forests

Article

Detection of Moisture Content of *Pinus massoniana* Lamb. Seedling Leaf Based on NIR Spectroscopy with a Multi-Learner Model

Yurong Li †, Haifei Xia †, Ying Liu *, Lintao Huo, Chao Ni and Binli Gou

Jiangsu Co-Innovation Center of Efficient Processing and Utilization of Forest Resources, College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China
* Correspondence: liuying@njfu.edu.cn
† These authors contributed equally to this work.

Abstract: The growth quality of *Pinus massoniana* (Lamb.) seedlings is closely related to the survival rate of afforestation. Moisture content detection is an important indicator in the cultivation of forest seedlings because it can directly reflect the adaptability and growth potential of the seedlings to the soil environment. To improve the accuracy of quantitative analysis of moisture content in *P. massoniana* seedlings using near-infrared spectroscopy, a total of 100 *P. massoniana* seedlings were collected, and their near-infrared diffuse reflectance spectra were measured in the range of 2500 to 800 nm (12,000 to 4000 cm^{-1}). An integrated learning framework was introduced, and a quantitative detection model for moisture content in *P. massoniana* seedlings was established by combining preprocessing and feature wavelength selection methods in chemometrics. Our results showed that the information carried by the spectra after multiple scattering correction (MSC) preprocessing had a good response to the target attribute. The stacking learning model based on the full-band spectrum had a prediction coefficient of determination R^2 of 0.8819, and the prediction accuracy of moisture content in *P. massoniana* seedlings could be significantly improved compared to the single model. After variable selection, the spectrum processed by MSC and feature selection with uninformative variable elimination (UVE) showed good prediction effects in all models. Additionally, the prediction coefficient of determination R^2 of the support vector regression (SVR)—adaptive boosting (AdaBoost)—partial least squares regression (PLSR) + AdaBoost model reached 0.9430. This indicates that the quantitative analysis model of moisture content in *P. massoniana* seedlings established through preprocessing, feature selection, and stacking learning models can achieve high accuracy in predicting moisture content in *P. massoniana* seedlings. This model can provide a feasible technical reference for the precision cultivation of *P. massoniana* seedlings.

Keywords: NIR spectroscopy; *Pinus massoniana* seedlings; non-destructive detection; multi-learner model

Citation: Li, Y.; Xia, H.; Liu, Y.; Huo, L.; Ni, C.; Gou, B. Detection of Moisture Content of *Pinus massoniana* Lamb. Seedling Leaf Based on NIR Spectroscopy with a Multi-Learner Model. *Forests* **2023**, *14*, 883. https://doi.org/10.3390/f14050883

Academic Editor: Laurie Schimleck

Received: 2 March 2023
Revised: 19 April 2023
Accepted: 23 April 2023
Published: 25 April 2023

1. Introduction

Pinus massoniana (Lamb.) is highly adaptable and widely distributed, with a horizontal distribution spanning about 20° longitude and 12° latitude. It is a major afforestation species with high economic value in southern China [1,2]. The use of advanced technical tools for monitoring and quantitatively analyzing dynamic phenotypic changes in seedlings is crucial for achieving rapid and accurate evaluation of seedling vigor. This is particularly important for the cultivation of intensive industrial timber forests and the improvement of precise forestry management in China [3].

The moisture content of seedling leaves is an important indicator of vitality and plays an important role in photosynthesis, material transport, and the maintenance of physiological functions [4]. To improve the survival rate of *P. massoniana* plantations, leaf moisture content should be detected before seedling transplanting. Commonly used methods for moisture content testing include the constant weight method, vacuum drying method, fixed temperature and time drying, and other chemical methods, etc. The mentioned testing methods are more accurate; however, they are cumbersome, have a long cycle time, and result in irreversible damage to seedling samples, leading to wastage of resources and not being conducive to the sustainable development of seedlings [5,6].

Near-infrared (NIR) spectroscopic imaging combines the advantages of spectroscopic and imaging techniques and is widely used as a mature non-destructive detection technique in the fields of agriculture and forestry, pharmaceuticals, food, petrochemicals, and tobacco [7–11]. There are information redundancy, noise, and background factors in the variables of spectral measurements. Extracting effective information from complex spectral data to identify component changes is a hot topic in spectral analysis research. NIR spectra (NIRS) were processed using discrete wavelet transforms (DWT), which decomposed the original spectrum into six layers of denoising. Information variables were then selected by reducing the dimensionality of sub-layer reconstruction spectra through the bootstrap soft shrinkage algorithm (BOSS), leading to the establishment of an effective non-destructive prediction model for tea moisture content [12]. Various preprocessing strategies and feature variable selection methods based on visible and near-infrared (Vis–NIR) technology, such as competitive adaptive reweighted sampling (CARS) and uninformative variable elimination (UVE), have been used to develop regression analysis models such as partial least squares regression (PLSR), support vector regression (SVR), and random forest (RF) to significantly improve the accuracy of predicting soil organic matter (SOM) content [13]. A model was established between leaf moisture content (LWC) and water index (WI) based on NIR reflectance, which provided a non-destructive and immediate measurement method for monitoring the water status of sunflower plants [14]. Moisture content detection and visualization of peanuts were achieved in the 900–1700 nm band, but only the weighted regression coefficient method was used to extract the characteristic wavelength [15]. The above studies show that NIR combined with different preprocessing, data dimension reduction, and feature extraction methods can achieve non-destructive detection of components.

