frontiers
in Plant Science

ORIGINAL RESEARCH
published: 11 February 2022
doi: 10.3389/fpls.2022.821365

Poplar's Waterlogging Resistance Modeling and Evaluating: Exploring and Perfecting the Feasibility of Machine Learning Methods in Plant Science

Xuelin Xie[1†], Xinye Zhang[2†], Jingfang Shen[1] and Kebing Du[3]**

[1] College of Sciences, Huazhong Agricultural University, Wuhan, China, [2] Hubei Academy of Forestry, Wuhan, China, [3] College of Horticulture and Forestry Sciences, Hubei Engineering Technology Research Center for Forestry Information, Huazhong Agricultural University, Wuhan, China

Floods, as one of the most common disasters in the natural environment, have caused huge losses to human life and property. Predicting the flood resistance of poplar can effectively help researchers select seedlings scientifically and resist floods precisely. Using machine learning algorithms, models of poplar's waterlogging tolerance were established and evaluated. First of all, the evaluation indexes of poplar's waterlogging tolerance were analyzed and determined. Then, significance testing, correlation analysis, and three feature selection algorithms (Hierarchical clustering, Lasso, and Stepwise regression) were used to screen photosynthesis, chlorophyll fluorescence, and environmental parameters. Based on this, four machine learning methods, BP neural network regression (BPR), extreme learning machine regression (ELMR), support vector regression (SVR), and random forest regression (RFR) were used to predict the flood resistance of poplar. The results show that random forest regression (RFR) and support vector regression (SVR) have high precision. On the test set, the coefficient of determination (R^2) is 0.8351 and 0.6864, the root mean square error (RMSE) is 0.2016 and 0.2780, and the mean absolute error (MAE) is 0.1782 and 0.2031, respectively. Therefore, random forest regression (RFR) and support vector regression (SVR) can be given priority to predict poplar flood resistance.

Keywords: flood disaster, prediction of waterlogging tolerance, machine learning, feature selection, model establishment and evaluation

OPEN ACCESS

Edited by:
Khawar Jabran,
Niğde Ömer Halisdemir University,
Turkey

Reviewed by:
Milind Ratnaparkhe,
ICAR Indian Institute of Soybean Research, India
Shixiong Zhang,
Xidian University, China

****Correspondence:***
Jingfang Shen
sjf_712@mail.hzau.edu.cn
Kebing Du
kebingdu@mail.hzau.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Sustainable and Intelligent Phytoprotection, a section of the journal Frontiers in Plant Science

***Received:** 01 December 2021*
***Accepted:** 20 January 2022*
***Published:** 11 February 2022*

Citation:
Xie X, Zhang X, Shen J and Du K (2022) Poplar's Waterlogging Resistance Modeling and Evaluating: Exploring and Perfecting the Feasibility of Machine Learning Methods in Plant Science. Front. Plant Sci. 13:821365. doi: 10.3389/fpls.2022.821365

INTRODUCTION

Natural disasters are inherently a phenomenon that has adverse consequences for society (Paprotny et al., 2018). It damages the living environment and life of human beings. Flood disasters, as one of the most common and expensive natural disasters, have caused huge losses to human lives and property (Hu et al., 2018; Ao et al., 2020). With the development of social industry and economy, the warming of the atmosphere caused by greenhouse gas emissions may increase the risk of river flooding (Hallegatte et al., 2013; Hirabayashi et al., 2013; Willner et al., 2018;

Bloeschl et al., 2019). Therefore, many studies want to build a system for predicting flood risk (Alfieri et al., 2017; Shafizadeh-Moghadam et al., 2018; Choubin et al., 2019; Khosravi et al., 2019), and a variety of machine learning methods are used in these studies. Choubin et al. (2019) used multivariate discriminant analysis (MDA), classification and regression trees (CART), and support vector machine (SVM) algorithms to predict flood risk in Iran's Khiyav Chai drainage basin. The results show that the residential areas at the outlet of the drainage basin are very susceptible to floods. Khosravi et al. (2019) adopted three Multi-Criteria Decision-Making techniques (VIKOR, TOPSIS, and SAW) and two Machine Learning methods (NBT and NB) to test the flood sensitivity modeling of the Ningdu River Basin in China. Finally, their research shows that the NBT model is a powerful tool for evaluating flood-prone areas, and can properly plan and manage flood disasters. Nevertheless, predicting flood risk cannot substantially reduce the life and economic losses of human society. Afforestation can strengthen the stability of water, soil, and carbon sinks in the forest ecosystem, thereby effectively coordinating the relationship between humans and the natural environment. A considerable number of studies have shown that afforestation can weaken the impact of global warming and effectively reduce the risk of river flooding (Hong et al., 2018, 2020; Liu X. et al., 2018; Forster et al., 2021). Thus, afforestation is widely used to resist flood disasters.

Plants have evolved numerous resistance mechanisms to resist flood disasters, including plant morphological Screening of Candidate Genesristics, metabolic responses, and molecular transcriptional regulation (Loreti et al., 2016; Du et al., 2017; Yin et al., 2017; Zeng et al., 2019; Lukic et al., 2020; Lee et al., 2021). Among the diverse plant populations, poplar has become the main flood-resistant tree varieties in flood-prone areas due to its rapid growth and flood resistance features. Many studies have shown that the root system is the key organ of poplar responding to Flooding stress (Coleman et al., 2000; Major and Constabel, 2007; Berhongaray et al., 2013; Ye et al., 2018; Gerjets et al., 2021). Flooding stress affects the diffusion of oxygen in plant root tissues. At the same time, it limits the mitochondrial respiration of root cells and accumulates toxic substances, which seriously affects its normal physiological activities (Arbona et al., 2008; Voesenek and Bailey-Serres, 2013; Tian et al., 2019). In addition, flooding stress will destroy the photosynthesis performance of plants, which will inhibit plant growth and biomass accumulation (Du et al., 2012; Zhu et al., 2016; Zheng et al., 2017; Xiong et al., 2019; Zhou et al., 2020). Flooding stress not only reduces the chlorophyll content of plants, but also reduces the carotenoid content (Zhou et al., 2017). Kreuzwieser et al. (2009) found that the metabolite changes occurred in leaves and roots of submerged poplar. Du et al. (2012) compared the physiological and morphological adaptability of two poplar clones (hypoxia-resistant and hypoxia-sensitive) to flooding, and Peng et al. (2018) monitored the different response mechanisms of these two clones of poplar to flooding stress. These studies have greatly promoted people's understanding of the waterlogging resistance mechanism, and to a considerable extent, strengthened people's resistance to flood risks. Thus far, there are still few studies on the influence of poplar on the waterlogging resistance factors. These factors include the intrinsic features of poplar trees (photosynthesis and chlorophyll fluorescence, etc.) and external environmental features (ambient temperature, humidity, etc.). As a popular research direction, machine learning has recently been gradually introduced into the field of plant science. For the research on the resistance of poplar to waterlogging, Xie and Shen (2021) used poplar photosynthesis features and external environmental factors to predict the waterlogging tolerance of poplar. By using the SVR method in machine learning, they confirmed the feasibility of applying photosynthesis and other characteristic parameters to predict poplar flood resistance. However, previous prediction studies did not consider important parameters such as chlorophyll fluorescence. Additionally, the related forecasting research is not systematic enough, and the corresponding investigation and research are still lacking.

Based on the above considerations, the main purpose of this article is to consider more comprehensive feature parameters and use a variety of machine learning methods to predict the flood resistance of poplar. At the same time, it aims to supplement and improve the key content and procedures of poplar flood resistance prediction. First of all, the evaluation indicators of waterlogging tolerance were well defined and explained. Then, 26 internal characteristics and external environmental factors of poplars were screened by using feature selection algorithms such as significance test and stepwise regression. Finally, four machine learning methods were used to establish the flood resistance models of poplar, and the results were comprehensively evaluated in detail. Compared with previous studies, this study supplements the evaluation index and prediction system of poplar waterlogging tolerance. The main contribution is that the definition and analysis of evaluation indicators for waterlogging tolerance have been improved, and more comprehensive characteristic parameters have been considered. Moreover, the feature selection, prediction methods, and evaluation indicators were adjusted, and more machine learning methods and results have been considered and analyzed. This research has enriched the prediction of poplar's flood resistance, which is of great significance to poplar's accurate flood resistance, intelligent selection of seedlings, and cultivation of high-quality saplings. Furthermore, to a considerable extent, it promotes the research of flood resistance mechanisms, which have great theoretical and practical value.

MATERIALS AND METHODS

Experimental Area and Materials

Research area: Huazhong Agricultural University, Wuhan, China (114°35′E, 30°49′E), subtropical humid monsoon climate. This area has four distinct seasons, with plenty of sunshine and plenty of rainfall. The annual average temperature is 15.8–17.5°C, rainfall is 1,269 mm, and total sunshine hours are between 1,810 h to 2,100 h.

Experimental objects: There were 20 poplar varieties in total. The scientific names corresponding to the 20 poplar varieties are shown in **Table 1**.

TABLE 1 | Scientific names of 20 poplar varieties.

Varieties	Scientific names
LS68	*Populus deltoides "Lux" × P. simonii (LS68)*
LS81	*P. deltoides "Lux" × P. simonii (LS81)*
NL895	*P. × euramericana "Nanlin 895"*
I-63	*P. deltoides "Harvard"*
I-69	*P. deltoides "Lux"*
I-72	*P. euramaricana "an Martino"*
I-214	*P. × euramericana "I-214"*
I-45-51	*P. × euramaricana "I-45/51"*
Flevo	*P. euramericana "Flevo"*
Juba	*P. deltoides "55/56" × P. deltoides "2KEN8"*
LH04-13	*P. deltoides "Lux" × P. deltoides "Harvard" (LH04-13)*
LH04-17	*P. deltoides "Lux" × P. deltoides "Harvard" (LH04-17)*
Triplo	*P. euramericana "Triplo"*
DD102-4	*P. deltoides "DD102-4"*
Raspalje	*P. deltoides "Raspalje"*
Danhong	*P. deltoides "Danhong"*
Canadensis	*P. canadensis Moench.*
2L2025	*P. deltoides "Lux" × P. deltoides "Shanhaiguan"*
Ningshanica	*P. ningshanica*
Lushan	*P. × liaoningensis*

Experimental Process and Parameter Measurement

The 1-year-old branches of 20 poplar clones were cut into about 15 cm cuttings with 3–4 buds. There were 4 experimental groups and 4 control groups for each variety, with a total of 160 experimental materials. After being soaked in water for 24 h, the cuttings were planted in mixed soil. The container was seedling pots (150 mm × 100 mm × 130 mm), and the soil was 1:1 substrate soil and peat soil (The soil consisted of 2–5% N, P_2O_5 and K_2O, pH = 6.2, total organic matter of nutrient soil was ≥ 28%, and the total nutrient was ≥ 2%). The morphological changes of the plants were observed every day, including the chlorosis and shedding of leaves. We measured the height, biomass, photosynthesis, and chlorophyll fluorescence parameters of poplar seedlings on the 0th and 60th days. The characteristic parameters were measured by the LI-6400 photosynthesis analyzer (LI-COR, Lincoln, NE, United States), and the time was concentrated between 9:00 am and 11:30 am. In the experiment, a standard LI-COR leaf chamber and red and blue light sources (6400-02 LED light sources) were used. The light intensity was 1,000 $\mu mol \cdot m^{-2} \cdot s^{-1}$, and the air velocity was 500 $\mu mol \cdot s^{-1}$. 26 characteristic parameters of poplar samples were measured, including photosynthesis, chlorophyll fluorescence features, and environmental variables. The specific information of these features is shown in **Appendix Table A1**, and the treatment process of the experimental group and the control group is as follows.

