.00	0.04
Length/volume	613.63	416.82	3.57	0.72
Root tips	790.99	589.73	-6.46	0.56
L1	477.26	334.94	5.68	0.68
L2	122.93	81.76	1.40	0.73
L3	35.34	24.20	0.51	0.73

L1, L2 and L3 indicate root length per root diameter (d) classes, for d<0.5, 0.5<d<1.0, 1.0<d<1.5, respectively.

When developing the SVM model using only the two most important input variables, it was observed that there was only a slight increase in the RMSE and MAE values in relation to the SVM model developed using all the six input variables. These results confirm the benefits of analyzing feature importance. Thus, it can be confirmed that root volume and dry weight of the aerial part are important to predict more complex variables of the root system, such as those evaluated in this study.

Given the high potential of machine learning models, as seen in this study, these tools have a great potential and can contribute to studies involving root architecture. With the use of these techniques, researchers have many possibilities, being possible to predict complex variables from simple variables. In the soybean context, there is still a lot to explore for a better understanding of the root system. In this study, only the vegetative stage was evaluated. However, research involving the entire growth cycle, or even the reproductive stage, can contribute positively to inferences regarding root architecture. In addition, machine learning models allow to verify the importance of input variables in the prediction of target variables.

Conclusions

In this work, data from 100 soybean cultivars under two water availability conditions were used with the objective of predicting root variables that are difficult to measure from simpler variables. Four machine learning models (ANN, RF, XGBoost and SVM) were used, two of them (ANN, RF) were applied using the multitask learning approach.

In general, the machine learning models, especially SVM, showed adequate potential to predict root architecture of soybean cultivars. This algorithm showed better values for RMSE, MAE and R^2. RF and XGBoost allowed to verify the importance of the input variables. It was found that among the input variables used, dry weight of the aerial part and root

volume were the most important ones. When using multitask learning, similar results were found in relation to the conventional approach. The main advantage of this strategy was to facilitate the predictive process, requiring only a single model to predict the nine root variables analyzed. Therefore, it is concluded that machine learning models, especially SVM, can be used to predict root variables of soybean cultivars, using easily measurable variables, such as dry weight of the aerial part and root volume.

Ackowledgments

We thank the Fundação de Apoio a Pesquisa do Estado de Minas Gerais, the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior and the Conselho Nacional de Desenvolvimento Científico e Tecnológico for financial support.

References

Belgiu, M., & Drăgu, L. (2016). Random forest in remote sensing: a review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing, 114,* 24-31. doi: 10.1016/j.isprsjprs.2016.01.011

Bressan, T. S., Souza, K. M., Girelli, T. J., Chemale, F. Jr. (2020). Evaluation of machine learning methods for lithology classification using geophysical data. *Computers and Geosciences, 139,* 104475. doi: 10.1016/j.cageo.2020.104475

Carmona, P., Climent, F., & Momparler, A. (2019). Predicting failure in the U.S. banking sector: an extreme gradient boosting approach. *International Review of Economics and Finance, 61,* 304-323. doi: 10.1016/j.iref.2018.03.008

Chen, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* Washington, United States of América.

Dubey, A., Kumar, A., Abd-Allah, E. F., Hashem, A., & Khan, M. L. (2019). Growing more with less: breeding and developing drought resilient soybean to improve food security. *Ecological Indicators, 105,* 425-437. doi: 10.1016/j.ecolind.2018.03.003

Falk, K. G., Jubery, T. Z., Mirnezami, S. V., Parmley, K. A., Sarkar, S., & Singh, A.,... Singh, A. K. (2020). Computer vision and machine learning enabled soybean root phenotyping pipeline. *Plant Methods, 16*(1), 1-19. doi: 10.1186/s13007-019-0550-5

Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X.,... Xiang, Y. (2018). Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: a case study in China. *Energy Conversion and Management, 164,* 102-111. doi: 10.1016/j.enconman.2018.02.087

Fenta, B., Beebe, S., Kunert, K., Burridge, J., Barlow, K., Lynch, J., & Foyer, C. (2014). Field phenotyping of soybean roots for drought stress tolerance. *Agronomy, 4*(3), 418-435. doi: 10.3390/agronomy4030418