Classical prediction models such as PLSR, SVR, back propagation neural network (BPNN), and RF are commonly used. However, to overcome the limitations of a single prediction model in various application scenarios, researchers have attempted to improve the accuracy and convergence speed of the model through two approaches: optimizing the core parameters of the algorithm or integrating different models. As a result, they have achieved good results [16–19]. The use of a radial basis function network (RBF) based on a self-organizing feature map (SOM) resulted in the successful prediction of key nutrient content in Lanzhou lily. This approach combined the self-organizing clustering features of SOM with the nonlinear approximation ability of RBF, leading to prediction results that were 5.6% higher than the correlation coefficient of prediction (Rp) obtained using the PLSR method [20]. Mixed linear regression and non-linear regression models, namely PLSR + general regression neural network (GRNN) and PLSR + BPNN, were used to predict the plant leaf nitrogen to phosphorus ratio (N/P). Both mixed models showed higher accuracy and stability as compared to PLSR, while also overcoming the overfitting problem of single regression models [21].

The aforementioned models can be classified into conventional regression models (PLSR, SVR), stacking learning models (RF), and neural network models (BPNN, GRNN). When dealing with high-dimensional and nonlinear spectral data, single regression models are no longer suitable due to their poor prediction performance and inability to effectively fit nonlinear or high-dimensional data. Combining multiple models and utilizing the stacking learning method can enable the mapping of data features from multiple perspectives, breaking the singularity of data domain analysis, and integrating the results of multiple learners to achieve more accurate predictions and higher robustness [22]. In this study, we utilized the NIRS data from *P. massoniana* seedlings as the primary source of information and integrated the stacking model to establish a quantitative analysis model for predicting the moisture content of *P. massoniana* seedlings. We compare and analyze our method with the chemometric method of multiple corrections and address the issue of high-dimensional NIRS data by selecting the optimal feature variables. By comparing the prediction results, we choose a suitable system modeling method for predicting the moisture content of *P. massoniana* seedlings. This provides a new method for moisture content detection in the cultivation process of seedlings.

2. Materials and Methods

2.1. Experimental Materials

In this paper, all of the *P. massoniana* samples are obtained from Qingyuan Nursery, Yizhou District, Hechi City, Guangxi Province, China, for a total of 100 plants. These samples are all annual *P. massoniana* seedlings and the height of the seedlings ranges from 15 to 20 cm. Moreover, the selected samples are in good growth condition and are not infected with insects or diseases.

2.1.1. Spectral Data Acquisition

The diffuse reflection method is commonly used for collecting the NIR spectra of solid samples. In this method, NIR light enters the sample and undergoes several reflections, refractions, scatterings, and absorptions inside the sample. This process carries rich information about the sample′s structure and tissue, and the resulting NIR signal is finally captured by the NIR spectrometer, completing the NIR spectral acquisition of the sample [23].

The spectral data of each *P. massoniana* sample were collected by a multi-purpose analyzer Fourier transform (MPA) NIR spectrometer (Figure 1a) equipped with a PbS detector in the wavelength range of 2500 to 800 nm (12,000 to 4000 cm^{-1}), with reflection mode and a spectral resolution of 4 cm^{-1}. The number of spectra per *P. massoniana* sample is 2203. The NIR optical fiber probe was aligned to the different parts of the *P. massoniana* samples to obtain the spectral data (Figure 1b). To avoid experimental errors as much as possible, the high, middle, and low areas of the *P. massoniana* seedling samples were selected for two repeated scans. We then took the average value of the data results from the six scans as the final value of the spectral data.

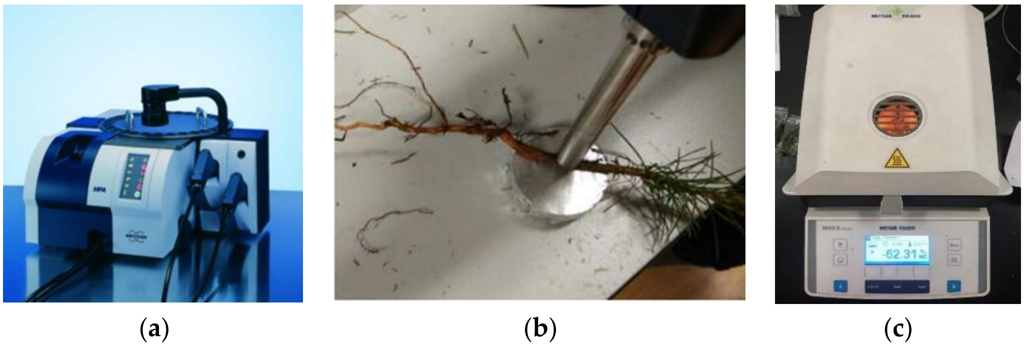

(**a**) (**b**) (**c**)

Figure 1. Equipment and collection process: (**a**) Bruker-MPA NIR spectrometer; (**b**) the process of spectral collection of *P. massoniana* seedlings; (**c**) HB43-S halogen moisture meter.

2.1.2. Moisture Content Determination

The moisture content of the leaves of *P. massoniana* was measured using the HB43-S halogen moisture meter (Figure 1c). This device determined the moisture contained in a sample by measuring the weight loss of the sample after heating and drying [24]. After placing the *P. massoniana* leaf samples into the sampling chamber, the temperature inside the instrument was increased to 125 °C, and the *P. massoniana* samples were heated by a halogen lamp until the sample mass no longer decreased, and then the moisture content of the samples was recorded.

2.2. Preprocessing and Feature Selection

2.2.1. Spectral Preprocessing

The raw spectral data may contain interference noise and irrelevant information generated under the influence of factors such as ambient temperature and humidity in the process of acquisition. Therefore, before model training, preprocessing operations are needed to be performed on the raw NIR spectral data to remove irrelevant noise that can interfere with the prediction results and improve the accuracy of the prediction model.

To reduce the random errors of the samples and filter out some systematic errors, five classical preprocessing methods are proposed in this experiment, including the moving window spectral matrix smoothing algorithm (Nirmaf), L2-normalize, multiple scattering correction (MSC), Savitzky–Golay smoothing (SG smoothing), and standard normal variate (SNV).