- Control group: Watered normally (CK). There were drainage holes at the bottom of the flower pots in the Control group. Watered the plants according to the needs of normal plant growth, and the soil moisture was maintained at about 75% of the maximum water holding capacity in the field.
- Experimental group: Shallow flooded (FL). The waterlogging test was started 5–6 weeks after cuttings, and the water surface was 10 cm higher than the soil surface. The experiment lasted for 60 days, of which, the flooding time was 45 days, and the drainage recovery time was 15 days.

Programming Environment

In this article, R 4.0.5 was used to perform data Processing and Feature selection process, and MATLAB R2018a was used to implement the Model building and evaluation.

METHODOLOGY

The methodology is divided into data processing, feature selection, model establishment and evaluation. The main procedures are shown in **Figure 1**, and the specific implementation steps will be introduced one by one below.

Data Processing

Evaluation Index of Waterlogging Tolerance

The changes in biomass and seedling height can reflect the waterlogging tolerance of plants. In previous studies, Xie and Shen (2021) proposed the waterlogging tolerance evaluation index Zscore. This article supplemented the definition of the other two waterlogging tolerance evaluation indicators, and used the three waterlogging tolerance evaluation indicators for outlier analysis. Finally, the most suitable evaluation index for waterlogging tolerance was selected. The definitions of the three evaluation indicators are given below.

The first evaluation index for waterlogging tolerance is Zbio, which is obtained based on changes in biomass. This indicator is based on the change in biomass of the test group within 60 days to judge the flood resistance of poplar, and it is dimensionless. The stronger the waterlogging resistance performance, the larger the corresponding Zbio. The calculation method is shown in Formula (1):

$$Zbio = \frac{bio(x_i) - E(bio)}{Std(bio)} \tag{1}$$

The second waterlogging tolerance evaluation index is Zsap, which is similar to the definition of Zbio. This index only considers the change of poplar seedling height, and its calculation method is shown in Formula (2):

$$Zsap = \frac{sap(x_i) - E(sap)}{Std(sap)} \tag{2}$$

The third evaluation index of waterlogging tolerance is Zscore. This indicator takes into account the changes in biomass, as well as changes in seedling height. Compared with Zbio and Zsap, this index can more comprehensively reflect the flood resistance of poplar, and its calculation formula is shown in Formula (3):

$$Zscore(x_i) = \omega_{bio} \times Zbio + \omega_{sap} \times Zsap \tag{3}$$

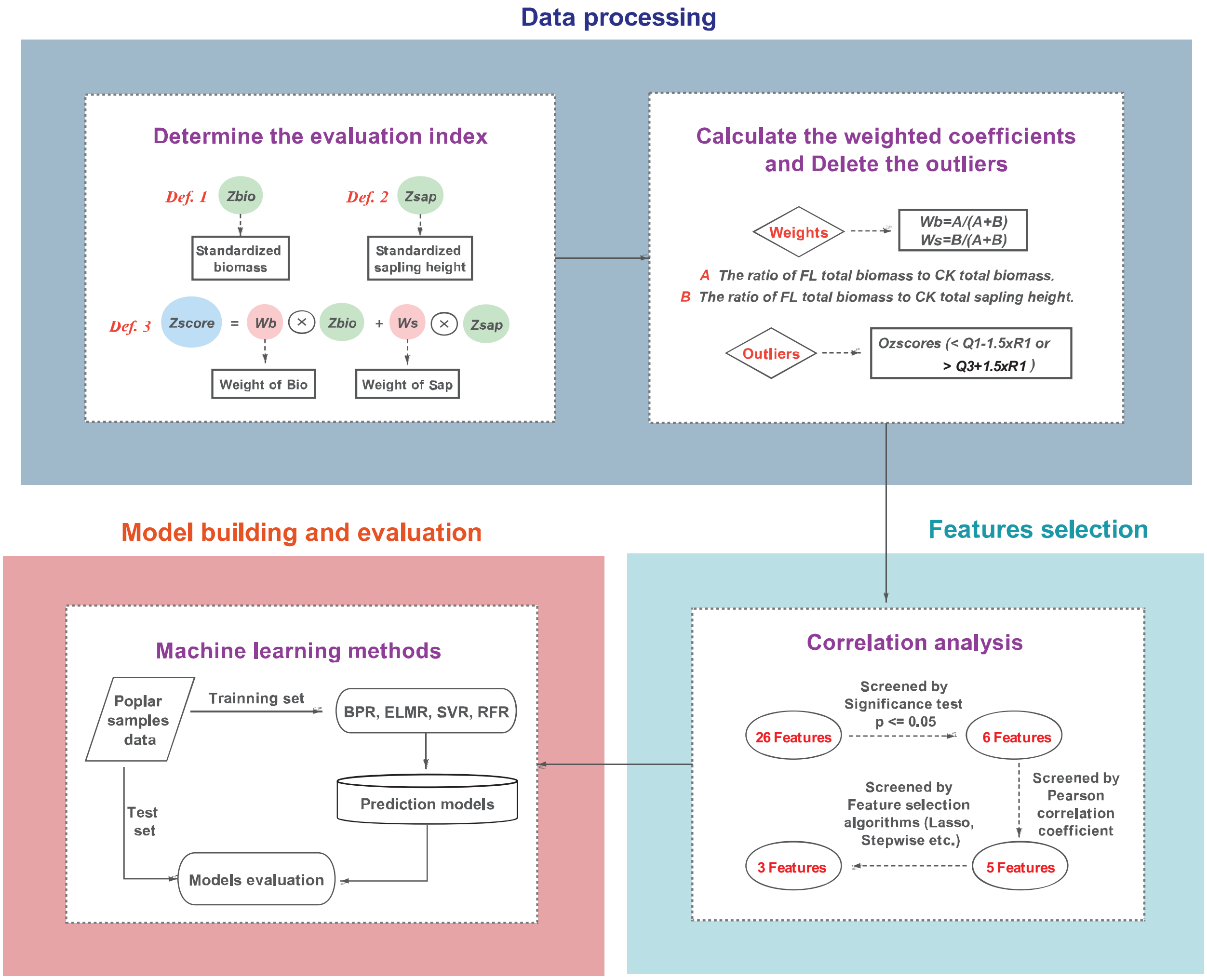

Data processing
Determine the evaluation index
Def. 1
Zbio
Standardized biomass
Def. 2
Zsap
Standardized sapling height
Def. 3
Zscore
=
Wb
×
Zbio
+
Ws
×
Zsap
Weight of Bio
Weight of Sap
Calculate the weighted coefficients and Delete the outliers
Weights
Wb=A/(A+B)
Ws=B/(A+B)
A The ratio of FL total biomass to CK total biomass.
B The ratio of FL total biomass to CK total sapling height.
Outliers
Ozscores (< Q1-1.5xR1 or > Q3+1.5xR1)
Model building and evaluation
Features selection
Machine learning methods
Poplar samples data
Trainning set
BPR, ELMR, SVR, RFR
Prediction models
Test set
Models evaluation
Correlation analysis
Screened by Significance test p <= 0.05
26 Features
6 Features
Screened by Pearson correlation coefficient
Screened by Feature selection algorithms (Lasso, Stepwise etc.)
3 Features
5 Features

FIGURE 1 | Flow chart of Methodology.

where ω_{bio} and ω_{sap} are the weight coefficients, which satisfy the condition $\omega_{bio}\omega_{sap} = 1$. The calculation method of the weight is shown in equation (4). Zbio and Zsap are the two evaluation indicators mentioned in above.

$$\omega_{bio} = \frac{A}{A+B}, \ \omega_{sap} = \frac{B}{A+B} \tag{4}$$

where $A = \frac{FL(Sum(bio(x_i))}{CK(Sum(bio(x_i))}$, $B = \frac{FL(Sum(sap(x_i))}{CK(Sum(sap(x_i))}$.

Treatment of Outliers

Extremely different from other observations, the outliers often cause anomalies (Aggarwal and Yu, 2005). Outliers may affect the accuracy of the final model (Domingues et al., 2018; Zhao et al., 2020). Consequently, before feature selection and models establishment, outliers in the data should be eliminated. The outlier Ozscore is defined in formula (5):

$$\text{Ozscore} > Q_3 + 1.5 \times R_1 \ or \ \text{Ozscore} < Q_1 - 1.5 \times R_1 \tag{5}$$

where Q_3 and Q_1 are the upper and lower quartiles, and the quartile range $R_1 = Q_3 - Q_1$.

Feature Selection

Feature selection is to effectively remove irrelevant and redundant features (Arora and Anand, 2019; Sayed et al., 2019). It can improve the performance of the model and reduce the cost of calculation (Li et al., 2018; Angulo and Shin, 2019). The 26 characteristic parameters considered in this study meet the conditions of multi-dimensional data. Therefore, these features need to be selected.

Hierarchical Clustering

Hierarchical clustering is a clustering method used to describe the hierarchical structure of samples in a group (Wu et al., 2009). The result of hierarchical clustering is usually represented by a dendrogram. The tree diagram shows the organization and relationship of the sample in the form of a tree, which is convenient for people to divide intuitively (Granato et al., 2018).

For related clustering research work, refer to Xu and Wunsch (2005) and Murtagh and Contreras (2012). A hierarchical clustering method was adopted to cluster the poplar varieties and the five features selected by correlation analysis, and the measurement method was Euclidean distance.

Lasso and Stepwise Regression

The Lasso method is proposed by Tibshirani (1996) by combining the advantages of both ridge regression and subset selection meth. It not only has the interpretability of subset selection, but also has the stability of ridge regression. To achieve the purpose of feature selection, this method compresses the coefficients of insignificant variables to 0 (Zou and Hastie, 2005; Cui and Gong, 2018). Stepwise regression uses collinearity and variance contribution tests to gradually find all the significant features, thereby obtaining the optimal model. The basic idea of stepwise regression is to add new variables one by one, each time a new variable is added, consider whether to eliminate the selected variable until no more variables are introduced. Stepwise regression is mainly used to solve the problem of multicollinearity. For related research, refer to Guidolin and Pedio (2021), Ou et al. (2016), and Yang et al. (2019).

Establishment and Evaluation of Regression Model

Machine Learning Methods

BP Neural Network

BP neural network is a multi-layer network structure composed of an input layer, an output layer, and one or more hidden layers (Yang et al., 2018), which can effectively deal with linear and non-linear relationships between data (Moghadassi et al., 2010). BP is called the error back propagation algorithm. In essence, the BP algorithm takes the error square as the objective function, and uses the gradient descent method to calculate the minimum value of the objective function. BP neural network can systematically solve the hidden layer connection weight learning problem of multilayer neural network, and it is one of the most widely used neural networks at present.

Extreme Learning Machine

The extreme learning machine is a new single hidden layer feedforward neural network (Ding et al., 2015). This algorithm can produce good generalization performance in most cases, and its learning speed is thousands of times faster than the traditional feedforward neural network algorithm. Therefore, many studies apply extreme learning machines for regression and prediction (Miche et al., 2010; Huang et al., 2011; Yao and Ge, 2018; Yaseen et al., 2018).

Support Vector Regression

Support vector machine (SVM) is a supervised machine learning method proposed by Cortes and Vapnik (1995) in the mid-1990s, which is used to deal with binary classification problems. The core idea of SVM is to find a hyperplane or hypersurface to segment the sample points to maximize the interval between the segmentation points. Support vector regression (SVR) is an application model of support vector machine (SVM) on regression problems (Demir and Bruzzone, 2014). As a classic regression algorithm in machine learning, support vector regression has been widely used in many fields, such as plant science, data mining, and biomedicine (Khosravi et al., 2018; Moazenzadeh et al., 2018; Zhuo et al., 2018; Han et al., 2019; Mishra and Padhy, 2019).

Random Forest Regression

Random forests produce reliable classifications by using predictions from a set of decision trees (Breiman, 2001). It is composed of multiple decision trees, and there is no correlation between each decision tree. The final output of the model is jointly determined by each decision tree in the forest. When dealing with regression problems, random forest uses the mean value of each decision tree as the final result. Due to the excellent regression results and the relatively fast processing speed, the use of random forest regression has also received extensive attention (Du et al., 2015; Chen et al., 2018; Dou et al., 2019).