García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2013). Classifying patterns with missing values using Multi-Task Learning perceptrons. *Expert Systems with Applications, 40*(4), 1333-1341. doi: 10.1016/j.eswa.2012.08.057

Goncalves, A., Ray, P., Soper, B., Widemann, D., Nygård, M., Nygård, J. F., & Sales, A. P. (2019). Bayesian multitask learning regression for heterogeneous patient cohorts. *Journal of Biomedical Informatics, 100,* (Suppl.), 100059. doi: 10.1016/j.yjbinx.2019.100059

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron, 105*(3), 416-434. doi: 10.1016/j.neuron.2019.12.002

Ito, K., Tanakamaru, K., Morita, S., Abe, J., & Inanaga, S. (2006). Lateral root development, including responses to soil drying, of maize (*Zea mays*) and wheat (*Triticum aestivum*) seminal roots. *Physiologia Plantarum, 127*(2), 260-267. doi: 10.1111/j.1399-3054.2006.00657.x

Iyer-Pascuzzi, A. S., Symonova, O., Mileyko, Y., Hao, Y., Belcher, H., Harer, J.,... Benfey, P. N. (2010). Imaging and analysis platform for automatic phenotyping and trait ranking of plant root systems. *Plant Physiology, 152*(3), 1148-1157. doi: 10.1104/pp.109.150748

Izquierdo-Verdiguier, E., & Zurita-Milla, R. (2020). An evaluation of guided regularized random forest for classification and regression tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation, 88,* 102051. doi: 10.1016/j.jag.2020.102051

Karandish, F., & Shahnazari, A. (2016). Soil temperature and maize nitrogen uptake improvement under partial root-zone drying irrigation. *Pedosphere, 26*(6), 872-886. doi: 10.1016/S1002-0160(15)60092-3

Krishnan, P., Singh, R., Verma, A. P. S., Joshi, D. K., & Singh, S. (2014). Changes in seed water status as characterized by NMR in developing soybean seed grown under moisture stress conditions. *Biochemical and Biophysical Research Communications, 444*(4), 485-490. doi: 10.1016/j.bbrc.2014.01.091

Laurett, L., Fernandes, A. A., Schmildt, E. R., Almeida, C. P., & Pinto, M. L. P. B. (2017). Desempenho da alface e da rúcula em diferentes concentrações de ferro na solução nutritiva. *Revista de Ciências Agrarias - Amazon Journal of Agricultural and Environmental Sciences, 60*(1), 45-52.

Ni, L., Wang, D., Wu, J., Wang, Y., Tao, Y., Zhang, J., & Liu, J. (2020). Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model. *Journal of Hydrology, 58,* 124901. doi: 10.1016/j.jhydrol.2020.124901

Park, C., Kim, Y., Park, Y., & Kim, S. B. (2018). Multitask learning for virtual metrology in semiconductor manufacturing systems. *Computers and Industrial Engineering, 123,* 209-219. doi: 10.1016/j.cie.2018.06.024

Patil, A. P., & Deka, P. C. (2016). An extreme learning machine approach for modeling evapotranspiration using extrinsic inputs. *Computers and Electronics in Agriculture, 121,* 385-392. doi: 10.1016/j.compag.2016.01.016

Puspasari, R., Hashiguchi, M., Ushio, R., Ishigaki, G., Tanaka, H., & Akashi, R. (2020). Evaluation of root traits in F2-progeny of interspecific hybrid between Lotus corniculatus "Super-Root" and tetraploid Lotus japonicus. *Plant and Soil, 446*(1-2), 613-625. doi: 10.1007/s11104-019-04332-2

Raghavendra, S., & Deka, P. C. (2014). Support vector machine applications in the field of hydrology: a review. *Applied Soft Computing Journal, 19,* 372-386. doi: 10.1016/j.asoc.2014.02.002

Rahmati, O., Falah, F., Dayal, K. S., Deo, R. C., Mohammadi, F., Biggs, T.,... Bui, D. T. (2020). Machine learning approaches for spatial modeling of agricultural droughts in the south-east region of Queensland Australia. *Science of the Total Environment, 699,* 134230. doi: 10.1016/j.scitotenv.2019.134230

Ratke, R. F., Santos, J. D. D. G. dos, & Souza, J. G. P. de. (2019). Métodos para estudo da dinâmica de raízes. In A. M., Zuffo, J. G. Aguilera, R. de & Oliveira (Orgs.), *Ciência em Foco* (Cap. 11, pp. 120-137). Nova Xavantina, MT: Pantanal Editora.