The Nirmaf preprocessing method uses a shifted average of individual sample data, thus denoising the data. The L2-normalize preprocessing method scales and pans the data so that they fall into a small, specific interval, which serves to remove the effect of data magnitude and make the data metrics comparable with each other. The L2-parametric normalization operation divides each dimension $(x_1, x_2, \ldots, x_n)$ of the spectral data X by the second parametric number $\|X\|_2$ of the vector X to obtain a new normalized vector, as shown in Equation (1):

$$X_{L2} = \left(\frac{x_1}{\sqrt{{x_1}^2 + {x_2}^2 + \ldots + {x_n}^2}}, \frac{x_2}{\sqrt{{x_1}^2 + {x_2}^2 + \ldots + {x_n}^2}}, \ldots, \frac{x_n}{\sqrt{{x_1}^2 + {x_2}^2 + \ldots + {x_n}^2}} \right) \quad (1)$$

MSC enhances the correlation between spectra and data [25]. SG smoothing has a strong filtering effect on noise points [26]. Additionally, SNV eliminates surface scattering and the effect of light range variation on diffuse reflectance spectra [27]. Considering the diversity of NIR spectral preprocessing results, MSC, SG smoothing, and SNV are also applied in this study. Spectral preprocessing operations are performed on MATLAB R2021b.

2.2.2. Feature Selection

The spectral data of *P. massoniana* seedlings contains a large number of spectral data features, including noise and a significant amount of unrelated information. Variable selection can be used to remove noise and interference variables that are unrelated to the target attribute from the spectral data. The proposed feature selection method not only reduces the number of model variables but also decreases the model′s complexity, thereby improving the predictive performance and robustness of the model. Based on spectral preprocessing, we utilized five conventional methods, namely genetic algorithm (GA), successive projections algorithm (SPA), UVE, CARS, and least angle regression (LARS) [28–32], for variable selection of the spectral data. The goal was to select appropriate spectral variables to be used in the quantitative analysis of *P. massoniana* seedling moisture content. Spectral feature selection is performed on PyCharm 2022.2.

2.3. Model Selection and Optimization

The stacking integrated model is a multi-layer learning model, with the first layer being the base learner layer consisting of different regression models and the second layer being the final output meta-learner layer. The difference in the performance of the models is evident in the spectral data. Nevertheless, the stacking integrated model can further improve the model′s prediction accuracy by combining the algorithmic strengths of each base learner and eliminating their respective prediction errors [33].

There are numerous ways to combine models in a stacking-integrated model. To select the appropriate learners for the stacking model, the prediction performance and correlation of every single model were compared and analyzed. Then, four candidate learners (adaptive boosting (AdaBoost), ExtraTree, PLSR, and SVR) are selected. Finally, the optimal stacking model construction method for this study is selected by comparing different combinations. The hardware and software used in this study were as follows: operating system: Windows 10, CPU: Intel I7-11700F 2.50 GHz, GPU: Nvidia GeForce RTX 3080Ti (12 GB), and environment configuration: PyCharm+Pytorch1.8+Python 3.7.4+Cuda 12.1.

3. Results and Discussion

3.1. Sample Moisture Content Data

The experimental sample dataset in this study was divided into a training set and a test set using a division ratio of 8:2 through the sample set partitioning based on the joint X-Y distance sampling (SPXY) algorithm. This was conducted to ensure that each data set could characterize the sample distribution to the maximum extent, increase the variability and representativeness among the samples, and further improve the stability of the model. The results of leaf moisture content measurements of *P. massoniana* seedlings are shown in Figure 2.

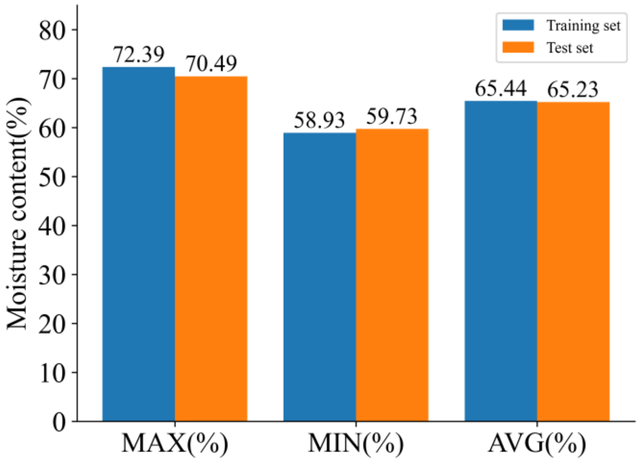


Figure 2. Statistics of the moisture content of *P. massoniana* samples. MAX (%) = the maximum value of moisture content in the training and test sets; MIN (%) = the minimum value of moisture content in the training and test sets; AVG (%) = the average value of moisture content in the training and test sets.

3.2. Raw Spectra and Spectral Preprocessing Results

The spectral data of *P. massoniana* measured by NIRS has certain noise and irrelevant information, so the raw spectral data needs to be preprocessed to eliminate them. The raw spectra and the NIR spectral curves after the five preprocessing methods of Nirmaf, L2-normalize, MSC, SG smoothing, and SNV are shown in Figure 3, in which the horizontal axis is the wavenumber (cm^{-1}) and the vertical axis is the absorbance (%). There are obvious absorbance peaks in the wavenumber 8000~4000 cm^{-1}, indicating that the use of NIR spectra to predict the moisture content of *P. massoniana* seedling samples is feasible [34,35].

As shown in Figure 3, the spectral curve becomes smoother after the preprocessing operation. To ensure the stability of the regression model and the accuracy of the prediction results, it is necessary to perform normalization on the preprocessed data and eliminate magnitude differences with the maximum–minimum normalization method. This is to avoid excessive differences in magnitude between data of different dimensions and improve the accuracy of the model.