The relationship between variables is often non-linear. Thus, compared with traditional linear regression, machine learning algorithms may have higher accuracy. There may be a non-linear relationship between poplar resistance to flooding and features. Consequently, the four machine learning methods mentioned above will be used to predict the waterlogging resistance of poplar.

Model Parameters

Manual tuning is the traditional method of adjusting the hyperparameters of machine learning models (Yang and Shami, 2020). With the improvement of automatic optimization methods, grid search (GS), particle swarm optimization (PSO), genetic algorithm (GA), and other optimization methods were proposed to find the optimal hyperparameters. to find the optimal hyperparameters. However, there are still problems such as complex optimization processes and slow convergence speed. Based on experience, we have selected the parameters of the four machine learning methods. The parameter sensitivity and parameter selection of each method will be analyzed below.

There are many kinds of training functions for the BPR algorithm, and most of the data sets are very sensitive to the training function. In the experiment, a variety of training functions were selected. Compared with other training functions such as trainlm function (based on Levenberg-Marquardt algorithm), the trainbr function based on Bayes rule has better network generalization ability and higher accuracy. Hence, the trainbr function was finally used in the BPR method. In addition, previous studies have shown that the number of hidden layer nodes is a key factor affecting the accuracy of BPR and ELMR models (Liu Z. T. et al., 2018; Zhang et al., 2018). For the number of hidden layer nodes, 3, 5, 7, 9, and 11 hidden layer nodes were used to train the BPR model, 2, 3, 4, 5, and 6 hidden layer nodes were used to train the ELMR model. The root mean square error (RMSE) of the training is shown in **Table 2**. When the hidden layer nodes of the BPR and ELMR methods were 9 and 5, respectively, the RMSE was considered to be the smallest. Therefore, the number of hidden layer nodes of BPR was set to 9, and the number of hidden layer nodes of ELMR was set to 5.

TABLE 2 | RMSE of models with different Nodes or *Mtry*.

BPR		ELMR		RFR	
Nodes	**RMSE**	**Nodes**	**RMSE**	**Mtry**	**RMSE**
3	0.5654	2	0.3836	1	0.1940
5	0.4221	3	0.3785	2	0.2654
7	0.3703	4	0.3591	3	0.3210
9	0.3680	5	0.3426	4	0.3558
11	0.3939	6	0.3438	5	0.3982

TABLE 3 | Machine learning model parameters.

Methods	Model parameter
BPR	Training function: trainbr
	Number of input layers: 3
	Number of hidden layers: 9
	Number of output layers: 1
	Transfer function: logsig, purelin (Input-Hidden, Hidden-Output)
	net.trainparam.goal: 0.0001
	net.trainparam.lr: 0.01
	net.trainparam.epochs: 1000
ELMR	Training function: elmtrain
	Number of input layers: 3
	Number of hidden layers: 5
	Number of output layers: 1
	Activation function: sigmoid
SVR	Training function: svmtrain
	Model: ε-SVR
	Kernel function: RBF
	Regularization parameter C: 65
	Gamma: 0.001
	p: 0.01
RFR	Training function: TreeBagger
	Number of decision trees: 200
	Minimum number of leaves: 1
	Fraction of in-bag observations (FBoot): 1

For the SVR method, two SVR models (nu-SVR and epsilon-SVR) and four kernel functions (linear, polynomial, sigmoid, and radial basis functions) of the LibSVM toolbox were selected. Due to the higher precision of the model on the training set, the epsilon-SVR model (ε-SVR, a model that minimizes the RMSE) based on the RBF kernel function was finally selected. The regularization parameter C and the penalty coefficient gamma were determined by fivefold cross validation. The minimum number of leaves (*Mtry*) is the sensitive parameter of the RFR model, and the value of *Mtry* is generally set to 2 (Probst et al., 2018). In the experiment, we set the value of *Mtry* to 1, 2, 3, 4, and 5. The RMSE of the training function is shown in **Table 2**. When the value of *Mtry* was 1, RMSE was considered to be the smallest. Therefore, the value of *Mtry* was set to 1. Other parameters of the machine learning method were set as common parameters. The specific values of the parameters are shown in **Table 3**.

Evaluation of Model Performance

The three evaluation indexes of coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE) were used to evaluate the performance of the model. The corresponding calculation formulas are shown in (6)-(8):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y_{\mathrm{i}}})^2}{\sum_{i=1}^{n} (y_i - \overline{y_{\mathrm{i}}})^2} \tag{6}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \widehat{y_{\mathrm{i}}})^2}{n}} \tag{7}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \widehat{y_{\mathrm{i}}} \right| \tag{8}$$

where n is the number of varieties, y_i is the actual value, $\widehat{y_{\mathrm{i}}}$ is the predicted value, and $\overline{y_{\mathrm{i}}}$ is the mean of the true y_i.

RESULTS

Treatment of Outliers and Selection of Evaluation Index

Three evaluation indicators are used to deal with outliers in the data. The calculated descriptive statistics are shown in **Table 4**, where Max and Min are the maximum and minimum values, and Med is the median. The results of deleting outliers are shown in **Figure 2**.

As shown in **Figure 2**, the points outside the red dotted line in the figure are outliers. It can be observed that for all samples, the defined Zscore roughly ranges from [−2, 2], while the ranges of Zbio and Zsap are larger than Zscore, and only Zscore has outliers. In addition, from the definition of the waterlogging tolerance evaluation index, we know that Zscore not only considers the biomass but also the change of seedling height, which can more comprehensively reflect the flood resistance of poplar. Thus, based on the above viewpoints, Zscore was finally selected as the waterlogging tolerance evaluation index in this article.

Screening of Features

Significance Test and Correlation Analysis

According to the significance level of the correlation between the features and the poplar waterlogging tolerance score Zscores, 6 features were selected from 26 features, and these 6 features were all established under the condition that the significance level $p = 0.05$. We calculated the Pearson correlation coefficient, and

TABLE 4 | Descriptive statistics of the three evaluation indicators.

Methods	Min	Q_1	Med	Q_3	Max
Zbio	−2.409917	−0.748933	−0.095463	0.615642	2.614667
Zsap	−1.984857	−0.799594	−0.033385	0.722554	2.419308
Zscore	−2.076712	−0.554362	−0.039611	0.466923	2.257776

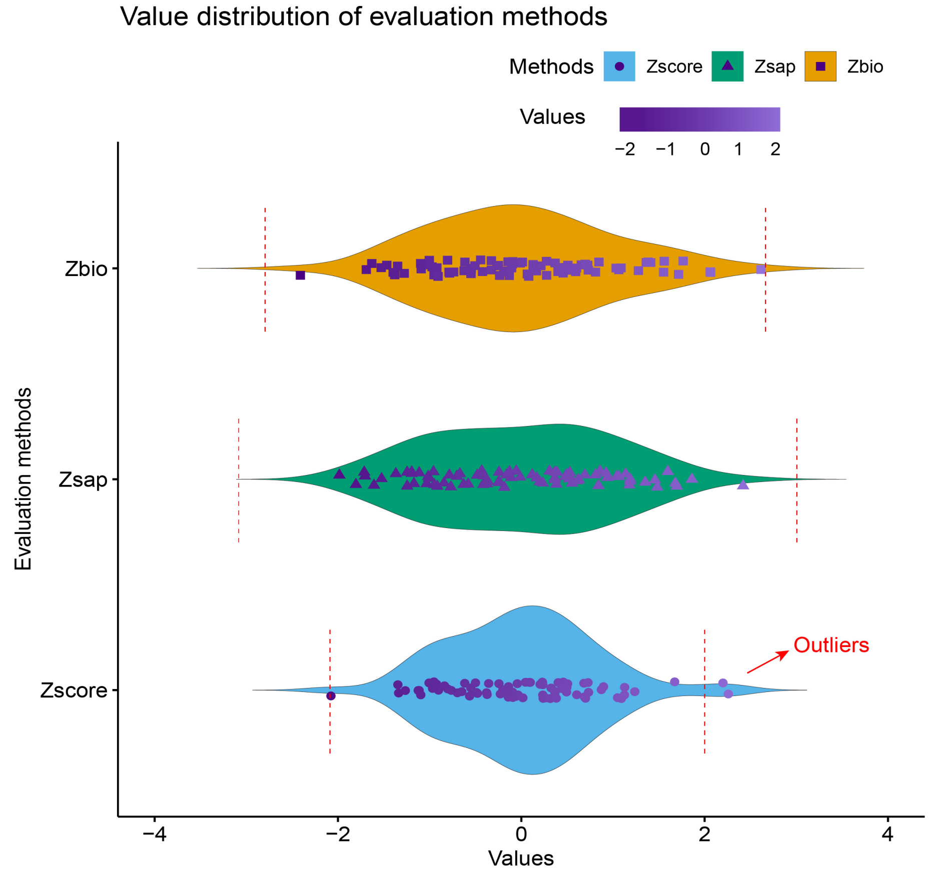


FIGURE 2 | Distribution of three evaluation indexes.

the results are shown in **Figure 3**. **Figure 3A** is the heat map of 26 features, the blank part is the case of $p = 0.05$, that is, it is not significant. **Figure 3B** is a heat map of the correlation coefficient that satisfies the condition of $p = 0.05$, and **Figure 3C** is the exact value of Pearson's correlation coefficient between 6 significant features. The specific meanings corresponding to the 6 significant features are shown in **Appendix Table A1**.

From **Figure 3C**, it can be found that the correlation between qN_Fo and H_2OS is particularly strong. The correlation coefficient between them exceeds 0.8. Thus, the feature with the largest coefficient is selected from these related features, and the highly related features are excluded. After this operation, the retained features are Fv, qN_Fo, Fm, H_2OS/H_2OR, and RH_S/RH_R.

Before establishing the regression model, univariate regression prediction was carried out on the features of significance test and correlation screening, and the result is shown in **Figure 4**. It can be observed that the five variables all meet the significance level of $p = 0.05$, and there is a considerable proportional relationship between them. Nevertheless, the results of univariate regression were general, and the highest coefficient of determination (R^2) is 0.57. For this reason, other methods should be chosen for regression analysis, such as multiple linear regression and machine learning regression methods.

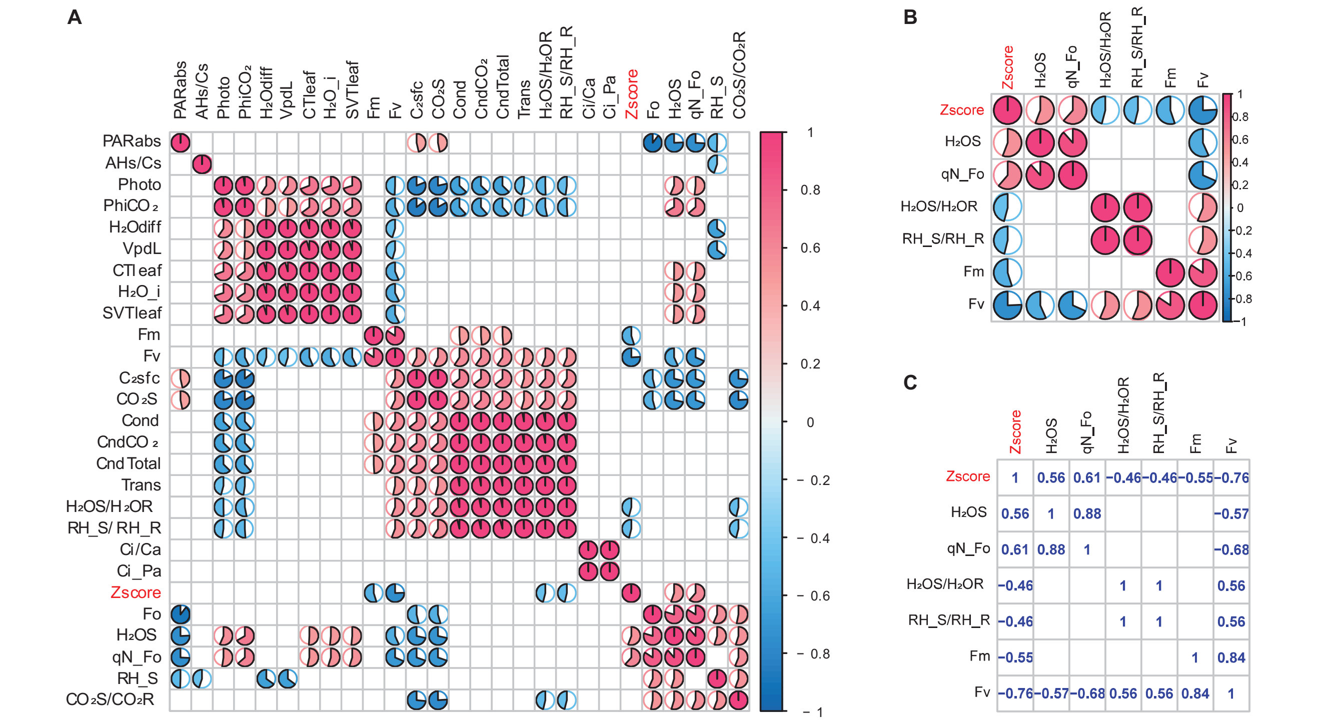
It is undeniable that the 5 features of significance testing and correlation screening may still have multicollinearity. To implement machine learning modeling more reasonably and accurately, three methods of hierarchical clustering, Lasso, and Stepwise regression were adopted for further feature selection. Before predicting the waterlogging tolerance of different poplar varieties and further feature screening, the characteristic parameters and Zscore of each sample were averaged according to the variety.