Ruiming, F., & Shijie, S. (2020). Daily reference evapotranspiration prediction of Tieguanyin tea plants based on mathematical morphology clustering and improved generalized regression neural network. *Agricultural Water Management, 236,* 106177. doi: 10.1016/j.agwat.2020.106177

Sandhu, N., Anitha Raman, K., Torres, R. O., Audebert, A., Dardou, A., Kumar, A., & Henry, A. (2016). Rice root architectural plasticity traits and genetic regions for adaptability to variable cultivation and stress conditions. *Plant Physiology, 171*(4), 2562-2576. doi: 10.1104/pp.16.00705

Santos Silva, P. P. dos, Sousa, M. B. e, & Oliveira, E. J. de. (2019). Prediction models and selection of agronomic and physiological traits for tolerance to water deficit in cassava. *Euphytica, 215*(4), 1-18. doi: 10.1007/s10681-019-2399-0

Sariyer, G., Öcal Taşar, C., & Cepe, G. E. (2019). Use of data mining techniques to classify length of stay of emergency department patients. *Bio-Algorithms and Med-Systems, 15*(1), 20180044. doi: 10.1515/bams-2018-0044

Saruta, K., Hirai, Y., Tanaka, K., Inoue, E., Okayasu, T., & Mitsuoka, M. (2013). Predictive models for yield and protein content of brown rice using support vector machine. *Computers and Electronics in Agriculture, 99,* 93-100. doi: 10.1016/j.compag.2013.09.003

Singh, D., Sisodia, D. S., & Singh, P. (2020). Compositional framework for multitask learning in the identification of cleavage sites of HIV-1 protease. *Journal of Biomedical Informatics, 102,* 103376. doi: 10.1016/j.jbi.2020.103376

Strock, C. F., De La Riva, L. M., & Lynch, J. P. (2018). Reduction in root secondary growth as a strategy for phosphorus acquisition. *Plant Physiology, 176*(1), 691-703. doi: 10.1104/pp.17.01583

Sun, N., Liu, C., Mei, X., Jiang, D., Wang, X., Dong, E., ..., & Cai, Y. (2020). QTL identification in backcross population for brace-root-related traits in maize. *Euphytica, 216*(2), 32. doi: 10.1007/s10681-020-2561-8

Tang, F. H., Chan, J. L. C., & Chan, B. K. L. (2019). Accurate age determination for adolescents using magnetic resonance imaging of the hand and wrist with an artificial neural network-based approach. *Journal of Digital Imaging, 32*(2), 283-289. doi: 10.1007/s10278-018-0135-2

Wang, M. B., & Zhang, Q. (2009). Issues in using the WinRHIZO system to determine physical characteristics of plant fine roots. *Shengtai Xuebao/ Acta Ecologica Sinica, 29*(2), 136-138. doi: 10.1016/j.chnaes.2009.05.007

Zhang, L., Traore, S., Ge, J., Li, Y., Wang, S., Zhu, G.,... Fipps, G. (2019). Using boosted tree regression and artificial neural networks to forecast upland rice yield under climate change in Sahel. *Computers and Electronics in Agriculture, 166,* 105031. doi: 10.1016/j.compag.2019.105031

Zhang, W., Wu, C., Zhong, H., Li, Y., & Wang, L. (2020). Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geoscience Frontiers, 12,* 469-477. doi: 10.1016/j.gsf.2020.03.007

Zhong, D., Novais, J., Grift, T. E., Bohn, M., & Han, J. (2009). Maize root complexity analysis using a support vector machine method. *Computers and Electronics in Agriculture, 69*(1), 46-50. doi: 10.1016/j.compag.2009.06.013

Zhong, R., Johnson, R., & Chen, Z. (2020). Generating pseudo density log from drilling and logging-while-drilling data using extreme gradient boosting (XGBoost). *International Journal of Coal Geology, 220,* 103416. doi: 10.1016/j.coal.2020.103416