PLSR was used to establish a prediction model for the leaf moisture content of *P. massoniana* seedlings. The coefficient of determination R^2 and root mean square error RMSE were used as the evaluation metrics of the prediction model. Furthermore, the number of potential factors of the optimal model was determined through manual tuning. The number of potential factors in the model was chosen in the range of 1 to 20. The prediction results of the spectral data processed by each preprocessing algorithm in the PLSR model are shown in Table 1.

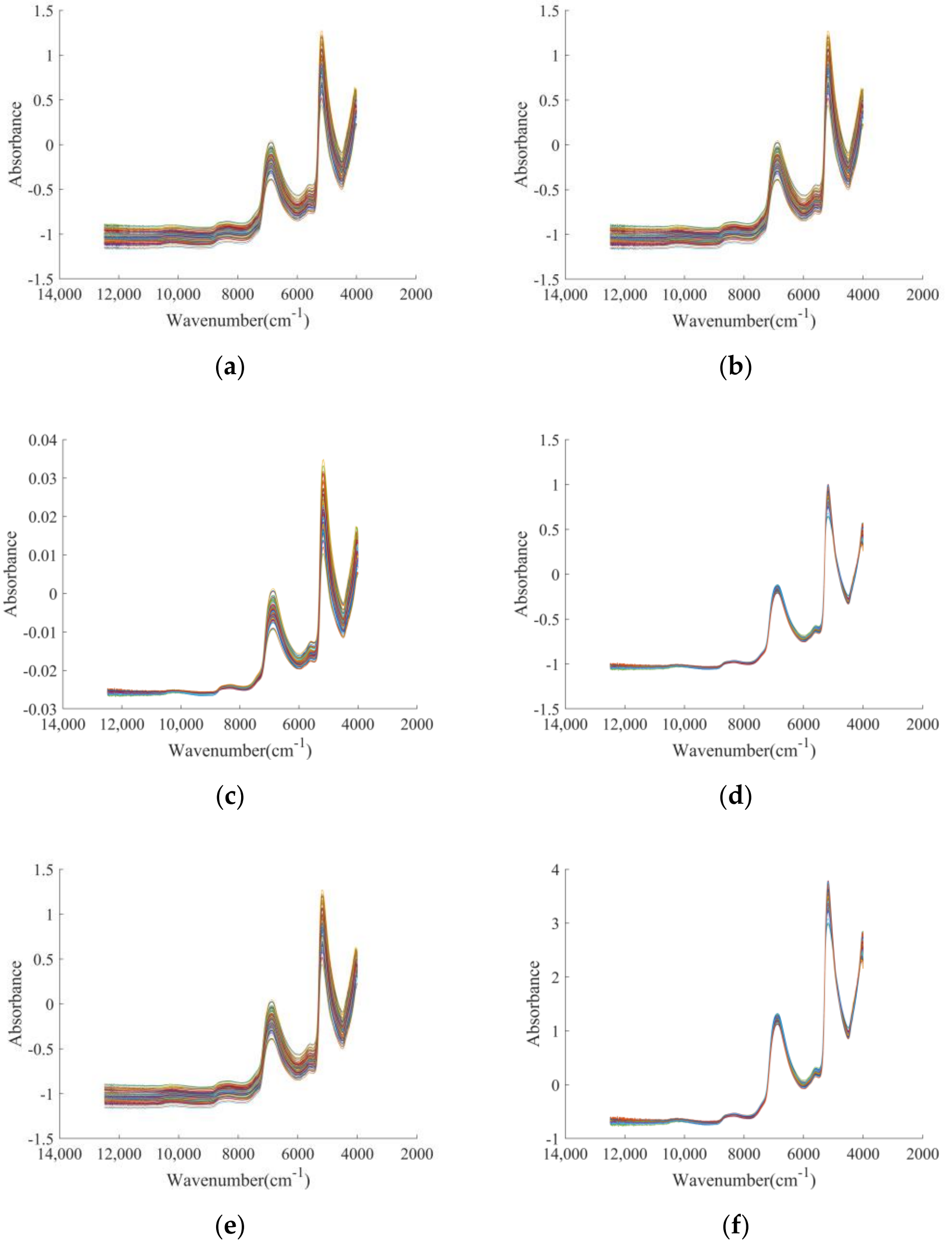


Figure 3. Spectral preprocessing effects: (**a**) raw spectrum; (**b**) Nirmaf−processed spectrum; (**c**) L2−normalize−processed spectrum; (**d**) MSC−processed spectrum; (**e**) SG smoothing−processed spectrum; and (**f**) SNV−processed spectrum.

Table 1. Comparison of PLSR-based preprocessing methods.

Preprocessing Methods	Number of Potential Factors	Training Set		Test Set	
		R^2	RMSE	R^2	RMSE
Raw data	10	0.8377	1.0040	0.7318	1.3384
Nirmaf	11	0.8447	0.9822	0.7513	1.2887
L2-normalize	10	0.9315	0.6523	0.7216	1.3637
MSC	10	0.8848	0.8458	0.8057	1.1391
SG smoothing	10	0.8150	1.0718	0.8009	1.1531
SNV	10	0.8848	0.8457	0.8011	1.1525

The coefficients of determination (R^2) of the raw spectral data and data processed by five preprocessing methods ranged from 0.7216 to 0.8057, and root mean square error (RMSE) ranged from 1.1391 to 1.3637 for the test set in the PLSR regression model. Generally, there is little difference in performance among PLSR models. The R^2 of the training and test sets of the MSC-PLSR are 0.8848 and 0.8057, respectively. The R^2 value of the training set is only 0.0467 lower than that of the L2-normalize-PLSR. Furthermore, the R^2 value of the test set improved by 0.0739 compared to the original spectrum, reaching 0.8057.