Clustering Results

The results of hierarchical clustering are shown in **Figure 5**. **Figure 5A** is the total clustering heat map, **Figure 5B** is the poplar varieties clustering, and **Figure 5C** is the poplar characteristic clustering.According to the clustering results in **Figure 5**, we can divide poplar varieties and features into 3 groups. The classification of poplar varieties is marked as A, B, and C, and the classification of characteristic parameters is marked as F1, F2, and F3.

Results of Lasso and Stepwise Regression

Lasso regression and backward stepwise regression are used to screen the 5 features (Fv, qN_Fo, Fm, H_2OS/H_2OR, and RH_S/RH_R) obtained by significance and correlation. The



	Zscore	H₂OS	qN_Fo	H₂OS/H₂OR	RH_S/RH_R	Fm	Fv
Zscore	1	0.56	0.61	−0.46	−0.46	−0.55	−0.76
H₂OS	0.56	1	0.88				−0.57
qN_Fo	0.61	0.88	1				−0.68
H₂OS/H₂OR	−0.46			1	1		0.56
RH_S/RH_R	−0.46			1	1		0.56
Fm	−0.55					1	0.84
Fv	−0.76	−0.57	−0.68	0.56	0.56	0.84	1

FIGURE 3 | Heat map of Pearson's correlation coefficient. **(A)** Heat map of 26 features. **(B)** Heat map of 6 features. **(C)** Specific values of 6 features.

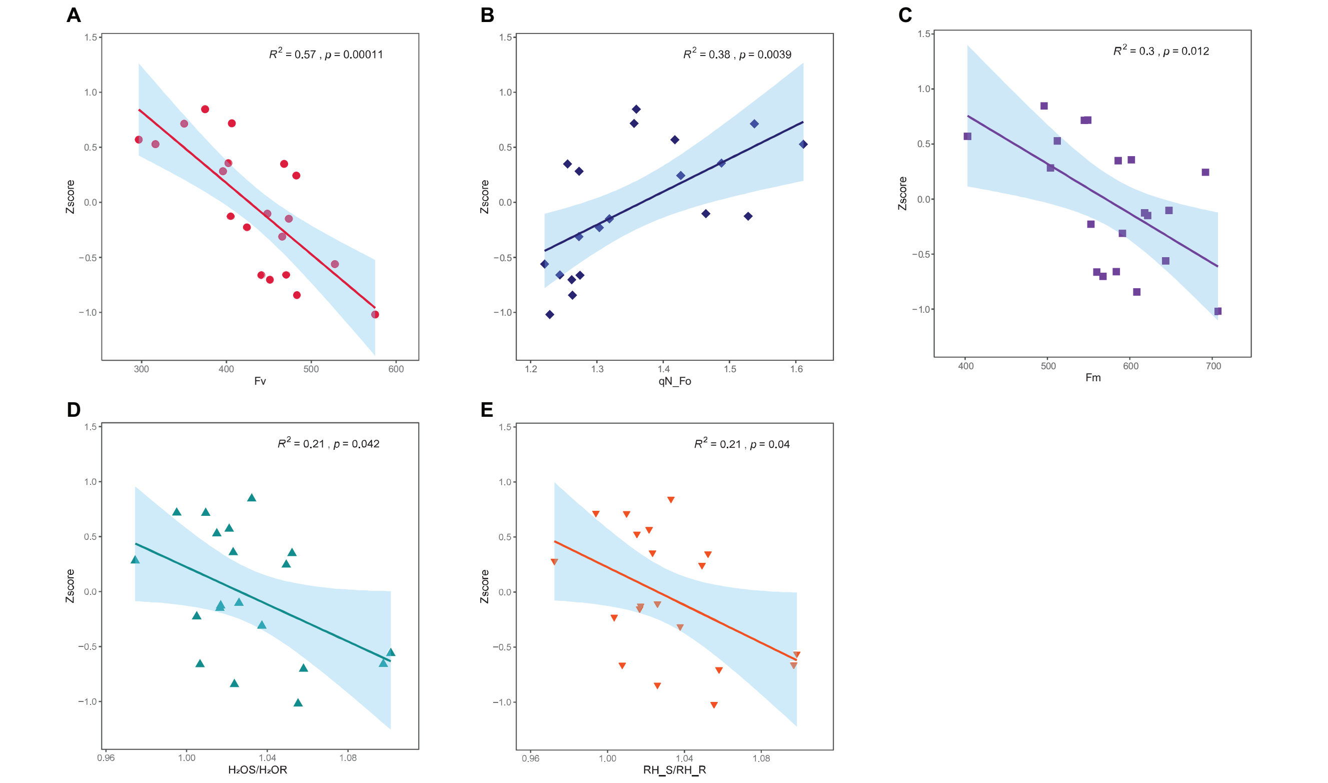


FIGURE 4 | Regression prediction of single feature. **(A)** Fv. **(B)** qN_Fo. **(C)** Fm. **(D)** H_2OS/H_2OR. **(E)** RH_S/RH_R.

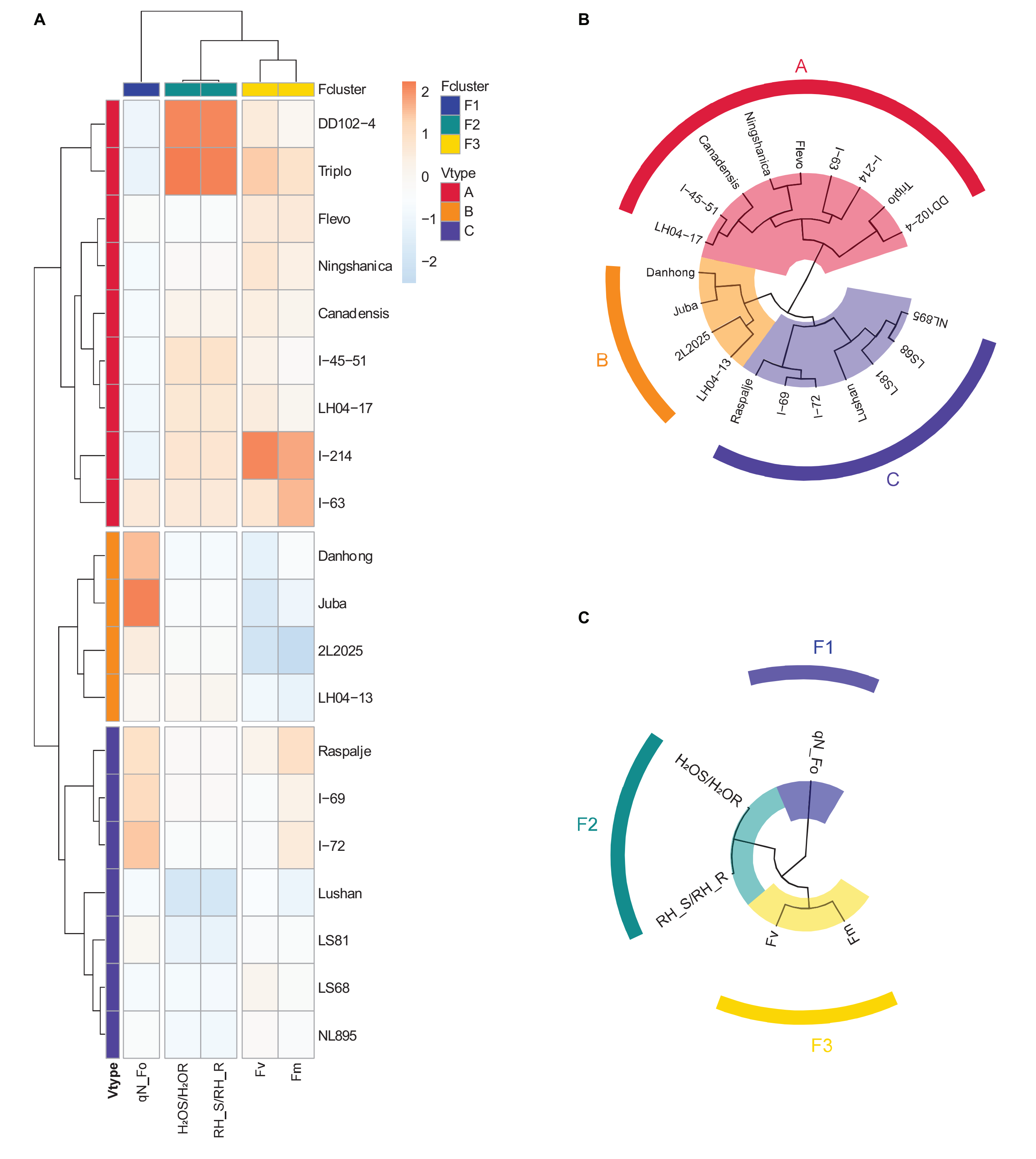


FIGURE 5 | Results of poplar features and varieties clustering. **(A)** Total. **(B)** Varieties. **(C)** Features.

results of the Lasso method are Fv, qN_Fo, and RH_S/RH_R. However, the screening result of stepwise regression only has the variable Fv. From the univariate regression analysis results in **Figure 4**, we know that the coefficient of determination (R^2) of Fv is 0.57. A single feature used for regression may lack interpretability and may affect the accuracy of the final model. In addition, according to the results of hierarchical clustering in **Figure 5**, a feature with the largest correlation coefficient was selected from each of the three groups (F1, F2, and F3), and the results obtained are consistent with the Lasso method. Therefore, combining the results of hierarchical clustering and univariate analysis, in the final machine learning modeling, we used the three characteristic parameters of Fv, qN_Fo and RH_S/RH_R.

Regression Results of Machine Learning Models

The Division of Test Set and Training Set

Before establishing the machine learning regression model, the poplar varieties were divided into training set and test set according to the ratio of 4:1 (the training set had 16 varieties, and the test set was 4 varieties). The four poplar varieties in the test set were selected from the three groups of A, B, C by stratified sampling based on the poplar hierarchical clustering results. The poplar varieties used and their corresponding Zscore and Vtype are shown in **Table 5**.