In contrast, other preprocessing methods, such as L2-normalize, reduce the original spectral modeling performance. Although this preprocessing has a higher R^2 in the training set than the other methods, it lacks robustness and does not significantly improve the model. Therefore, MSC is subsequently selected as the preprocessing method for the NIR spectra of *P. massoniana* seedlings.

3.3. Composition and Optimization of Stacking Integrated Model

The optimization of the prediction effect of the stacking integrated model depends on the selection of appropriate base learners and meta-learner, as well as the combination of different learners. In this study, the prediction performance of each single model and the Pearson correlation coefficient between each model were used to select the appropriate combination of base learner and meta-learner.

3.3.1. Selection of Base Learners

The component learners initially selected in this paper are AdaBoost, ExtraTree, RF, Ridge, PLSR, and SVR. Among them, AdaBoost and RF were improved with boosting and bagging integrated methods, respectively. Both of these models have a reasonable number of hyperparameters, and the practical application of these models does not require adjusting too many parameters. ExtraTree has a small number of key hyperparameters and reasonable heuristics for configuration parameters that can handle high-dimensional data. Ridge regression has high stability and can effectively improve the problem of model overfitting. PLSR is often used in spectral data, which can better handle data with dimensions much larger than the number of samples. The SVR algorithm is easy to implement and robust to outliers.

In integrated learning, the main challenge is how to synthesize multiple weak learners into one strong learner. In the initial selection of the base learner, two aspects need to be considered. Firstly, the learning strength of the base learner should be taken into account, and models with significant differences can be chosen as base learners to combine the advantages of different algorithms [36]. Secondly, the prediction performance of the base learner directly affects the overall performance of the stacking integrated model [37], so choosing base learners with better prediction performance can improve the integrated model′s performance.

In testing the performance of a single model, the grid search method was used to select the optimal hyperparameters. The experimental results are shown in Table 2.

PLSR, SVR, and Ridge were more effective in predicting the moisture content of *P. massoniana* seedlings; Adaboost had the lowest test set R^2 of 0.6425 with weak performance; and the prediction accuracy of the remaining learners had little difference.

In this paper, the correlation between the prediction results of each model on the test set was measured using the Pearson correlation coefficient to analyze the difference between models. The correlation can be expressed using Equation (2):

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}} \tag{2}$$

where μ_x and μ_y are the predicted mean values of model x and model y, respectively. n denotes the number of test set samples. x_i, y_i denote the predicted values of the i-th sample of models x and y.

Table 2. Model optimal hyperparameters and prediction results.

Model	Optimal Hyperparameters	Training Set		Test Set	
		R^2	RMSE	R^2	RMSE
AdaBoost	n_estimators = 50	0.9125	0.7369	0.6425	1.5452
ExtraTree	min_samples_leaf = 7 max_depth = 5	0.7433	1.2626	0.7362	1.3272
RF	n_estimators = 8 max_leaf_nodes = 14	0.9000	0.7881	0.7336	1.3340
Ridge	alpha = 1.6 solver = sag	0.7939	1.1314	0.8054	1.1400
PLSR	n_components = 10	0.8848	0.8458	0.8057	1.1391
SVR	kernel = linear C = 1.25	0.8107	1.0842	0.7884	1.1888

The Pearson correlation coefficients of each single model are shown in Figure 4.

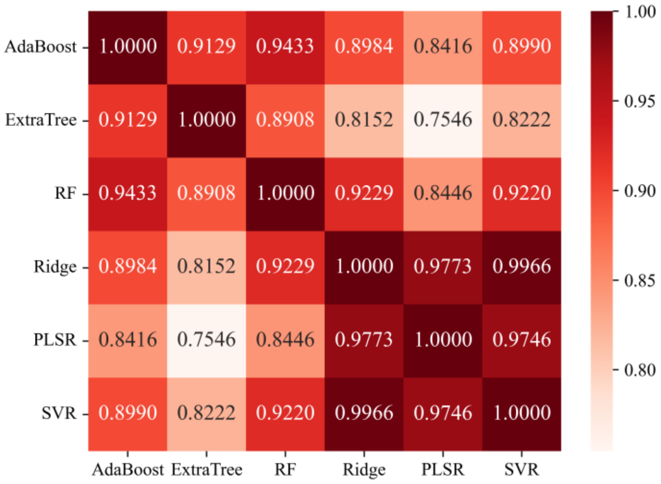


Figure 4. Correlation analysis of the prediction results of each model.

Different model algorithms extract features from different perspectives. Therefore, when constructing stacking models, it is essential to select models with a substantial difference to combine the advantages of various algorithms. As can be seen from Figure 4, the correlation between Ridge and SVR is the highest at 0.9966. This is because both models use L2 regularization terms and have similar hyperparameters, as well as the angles of the observed data being more similar. However, SVR has better robustness, so SVR is chosen as one of the candidate base learners. In contrast, AdaBoost and ExtraTree have large differences compared with other models, and the prediction results are less relevant. When constructing stacking models with large differences, the advantages between different model algorithms can be combined. Therefore, AdaBoost and ExtraTree are chosen as candidate-based learners. The prediction performance of the base learner directly affects the prediction performance of the stacking model. The candidate base learner was selected based on its prediction accuracy, and the PLSR with the highest accuracy was chosen as the base learner for the stacking model.

This paper selects AdaBoost, ExtraTree, PLSR, and SVR as the candidate base learners for the stacking model by combining the model prediction results and the correlation between the models. The mandatory base learners are chosen as the best and worst performers, PLSR and AdaBoost, respectively.

3.3.2. Comparison of Different Model Combinations in Stacking

As the hyperparameter of PLSR did not match the data structure, it was not used as the meta-learner in this paper. Instead, AdaBoost, ExtraTree, and SVR were used as meta-learners for comparison experiments to select the best combination of learners.

As can be observed from Table 3, there is a significant difference in the prediction of moisture content of *P. massoniana* seedlings for different combinations of stacking integrated models. Comparing the combination ways of numbers 1, 2, and 3 and numbers 4, 5, and 6, the stacking integrated model′s overall performance improved when AdaBoost was used as the meta-learner. This is because AdaBoost considers the weight assignment of each base learner, reducing the risk of overfitting and improving the model′s generalization ability. The overall prediction performance of the stacking integrated model was compared between the combinations of numbers 1 and 4, 2 and 5, and 3 and 6. It was found that using SVR, AdaBoost, and PLSR as base learners resulted in better performance than using ExtraTree, AdaBoost, and PLSR as base learners. Additionally, the inclusion of SVR was found to be more beneficial for improving the generalization ability of the model. Based on the above analysis, we can conclude that the combination of number 6 not only exhibits better prediction performance than each individual model, but it also achieves the best prediction performance among all the different combinations.

Table 3. Comparison of the combination method and performance of different learners.

Number	Base Learners	Meta-Learner	Training Set		Test Set	
			R^2	RMSE	R^2	RMSE
1	ExtraTree-AdaBoost-PLSR	SVR	0.9413	0.6039	0.7417	1.3135
2	ExtraTree-AdaBoost-PLSR	ExtraTree	0.7750	1.1819	0.7211	1.3647
3	ExtraTree-AdaBoost-PLSR	AdaBoost	0.9709	0.4253	0.8305	1.0639
4	SVR-AdaBoost-PLSR	SVR	0.9448	0.5856	0.7188	1.3705
5	SVR-AdaBoost-PLSR	ExtraTree	0.7740	1.1846	0.7708	1.2372
6	SVR-AdaBoost-PLSR	AdaBoost	0.9718	0.4187	0.8819	0.8879

Therefore, in this paper, we have chosen the combination of number 6, which includes SVR, AdaBoost, and PLSR as the base learners and AdaBoost as the meta-learner of the stacking integrated model. The stacking model framework and the complete training process are shown in Figure 5.

3.4. Feature Selection and Stacking Prediction Performance Analysis

Due to the large number of spectral bands in the raw spectral data of the leaves, the existence of interference information and redundant bands can result in a computationally intensive and less accurate prediction model. Therefore, it is necessary to perform feature band extraction on the raw spectral data to address this issue. The feature bands were selected by five feature selection algorithms: GA, SPA, UVE, CARS, and LARS (Figure 6).

Overall, SPA outputs more scattered feature bands and has the best dimensionality reduction effect, selecting 23 feature bands and reducing the number of bands by 98.96%; CARS has the second-best dimensionality reduction effect, with 46 feature bands, and a small number of overlapping bands can be seen when comparing CARS and SPA. UVE and LARS retain more features in the feature band selection process, mainly concentrated on 5500–4000 cm^{-1} and 12,500–8000 cm^{-1}, and the band reduction is 74.17% and 77.30%, respectively. GA has the worst dimensionality reduction effect, the features are uniformly scattered and dense, the discrimination of the strong interference information present in the spectral data is low, and the band reduction is 55.97%. The vibration generated by the O-H chemical bonds in plant water molecules is mainly in the 4600–4000 cm^{-1} interval [38], and the wavenumbers in this interval have a certain response relationship with plant moisture content. All five methods retained part of the wavelengths in this interval, which can be further analyzed.