Training Set

Four machine learning regression methods were used to perform regression prediction on the three screened features (Fv, qN_Fo and RH_S/RH_R). The results obtained on the training set, and the corresponding R^2, RMSE, and MAE are shown in **Figure 6** and **Table 6**.**Figure 6D** is a histogram of model evaluation indexes (R^2, RMSE, and MAE) of four machine learning methods on the training set. The colored columns correspond to the four machine learning methods of BPR, ELMR, SVR, and RFR, respectively. From the first subplot of **Figure 6D**, it can be noticed that on the training set, the highest R^2 of the four machine learning methods is random forest regression (RFR). Specifically, from **Figure 6B** and **Table 6**, we can observe that the coefficient of determination (R^2) of RFR is 0.8847. Then, the second one is support vector regression (SVR), the R^2 is 0.7027. In contrast, the performance of BP neural network regression (BPR) and Extreme learning machine regression (ELMR) methods are relatively poor, and their R^2 are 0.5847 and 0.6401, respectively. In addition, from **Figure 6D**, we can get similar results from the performance of RMSE and MAE. Similarly, from **Table 6**, we can find that the RMSE of the RFR method is the smallest with a value of 0.1940, and at the same time, the MAE of RFR is also the smallest, with a value of 0.1591. Therefore, for the four machine learning methods, the RFR method has the best regression effect. Then, the second is the SVR method. Correspondingly, the prediction effects of ELMR and BPR on the training set are relatively trivial.

Test Set

Similarly, the results of the four machine learning regression methods on the test set, and the corresponding R^2, RMSE and MAE are shown in **Figure 7** and **Table 7**.

Figure 7D is a histogram of model evaluation indexes (R^2, RMSE, and MAE) of four machine learning methods on the test set. The colored columns correspond to the four machine learning methods of BPR, ELMR, SVR, and RFR, respectively. As shown in **Figure 7D**, random forest regression (RFR) has the highest R^2 for the four machine learning methods on the test set. In addition, from **Figure 7B** and **Table 7**, we can observe that the R^2 of RFR is 0.8351. Then, the second one is SVR, the R^2 is 0.6864. The third and fourth are ELMR and BPR, their performance is relatively poor, and the corresponding R^2 are 0.6207 and 0.5703, respectively. Besides, from **Figure 7D** and **Table 7**, on the test set, the smallest root mean square error (RMSE) is RFR, followed by SVR and other methods. Similar results appear on the mean absolute error (MAE). Consequently, our results show that not all machine learning algorithms can show high accuracy. The best performance on the test set is RFR, followed by SVR. Then, the third and fourth are ELMR and BPR. This result is consistent with the training set.

In summary, according to the results of the training set and the test set, for the flood resistance of poplar, the best prediction effect of the four machine learning methods is random forest regression (RFR), and the second one is support vector regression (SVR). By contrast, the performance effects of BP neural network regression (BPR) and Extreme learning machine regression (ELMR) methods are poor. The prediction accuracy from high to low is RFR > SVR > ELMR > BPR. Hence, when predicting the flood resistance of poplar, random forest regression (RFR) and support vector regression (SVR) can be used first, and RFR can be given more consideration.

TABLE 5 | Main information of poplar varieties.

Samples	Z score	V type
Canadensis	−0.31136376	A
DD102-4	−0.659417544	A
Flevo	−0.149084082	A
I-214	−1.018704692	A
I-63	0.244020869	A
LH04-17	0.348729466	A
Ningshanica	−0.843303666	A
Danhong	0.714574083	B
Juba	0.528127585	B
LH04-13	0.845992953	B
I-69	0.356889463	C
I-72	−0.12622992	C
LS68	−0.662543236	C
LS81	0.717527146	C
Lushan	0.282405203	C
NL895	−0.227444975	C
I-45-51	−0.702652584	A
Triplo	−0.561219766	A
2L2025	0.570264018	B
Raspalje	−0.103562886	C

DISCUSSION

Machine learning is a field of artificial intelligence (AI). Compared with traditional statistical models, machine learning has higher performance, and at the same time, its complexity is relatively lower (Mekanik et al., 2013). In fact, before establishing the regression model of machine learning, we performed multiple linear regression (MLR) on the five variables (Fv, qN_Fo, Fm, H_2OS/H_2OR, and RH_S/RH_R) selected by the significance

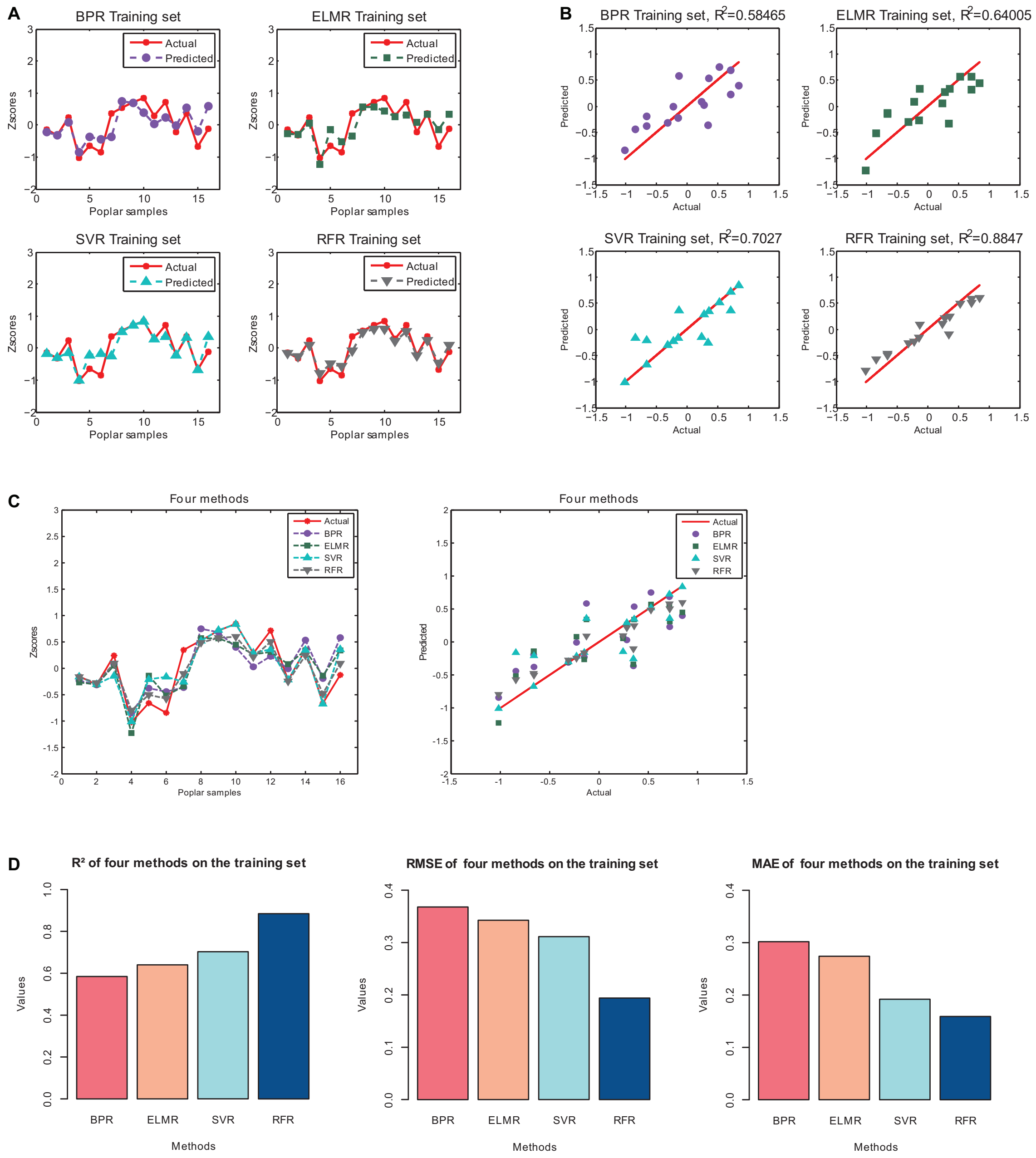


FIGURE 6 | The regression results of the four machine learning methods on the training set. **(A)** Results of Predicted. **(B)** Predicted and actual value. **(C)** Comparison of results of four methods. **(D)** The R^2, RMSE and MAE of the four machine learning methods on the training set.

TABLE 6 | The R^2, RMSE and MAE of the training set.

Methods	BPR	ELMR	SVR	RFR
R^2	0.5847	0.6401	0.7027	0.8847
RMSE	0.3680	0.3426	0.3113	0.1940
MAE	0.3019	0.2741	0.1920	0.1591

testing and correlation analysis. However, the results show that the coefficients of determination (R^2) of MLR on the training set and test set are 0.5616 and 0.5172, respectively. The regression results are shown in **Appendix Figure A1**. Many studies have compared machine learning models with traditional statistical models (Aertsen et al., 2010; Rezaeianzadeh et al., 2014; Idowu et al., 2016; Johnson et al., 2016; Wang and Srinivasan, 2017). In most cases, machine learning models are better than traditional statistical models, such as linear regression. The model and the variables are not linearly related in most situations, and the variables involved are also multivariate. Therefore, more and more fields have begun to use machine learning algorithms. Even

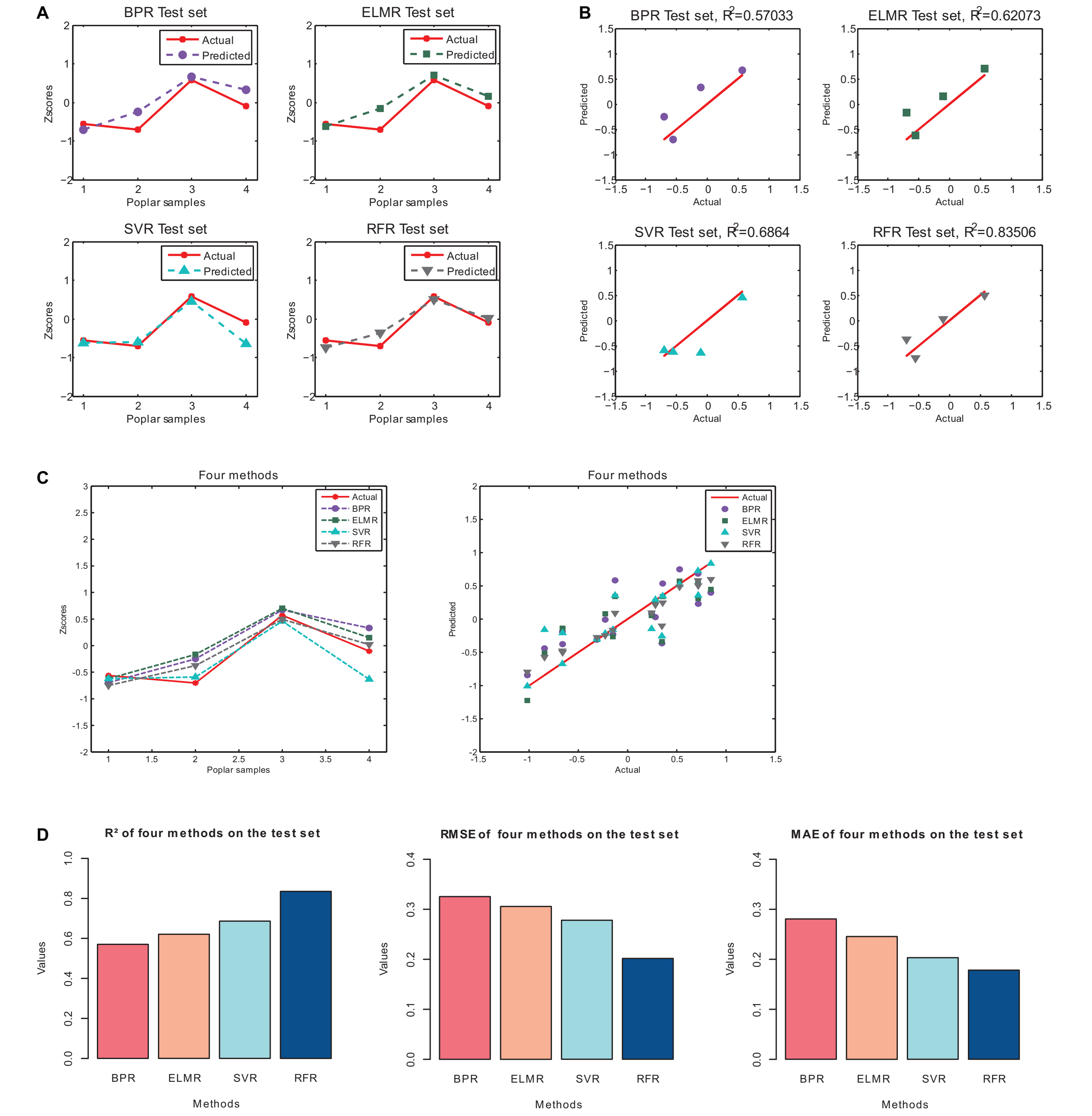


FIGURE 7 | The regression results of the four machine learning methods on the test set. **(A)** Results of Predicted. **(B)** Predicted and actual value. **(C)** Comparison of results of four methods. **(D)** The R^2, RMSE and MAE of the four machine learning methods on the test set.

so, while machine learning has many advantages, it also has limitations. For example, many machine learning models lack interpretability and are prone to overfitting. For this reason, these problems still need to be considered in practical applications.