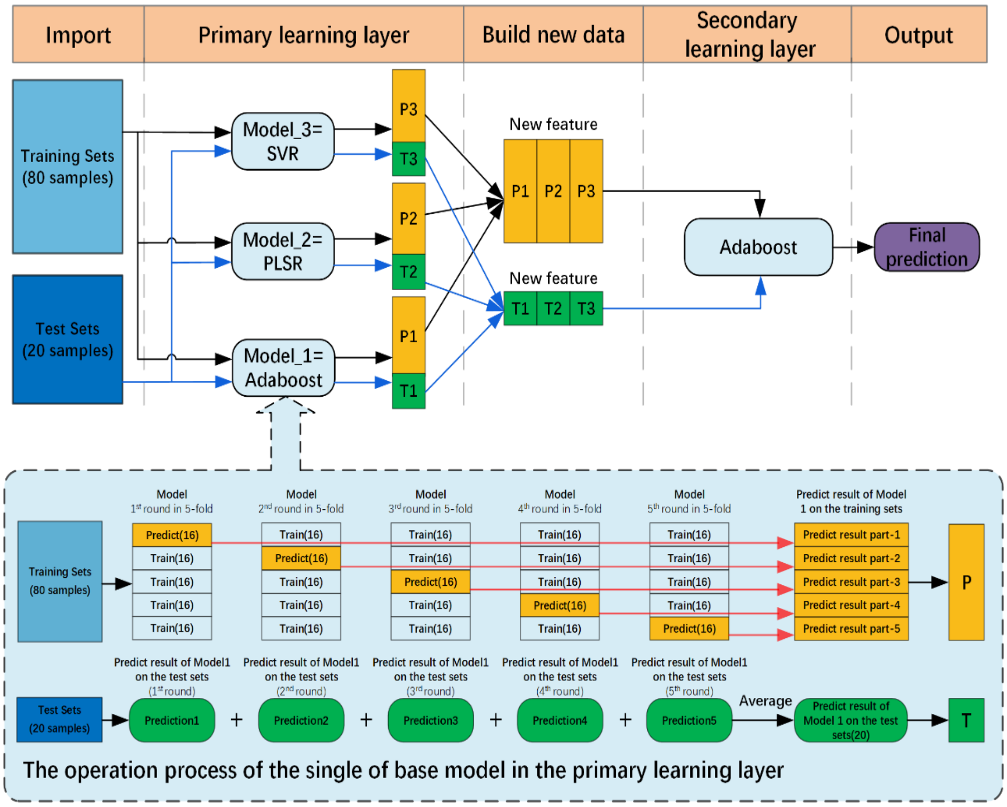

Import
Primary learning layer
Build new data
Secondary learning layer
Output
Training Sets (80 samples)
Test Sets (20 samples)
Model_3= SVR
Model_2= PLSR
Model_1= Adaboost
P3
T3
P2
T2
P1
T1
New feature
P1 P2 P3
New feature
T1 T2 T3
Adaboost
Final prediction
Model 1st round in 5-fold
Model 2nd round in 5-fold
Model 3rd round in 5-fold
Model 4th round in 5-fold
Model 5th round in 5-fold
Predict result of Model 1 on the training sets
Predict(16)
Train(16)
Predict result part-1
Predict result part-2
Predict result part-3
Predict result part-4
Predict result part-5
P
Predict result of Model1 on the test sets (1st round)
Predict result of Model1 on the test sets (2nd round)
Predict result of Model1 on the test sets (3rd round)
Predict result of Model1 on the test sets (4th round)
Predict result of Model1 on the test sets (5th round)
Prediction1
Prediction2
Prediction3
Prediction4
Prediction5
Average
Predict result of Model 1 on the test sets(20)
T
The operation process of the single of base model in the primary learning layer

Figure 5. Moisture content prediction model of *P. massoniana* leaves based on the stacking integrated learning model.

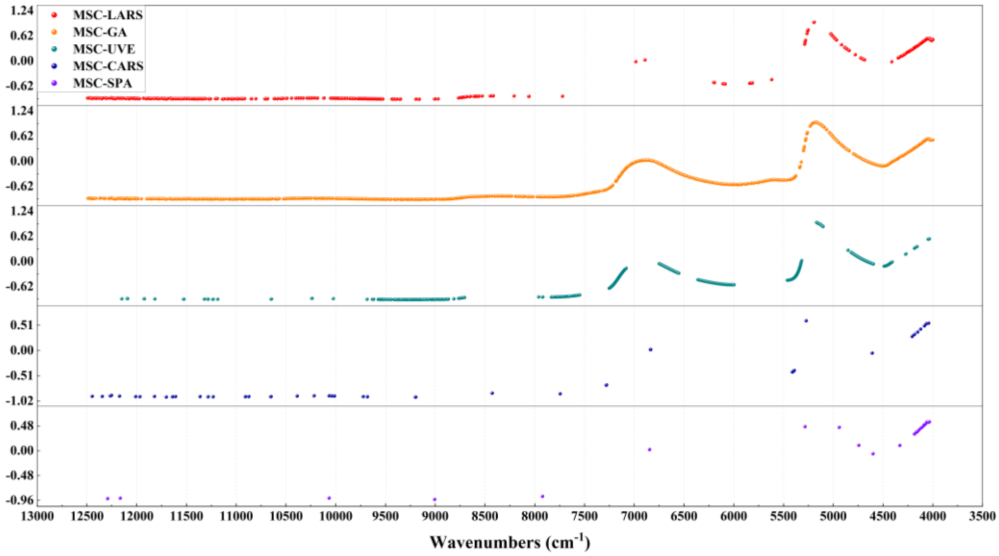


Figure 6. Spectral feature band selection effect. <Preprocessing> − <Feature Selection> = Spectral data after preprocessing and feature selection methods.

The PLSR, SVR, AdaBoost, and stacking integrated models for the moisture content of *P. massoniana* leaves were developed by using full-spectrum data and feature band spectral data, respectively. To ensure the best prediction results for each model, it is necessary to set the hyperparameters of each model. The model tuning process is conducted using the most widely used K-fold cross-validation, with K = 5. The final model hyperparameters are determined using a grid search method in combination with cross-validation. The prediction results of each model are shown in Table 4.

Table 4. Comparison of model results with different characteristic bands.