The risk of resisting flood disasters can be mainly divided into two aspects. One is to directly predict flood disasters in the risk areas, and take preventive measures before the disaster occurs, such as transferring personnel and valuable finances. Generally speaking, the key variables that need to be considered in flood forecasting include 25 factors such as water level, river flood, soil moisture, and rainfall (Maier et al., 2010). Among these key variables affecting flood forecasting, rainfall and the spatial examination of the hydrologic cycle have the most significant effects (Nourani and Komasi, 2013). Although many studies

TABLE 7 | The R^2, RMSE and MAE of the test set.

Methods	BPR	ELMR	SVR	RFR
R^2	0.5703	0.6207	0.6864	0.8351
RMSE	0.3254	0.3057	0.2780	0.2016
MAE	0.2806	0.2456	0.2032	0.1782

have predicted the risk of flooded areas (Sampson et al., 2015; Tehrany et al., 2015; Wang et al., 2015; Darabi et al., 2019), this method still cannot essentially eliminate the impact of flood disasters. At the same time, it is relatively difficult to predict flood disasters. Thus, people have to consider another method to resist flood disasters. Another way to resist the impact of flood disasters is mainly through building dams and afforestation. The key to afforestation is to understand the waterlogging resistance mechanism of plants. The research on the waterlogging resistance mechanism of plants mainly focuses on exploring the ways for plants to resist flood stress (Wang et al., 2013, 2021; Najeeb et al., 2015; Duy et al., 2016). These studies have analyzed the waterlogging resistance mechanism from the molecular level of proteins and metal ions. However, there are few studies on waterlogging resistance prediction, and a complete system is still lacking. Xie and Shen (2021) used the SVR method in machine learning to predict the waterlogging resistance of poplar. However, there are still some limitations in their studies, such as chlorophyll fluorescence features that have not been considered. Compared with the previous research, we considered more accurate feature parameters and more kinds of machine learning methods. Additionally, we improved the prediction system of poplar resistance to waterlogging and added two quantitative definitions of waterlogging resistance evaluation indexes, which has made considerable improvements.

This study used machine learning methods to predict the flood resistance of poplar. First, three indicators of flood resistance were defined and evaluated. Then, the data was processed, and feature selection and modeling evaluation were implemented. The whole process is intuitive and specific, which has perfected the research system of waterlogging tolerance prediction to a considerable extent, and at the same time, it has also promoted the research on the mechanism of waterlogging tolerance. This study helps researchers to screen out poplar varieties with strong waterlogging tolerance during the poplar sapling period. It can further cultivate high-quality poplar saplings to achieve the purpose of precise flood resistance. The results of the experiment show that the machine learning algorithm shows high accuracy in predicting the flood resistance of poplar, especially the random forest regression (RFR) and support vector regression (SVR) methods. The final result has certain practical value. In practical applications, these two algorithms can be used first. However, it must be mentioned that although 160 poplar samples were used throughout the experiment, only 20 poplar varieties were actually used for regression analysis. In addition, in the regression analysis, 80 poplar samples from the experimental group were used and averaged according to varieties. Since the waterlogging tolerance of different individuals may be quite different, the final result may deviate from the actual situation. But within the allowable range of error, our research mainly provides a way of predicting waterlogging tolerance and improving the system for predicting waterlogging tolerance. Future research can consider more poplar varieties to improve the universality and stability of the method. In addition, the quantitative relationship of poplar varieties' impact on flood disasters can be considered. In a word, this research has great theoretical value and practical significance, and the proposed method can meet the actual engineering needs in a considerable range.

CONCLUSION

To predict the flood resistance of poplar, the author first analyzed the differences between the three evaluation indexes of flood resistance. Then, the final evaluation index of waterlogging tolerance was determined, and outliers were eliminated. For the selection of feature parameters, the first screening was carried out according to the significance test and correlation analysis, and then the three methods of hierarchical clustering, Lasso, and stepwise regression were adopted to screen the features for the second time. The selected features are interpretable and promote the understanding of poplar's waterlogging resistance mechanism. Finally, four machine learning methods were used to predict and evaluate the flood resistance of poplar. The results show that the random forest regression and support vector regression methods are more precise. Nevertheless, it must be pointed out that there are only four groups of experiments and controls for each variety. Due to sample differences and randomness, the final result may deviate from the actual situation. Future research can consider more poplar species and sample sizes to improve the versatility and stability of the method.

This research has perfected the prediction system of plant resistance to waterlogging, and has important value for accurate flood resistance and scientific seedling selection. Meanwhile, it has also made a great contribution to a better understanding of the mechanism of waterlogging tolerance. The analysis process of this paper is clear and repeatable. When considering the features related to the flood resistance of poplar, the photosynthesis features, chlorophyll fluorescence features, and environmental features are comprehensively considered. After data processing, feature selection, and other operations, the machine learning models were used to predict the flood resistance of poplar. Finally, the regression results show that the random forest regression (RFR) and support vector regression (SVR) methods have high accuracy. On the test set, the coefficients of determination (R^2) of the two methods are 0.8351 and 0.6864, respectively, the root mean square errors (RMSE) are 0.2016 and 0.2780, and the mean absolute errors (MAE) are 0.1782 and 0.2031. Based on the above conclusions, our research shows that combining photosynthesis, chlorophyll fluorescence, and environmental variables before flooding experiments, modeling and prediction of machine learning methods against waterlogging can achieve high accuracy, which is suitable for actual engineering problems.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

XX, XZ, JS, and KD participated in the conception and design of this research and revised the manuscript. XZ and KD carried out experiments and organized the database. XX and JS performed the statistical analysis and proposed the methodology. XX implemented visualization and wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the National Natural Science Foundation of China (Grant No. 31570665), Fundamental Research Funds for the Central Universities (Grant No. 2662020YLPY017), and the High-end Foreign Experts Recruitment Plan of China (Grant No. G20200017074).

ACKNOWLEDGMENTS

We would like to thank the support of College of Horticulture and Forestry Sciences of Huazhong Agricultural University and other researchers who participated in this experiment.

REFERENCES

Aertsen, W., Kint, V., van Orshoven, J., Ozkan, K., and Muys, B. (2010). Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. *Ecol. Model.* 221, 1119–1130. doi: 10.1016/j.ecolmodel.2010.01.007

Aggarwal, C. C., and Yu, P. S. (2005). An effective and efficient algorithm for high-dimensional outlier detection. *VLDB J.* 14, 211–221. doi: 10.1007/s00778-004-0125-5

Alfieri, L., Bisselink, B., Dottori, F., Naumann, G., de Roo, A., Salamon, P., et al. (2017). Global projections of river flood risk in a warmer world. *Earths Future* 5, 171–182. doi: 10.1002/2016ef000485

Angulo, A. P., and Shin, K. (2019). Mrmr plus and Cfs plus feature selection algorithms for high-dimensional data. *Appl. Intell.* 49, 1954–1967. doi: 10.1007/s10489-018-1381-1

Ao, Y., Zhou, X., Ji, F., Wang, Y., Yang, L., Wang, Q., et al. (2020). Flood disaster preparedness: experience and attitude of rural residents in Sichuan, China. *Nat. Hazards* 104, 2591–2618. doi: 10.1007/s11069-020-04286-0

Arbona, V., Hossain, Z., Lopez-Climent, M. F., Perez-Clemente, R. M., and Gomez-Cadenas, A. (2008). Antioxidant enzymatic activity is linked to waterlogging stress tolerance in citrus. *Physiol. Plant.* 132, 452–466. doi: 10.1111/j.1399-3054.2007.01029.x

Arora, S., and Anand, P. (2019). Binary butterfly optimization approaches for feature selection. *Expert Syst. Appl.* 116, 147–160. doi: 10.1016/j.eswa.2018.08.051

Berhongaray, G., Janssens, I. A., King, J. S., and Ceulemans, R. (2013). Fine root biomass and turnover of two fast-growing poplar genotypes in a short-rotation coppice culture. *Plant Soil* 373, 269–283. doi: 10.1007/s11104-013-1778-x

Bloeschl, G., Hall, J., Viglione, A., Perdigão, R. A. P., Parajka, J., Merz, B., et al. (2019). Changing climate both increases and decreases European river floods. *Nature* 573, 108–111. doi: 10.1038/s41586-019-1495-6

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/a:1010933404324

Chen, W., Xie, X., Peng, J., Shahabi, H., Hong, H., Tien, B. D., et al. (2018). GIS-based landslide susceptibility evaluation using a novel hybrid integration approach of bivariate statistical based random forest method. *Catena* 164, 135–149. doi: 10.1016/j.catena.2018.01.012

Choubin, B., Moradi, E., Golshan, M., Adamowski, J., Sajedi-Hosseini, F., and Mosavi, A. (2019). An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. *Sci. Total Environ.* 651, 2087–2096. doi: 10.1016/j.scitotenv.2018.10.064

Coleman, M. D., Dickson, R. E., and Isebrands, J. G. (2000). Contrasting fine-root production, survival and soil CO_2 efflux in pine and poplar plantations. *Plant Soil* 225, 129–139. doi: 10.1023/a:1026564228951

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/bf00994018

Cui, Z., and Gong, G. (2018). The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. *Neuroimage* 178, 622–637. doi: 10.1016/j.neuroimage.2018.06.001

Darabi, H., Choubin, B., Rahmati, O., Haghighi, A. T., Pradhan, B., and Klove, B. (2019). Urban flood risk mapping using the GARP and QUEST models: a comparative study of machine learning techniques. *J. Hydrol.* 569, 142–154. doi: 10.1016/j.jhydrol.2018.12.002

Demir, B., and Bruzzone, L. (2014). A multiple criteria active learning method for support vector regression. *Pattern Recognit.* 47, 2558–2567. doi: 10.1016/j.patcog.2014.02.001

Ding, S., Zhao, H., Zhang, Y., Xu, X., and Nie, R. (2015). Extreme learning machine: algorithm, theory and applications. *Artif. Intell. Rev.* 44, 103–115. doi: 10.1007/s10462-013-9405-z

Domingues, R., Filippone, M., Michiardi, P., and Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: experiments and analyses. *Pattern Recognit.* 74, 406–421. doi: 10.1016/j.patcog.2017.09.037

Dou, J., Yunus, A. P., Bui, D. T., Merghadi, A., Sahana, M., Zhu, Z., et al. (2019). Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. *Sci. Total Environ.* 662, 332–346. doi: 10.1016/j.scitotenv.2019.01.221

Du, H., Zhu, J., Su, H., Huang, M., Wang, H., Ding, S., et al. (2017). Bulked segregant RNA-seq reveals differential expression and SNPs of candidate genes associated with waterlogging tolerance in maize. *Front. Plant Sci.* 8:1022. doi: 10.3389/fpls.2017.01022

Du, K., Xu, L., Wu, H., Tu, B., and Zheng, B. (2012). Ecophysiological and morphological adaption to soil flooding of two poplar clones differing in flood-tolerance. *Flora* 207, 96–106. doi: 10.1016/j.flora.2011.11.002

Du, P., Samat, A., Waske, B., Liu, S., and Li, Z. (2015). Random forest and rotation forest for fully polarized SAR image classification using polarimetric and spatial features. *ISPRS J. Photogramm. Remote Sens.* 105, 38–53. doi: 10.1016/j.isprsjprs.2015.03.002

Duy, N., Rieu, I., Mariani, C., and van Dam, N. M. (2016). How plants handle multiple stresses: hormonal interactions underlying responses to abiotic stress and insect herbivory. *Plant Mol. Biol.* 91, 727–740. doi: 10.1007/s11103-016-0481-8

Forster, E. J., Healey, J. R., Dymond, C., and Styles, D. (2021). Commercial afforestation can deliver effective climate change mitigation under multiple decarbonisation pathways. *Nat. Commun.* 12:3831. doi: 10.1038/s41467-021-24084-x

Gerjets, R., Richter, F., Jansen, M., and Carminati, A. (2021). Hydraulic redistribution by hybrid poplars (*Populus nigra* x *Populus maximowiczii*) in a greenhouse soil column experiment. *Plant Soil* 463, 145–154. doi: 10.1007/s11104-021-04894-0

Granato, D., Santos, J. S., Escher, G. B., Ferreira, B. L., and Maggio, R. M. (2018). Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: a critical perspective. *Trends Food Sci. Technol.* 72, 83–90. doi: 10.1016/j.tifs.2017.12.006

Guidolin, M., and Pedio, M. (2021). Forecasting commodity futures returns with stepwise regressions: do commodity-specific factors help? *Ann. Oper. Res.* 299, 1317–1356. doi: 10.1007/s10479-020-03515-w

Hallegatte, S., Green, C., Nicholls, R. J., and Corfee-Morlot, J. (2013). Future flood losses in major coastal cities. *Nat. Clim. Change* 3, 802–806. doi: 10.1038/nclimate1979

Han, L., Yang, G., Dai, H., Xu, B., Yang, H., Feng, H., et al. (2019). Modeling maize above-ground biomass based on machine learning approaches using UAV remote-sensing data. *Plant Methods* 15:10. doi: 10.1186/s13007-019-0394-z

Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., et al. (2013). Global flood risk under climate change. *Nat. Clim. Change* 3, 816–821. doi: 10.1038/nclimate1911

Hong, S., Piao, S., Chen, A., Liu, Y., Liu, L., Peng, S., et al. (2018). Afforestation neutralizes soil pH. *Nat. Commun.* 9:520. doi: 10.1038/s41467-018-02970-1

Hong, S., Yin, G., Piao, S., Dybzinski, R., Cong, N., Li, X., et al. (2020). Divergent responses of soil organic carbon to afforestation. *Nat. Sustain.* 3, 694–700. doi: 10.1038/s41893-020-0557-y

Hu, P., Zhang, Q., Shi, P., Chen, B., and Fang, J. (2018). Flood-induced mortality across the globe: spatiotemporal pattern and influencing factors. *Sci. Total Environ.* 643, 171–182. doi: 10.1016/j.scitotenv.2018.06.197

Huang, G.-B., Wang, D. H., and Lan, Y. (2011). Extreme learning machines: a survey. *Int. J. Mach. Learn. Cybern.* 2, 107–122. doi: 10.1007/s13042-011-0019-y

Idowu, S., Saguna, S., Ahlund, C., and Schelen, O. (2016). Applied machine learning: forecasting heat load in district heating system. *Energy Build.* 133, 478–488. doi: 10.1016/j.enbuild.2016.09.068

Johnson, M. D., Hsieh, W. W., Cannon, A. J., Davidson, A., and Bedard, F. (2016). Crop yield forecasting on the Canadian prairies by remotely sensed vegetation indices and machine learning methods. *Agric. For. Meteorol.* 218, 74–84. doi: 10.1016/j.agrformet.2015.11.003

Khosravi, A., Koury, R. N. N., Machado, L., and Pabon, J. J. G. (2018). Prediction of wind speed and wind direction using artificial neural network, support vector regression and adaptive neuro-fuzzy inference system. *Sustain. Energy Technol. Assess.* 25, 146–160. doi: 10.1016/j.seta.2018.01.001

Khosravi, K., Shahabi, H., ThaiPham, B., Adamowski, J., Shirzadi, A., Pradhan, B., et al. (2019). A comparative assessment of flood susceptibility modeling using Multi-Criteria Decision-Making analysis and machine learning methods. *J. Hydrol.* 573, 311–323. doi: 10.1016/j.jhydrol.2019.03.073

Kreuzwieser, J., Hauberg, J., Howell, K. A., Carroll, A., Rennenberg, H., Millar, A. H., et al. (2009). Differential response of gray poplar leaves and roots underpins stress adaptation during hypoxia. *Plant Physiol.* 149, 461–473. doi: 10.1104/pp.108.125989

Lee, E.-J., Kim, K. Y., Zhang, J., Yamaoka, Y., Gao, P., Kim, H., et al. (2021). *Arabidopsis* seedling establishment under waterlogging requires ABCG5-mediated formation of a dense cuticle layer. *New Phytol.* 229, 156–172. doi: 10.1111/nph.16816

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., et al. (2018). Feature selection: a data perspective. *ACM Comput. Surv.* 50:39. doi: 10.1145/3136625

Liu, X., Yang, T., Wang, Q., Huang, F., and Li, L. (2018). Dynamics of soil carbon and nitrogen stocks after afforestation in arid and semi-arid regions: a meta-analysis. *Sci. Total Environ.* 618, 1658–1664. doi: 10.1016/j.scitotenv.2017.10.009

Liu, Z. T., Wu, M., Cao, W. H., Mao, J., Xu, J. P., and Tan, G. Z. (2018). Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing* 273, 271–280. doi: 10.1016/j.neucom.2017.07.050

Loreti, E., van Veen, H., and Perata, P. (2016). Plant responses to flooding stress. *Curr. Opin. Plant Biol.* 33, 64–71. doi: 10.1016/j.pbi.2016.06.005

Lukic, N., Kukavica, B., Davidovic-Plavsic, B., Hasanagic, D., and Walter, J. (2020). Plant stress memory is linked to high levels of anti-oxidative enzymes over several weeks. *Environ. Exp. Bot.* 178:104166. doi: 10.1016/j.envexpbot.2020.104166

Maier, H. R., Jain, A., Dandy, G. C., and Sudheer, K. P. (2010). Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. *Environ. Model. Softw.* 25, 891–909. doi: 10.1016/j.envsoft.2010.02.003

Major, I. T., and Constabel, C. P. (2007). Shoot-root defense signaling and activation of root defense by leaf damage in poplar. *Can. J. Bot.* 85, 1171–1181. doi: 10.1139/b07-090

Mekanik, F., Imteaz, M. A., Gato-Trinidad, S., and Elmahdi, A. (2013). Multiple regression and artificial neural network for long-term rainfall forecasting using large scale climate modes. *J. Hydrol.* 503, 11–21. doi: 10.1016/j.jhydrol.2013.08.035

Miche, Y., Sorjamaa, A., Bas, P., Simula, O., Jutten, C., and Lendasse, A. (2010). OP-ELM: optimally pruned extreme learning machine. *IEEE Trans. Neural Netw.* 21, 158–162. doi: 10.1109/tnn.2009.2036259

Mishra, S., and Padhy, S. (2019). An efficient portfolio construction model using stock price predicted by support vector regression. *N. Am. J. Econ. Finance* 50:101027. doi: 10.1016/j.najef.2019.101027

Moazenzadeh, R., Mohammadi, B., Shahaboddin, S., and Chau, K.-W. (2018). Coupling a firefly algorithm with support vector regression to predict evaporation in northern Iran. *Eng. Appl. Comput. Fluid Mech.* 12, 584–597. doi: 10.1080/19942060.2018.1482476

Moghadassi, A. R., Nikkholgh, M. R., Parvizian, F., and Hosseini, S. M. (2010). Estimation of thermophysical properties of dimethyl ether as a commercial refrigerant based on artificial neural networks. *Expert Syst. Appl.* 37, 7755–7761. doi: 10.1016/j.eswa.2010.04.065

Murtagh, F., and Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 2, 86–97. doi: 10.1002/widm.53

Najeeb, U., Atwell, B. J., Bange, M. P., and Tan, D. K. Y. (2015). Aminoethoxyvinylglycine (AVG) ameliorates waterlogging-induced damage in cotton by inhibiting ethylene synthesis and sustaining photosynthetic capacity. *Plant Growth Regul.* 76, 83–98. doi: 10.1007/s10725-015-0037-y

Nourani, V., and Komasi, M. (2013). A geomorphology-based ANFIS model for multi-station modeling of rainfall-runoff process. *J. Hydrol.* 490, 41–55. doi: 10.1016/j.jhydrol.2013.03.024

Ou, C., Ray, R., Li, C., and Yong, H. (2016). Multi-index and two-level evaluation of shale gas reserve quality. *J. Nat. Gas Sci. Eng.* 35, 1139–1145. doi: 10.1016/j.jngse.2016.09.056

Paprotny, D., Sebastian, A., Morales-Napoles, O., and Jonkman, S. N. (2018). Trends in flood losses in Europe over the past 150 years. *Nat. Commun.* 9:1985. doi: 10.1038/s41467-018-04253-1

Peng, Y., Zhou, Z., Zhang, Z., Yu, X., Zhang, X., and Du, K. (2018). Molecular and physiological responses in roots of two full-sib poplars uncover mechanisms that contribute to differences in partial submergence tolerance. *Sci. Rep.* 8:12829. doi: 10.1038/s41598-018-30821-y

Probst, P., Wright, M., and Boulesteix, A. L. (2018). Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 9:e1301. doi: 10.1002/widm.1301

Rezaeianzadeh, M., Tabari, H., Yazdi, A. A., Isik, S., and Kalin, L. (2014). Flood flow forecasting using ANN, ANFIS and regression models. *Neural Comput. Appl.* 25, 25–37. doi: 10.1007/s00521-013-1443-6

Sampson, C. C., Smith, A. M., Bates, P. B., Neal, J. C., Alfieri, L., and Freer, J. E. (2015). A high-resolution global flood hazard model. *Water Resour. Res.* 51, 7358–7381. doi: 10.1002/2015wr016954

Sayed, G. I., Hassanien, A. E., and Azar, A. T. (2019). Feature selection via a novel chaotic crow search algorithm. *Neural Comput. Appl.* 31, 171–188. doi: 10.1007/s00521-017-2988-6

Shafizadeh-Moghadam, H., Valavi, R., Shahabi, H., Chapi, K., and Shirzadi, A. (2018). Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping. *J. Environ. Manage.* 217, 1–11. doi: 10.1016/j.jenvman.2018.03.089

Tehrany, M. S., Pradhan, B., Mansor, S., and Ahmad, N. (2015). Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. *Catena* 125, 91–101. doi: 10.1016/j.catena.2014.10.017

Tian, L., Li, J., Bi, W., Zuo, S., Li, L., Li, W., et al. (2019). Effects of waterlogging stress at different growth stages on the photosynthetic characteristics and grain yield of spring maize (*Zea mays* L.) under field conditions. *Agric. Water Manage.* 218, 250–258. doi: 10.1016/j.agwat.2019.03.054

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* 58, 267–288. doi: 10.1111/j.25176161.1996.tb02080.x

Voesenek, L. A. C. J., and Bailey-Serres, J. (2013). Flooding tolerance: O_2 sensing and survival strategies. *Curr. Opin. Plant Biol.* 16, 647–653. doi: 10.1016/j.pbi.2013.06.008

Wang, M., Zheng, Q., Shen, Q., and Guo, S. (2013). The critical role of potassium in plant stress response. *Int. J. Mol. Sci.* 14, 7370–7390. doi: 10.3390/ijms14047370

Wang, X., He, Y., Zhang, C., Tian, Y.-A., Lei, X., Li, D., et al. (2021). Physiological and transcriptional responses of *Phalaris arundinacea* under waterlogging conditions. *J. Plant Physiol.* 261:153428. doi: 10.1016/j.jplph.2021.153428

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., and Bai, X. (2015). Flood hazard risk assessment model based on random forest. *J. Hydrol.* 527, 1130–1141. doi: 10.1016/j.jhydrol.2015.06.008

Wang, Z., and Srinivasan, R. S. (2017). A review of artificial intelligence based building energy use prediction: contrasting the capabilities of single and ensemble prediction models. *Renew. Sustain. Energy Rev.* 75, 796–808. doi: 10.1016/j.rser.2016.10.079

Willner, S. N., Levermann, A., Zhao, F., and Frieler, K. (2018). Adaptation required to preserve future high-end river flood risk at present levels. *Sci. Adv.* 4:eaao1914. doi: 10.1126/sciadv.aao1914

Wu, J., Xiong, H., and Chen, J. (2009). Towards understanding hierarchical clustering: a data distribution perspective. *Neurocomputing* 72, 2319–2330. doi: 10.1016/j.neucom.2008.12.011

Xie, X., and Shen, J. (2021). Waterlogging resistance evaluation index and photosynthesis characteristics selection: using machine learning methods to judge poplar's waterlogging resistance. *Mathematics* 9:1542. doi: 10.3390/math9131542

Xiong, Q., Cao, C., Shen, T., Zhong, L., He, H., and Chen, X. (2019). Comprehensive metabolomic and proteomic analysis in biochemical metabolic pathways of rice spikes under drought and submergence stress. *Biochim. Biophys. Acta Proteins Proteomics* 1867, 237–247. doi: 10.1016/j.bbapap.2019.01.001

Xu, R., and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Trans. Neural Netw.* 16, 645–678. doi: 10.1109/tnn.2005.845141

Yang, L., and Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: theory and practice. *Neurocomputing* 415, 295–316. doi: 10.1016/j.neucom.2020.07.061

Yang, P. Y., Hui, C. J., Tien, D. J., Snowden, A. W., Derfus, G. E., and Opel, C. F. (2019). Accurate definition of control strategies using cross validated stepwise regression and Monte Carlo simulation. *J. Biotechnol.* 306:100006. doi: 10.1016/j.btecx.2019.100006

Yang, S., Feng, Q., Liang, T., Liu, B., Zhang, W., and Xie, H. (2018). Modeling grassland above-ground biomass based on artificial neural network and remote sensing in the Three-River Headwaters Region. *Remote Sens. Environ.* 204, 448–455. doi: 10.1016/j.rse.2017.10.011

Yao, L., and Ge, Z. (2018). Deep learning of semisupervised process data with hierarchical extreme learning machine and soft sensor application. *IEEE Trans. Ind. Electron.* 65, 1490–1498. doi: 10.1109/tie.2017.2733448

Yaseen, Z. M., Deo, R. C., Hilal, A., Abd, A. M., Bueno, L. C., Salcedo-Sanz, S., et al. (2018). Predicting compressive strength of lightweight foamed concrete using extreme learning machine model. *Adv. Eng. Softw.* 115, 112–125. doi: 10.1016/j.advengsoft.2017.09.004

Ye, H., Song, L., Chen, H., Valliyodan, B., Cheng, P., Ali, L., et al. (2018). A major natural genetic variation associated with root system architecture and plasticity improves waterlogging tolerance and yield in soybean. *Plant Cell Environ.* 41, 2169–2182. doi: 10.1111/pce.13190

Yin, X., Hiraga, S., Hajika, M., Nishimura, M., and Komatsu, S. (2017). Transcriptomic analysis reveals the flooding tolerant mechanism in flooding tolerant line and abscisic acid treated soybean. *Plant Mol. Biol.* 93, 479–496. doi: 10.1007/s11103-016-0576-2

Zeng, N., Yang, Z., Zhang, Z., Hu, L., and Chen, L. (2019). Comparative transcriptome combined with proteome analyses revealed key factors involved in alfalfa (*Medicago sativa*) response to waterlogging stress. *Int. J. Mol. Sci.* 20:1359. doi: 10.3390/ijms20061359

Zhang, D., Lin, J., Peng, Q., Wang, D., Yang, T., Soroosh, S., et al. (2018). Modeling and simulating of reservoir operation using the artificial neural network, support vector regression, deep learning algorithm. *J. Hydrol.* 565, 720–736. doi: 10.1016/j.jhydrol.2018.08.050

Zhao, X., Zhang, Y., Xie, S., Qin, Q., Wu, S., and Luo, B. (2020). Outlier detection based on residual histogram preference for geometric multi-model fitting. *Sensors* 20:3037. doi: 10.3390/s20113037

Zheng, X., Zhou, J., Tan, D.-X., Wang, N., Wang, L., Shan, D., et al. (2017). Melatonin improves waterlogging tolerance of *Malus baccata* (Linn.) Borkh. seedlings by maintaining aerobic respiration, photosynthesis and ROS migration. *Front. Plant Sci.* 8:483. doi: 10.3389/fpls.2017.00483

Zhou, C., Bai, T., Wang, Y., Wu, T., Zhang, X., Xu, X., et al. (2017). Morpholoical and enzymatic responses to waterlogging in three *Prunus* species. *Sci. Hortic.* 221, 62–67. doi: 10.1016/j.scienta.2017.03.054

Zhou, W., Chen, F., Meng, Y., Chandrasekaran, U., Luo, X., Yang, W., et al. (2020). Plant waterlogging/flooding stress responses: from seed germination to maturation. *Plant Physiol. Biochem.* 148, 228–236. doi: 10.1016/j.plaphy.2020.01.020

Zhu, M., Li, F. H., and Shi, Z. S. (2016). Morphological and photosynthetic response of waxy corn inbred line to waterlogging. *Photosynthetica* 54, 636–640. doi: 10.1007/s11099-016-0203-0

Zhuo, Y., Tehrani, A. M., and Brgoch, J. (2018). Predicting the band gaps of inorganic solids by machine learning. *J. Phys. Chem. Lett.* 9, 1668–1673. doi: 10.1021/acs.jpclett.8b00124

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

APPENDIX A

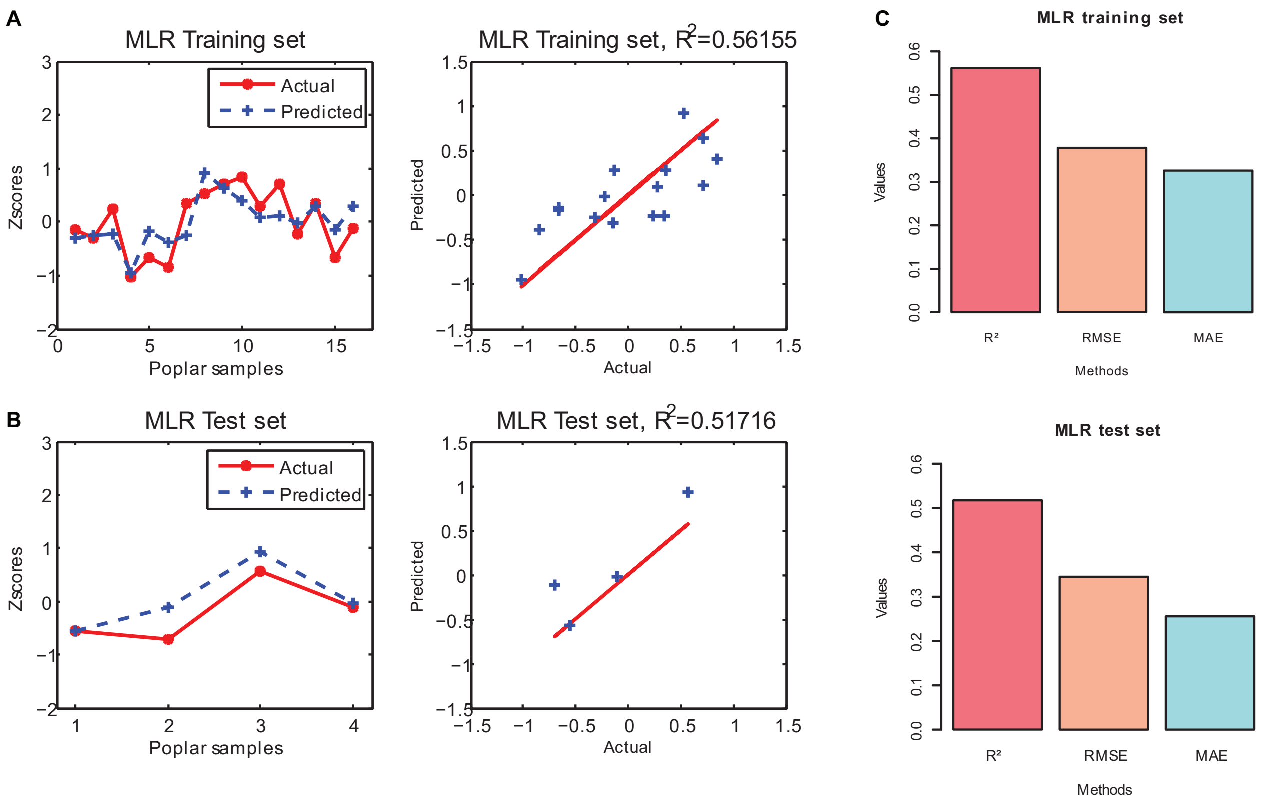


FIGURE A1 | Results of multiple linear regression.

TABLE A1 | The specific meanings and correlation coefficients of 26 features.

Features	Specific meaning	Unit
AHs/Cs	Ball-Berry parameter	Dimensionless
Cond	Conductance to H_2O	mol H_2O m^{-2} s^{-1}
$CndCO_2$	Total conductance to CO_2	mol CO_2 m^{-2} s^{-1}
CO_2S	CO_2 concentration on Sample cell	μmol CO_2 mol^{-1}
CO_2S/CO_2R	CO_2 concentration on Sample cell/CO_2 concentration on Reference cell	Dimensionless
C_2sfc	CO_2 concentration on Leaf Surface	μmol CO_2 mol^{-1}
Ci_Pa	Intercellular CO_2 partial pressure	Pa
Ci/Ca	Intercellular CO_2/Ambient CO_2	Dimensionless
CndTotal	Total conductance to water vapor	mol H2O m^{-2} s^{-1}
CTleaf	Computed leaf temperature	°C
Fo	Minimal fluorescence (dark)	bit
Fm	Maximal fluorescence (dark)	bit
Fv	Variable fluorescence	bit
H_2OS	H_2O concentration on Sample cell	mmol H_2O mol^{-1}
H_2OS/H_2OR	H_2O concentration on Sample cell/H_2O concentration on Reference cell	Dimensionless
H_2O_i	Intercellular H_2O concentration	mmol H_2O mol^{-1}
H_2Odiff	Difference between Intercellular H_2O and Sample cell H_2O	mmol H_2O mol^{-1}
Photo	Photosynthetic rate	μmol CO_2 m^{-2} s^{-1}
PARabs	Absorbed Photosynthetically active radiation	μmol m^{-2} s^{-1}
$PhiCO_2$	Quantum yield corresponding to CO_2 assimilation rate	Dimensionless
qN_Fo	Non-photochemical quenching (Calculated by Fo)	Dimensionless
RH_S	Relative humidity in the sample cell	%
RH_S/RH_R	Relative humidity on Sample cell/Relative humidity on Reference cell	Dimensionless
SVTleaf	Saturated vapor pressure calculated by leaf temperature	Pa
Trans	Transpiration rate	mol H_2O m^{-2} s^{-1}
VpdL	Vapor pressure deficit based on Leaf temperature	kPa