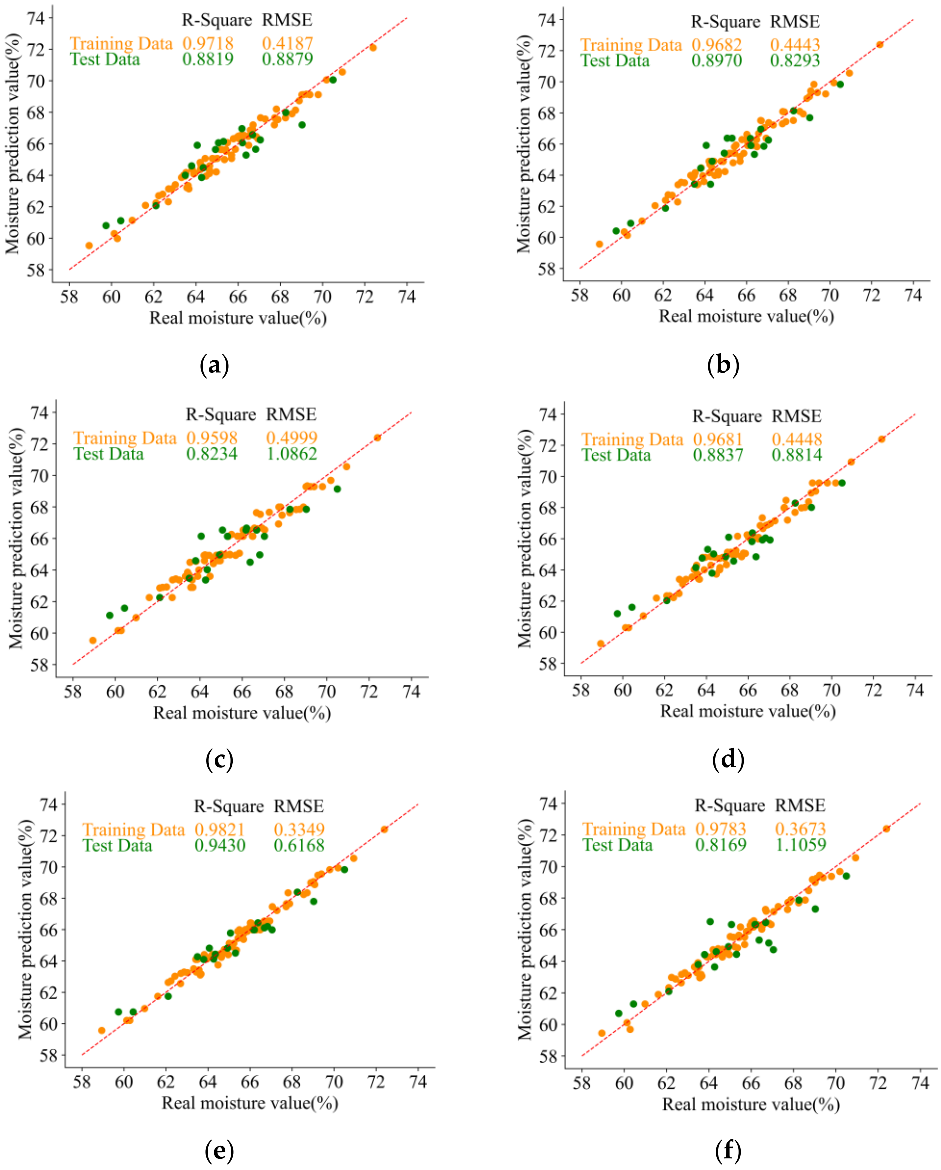
		PLSR				SVR				AdaBoost				Stacking			
Feature Selection	**Number of Wavenumbers**	**Training Set**		**Test Set**		**Training Set**		**Test Set**		**Training Set**		**Test Set**		**Training Set**		**Test Set**	
		R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
MSC	2203	0.8848	0.8458	0.8057	1.1391	0.8107	1.0842	0.7884	1.1888	0.9125	0.7369	0.6425	1.5452	0.9718	0.4187	0.8819	0.8879
MSC-GA	970	0.8927	0.8163	0.8344	1.0518	0.7592	1.2228	0.7840	1.2011	0.9214	0.6988	0.7582	1.2709	0.9682	0.4443	0.8970	0.8293
MSC-SPA	23	0.8319	1.0216	0.8138	1.1151	0.5864	1.6025	0.6539	1.5204	0.9137	0.7322	0.6219	1.5890	0.9598	0.4999	0.8234	1.0862
MSC-UVE	569	0.9029	0.7764	0.7933	1.1749	0.7519	1.2413	0.7872	1.1921	0.8985	0.7941	0.6436	1.5428	0.9681	0.4448	0.8837	0.8814
MSC-CARS	46	0.9663	0.4575	0.9265	0.7008	0.5607	1.6516	0.5969	1.6408	0.8867	0.8389	0.7131	1.3843	0.9821	0.3349	0.9430	0.6168
MSC-LARS	500	0.9402	0.6091	0.7124	1.3860	0.7171	1.3254	0.7710	1.2367	0.9039	0.7725	0.6460	1.5375	0.9783	0.3673	0.8169	1.1059

Note: <Preprocessing> − <Feature Selection> = Spectral data after preprocessing and feature selection methods.

Upon longitudinal comparison of Table 4, it is evident that the model prediction accuracy of the feature bands selected by both the SPA and LARS algorithms significantly decreases compared to the full bands on the different models. The R^2 of the test set for the feature bands selected by SPA decreases on average by 6.47% compared to the full spectral data on each model, while the R^2 of the bands selected by LARS decreases on average by 5.16% compared to the full spectral data on each model. This indicates that these two methods failed to effectively extract the useful feature bands, resulting in lower prediction accuracy for each model. The prediction effect of the feature bands selected by the UVE algorithm on each model is closer to that of the full band, indicating that the use of noise to eliminate invalid variables in the spectral data is beneficial and retains a large amount of valid spectral information. GA and CARS perform better in extracting valid spectral data, leading to an improvement in model prediction accuracy to different degrees compared to the full band. Among them, the CARS algorithm shows the most significant improvement, with the highest R^2 improvement of 14.99% and RMSE reduction of 38.48% on PLSR and the highest R^2 of 0.9430 and RMSE of 0.6168 in the stacking model. Based on the results, it is evident that the method of adaptively weighting the spectral bands for the selection of feature bands in *P. massoniana* leaf spectral data is the most effective approach.

The comparison in Table 4 shows that the prediction accuracy of the Adaboost model in the training set is generally higher than that in the test set for each dataset, indicating that the model is more sensitive to data imbalance, resulting in decreased prediction accuracy. In contrast, the prediction accuracy of the SVR on both the training and test sets has a smaller fluctuation range, is robust to outliers, and has excellent generalization ability. However, the overall model′s performance is low. It is possible that the low performance of the model is due to the excessive number of features. PLSR performed the best in the single model, especially in the MSC-CARS dataset, where R^2 reached 0.9265, and the MSC-CARS-Stacking model still had a 0.0165 improvement in R^2 in the same dataset. The stacking integrated learning model demonstrated improvement relative to the single model on each spectral dataset. The MSC-CARS-stacking model had the best overall improvement, with an average increase of 26.49% in R^2 and an average reduction of 50.34% in RMSE for the test set compared to the single learner. The worst performance improvement was observed in the MSC-GA-stacking integrated model, with an average test set R^2 improvement of 13.23% and an average RMSE reduction of 29.40%. This also confirms the effectiveness of the CARS method for spectral information extraction and the deficiency of the GA feature screening method. The CARS method selects the least number of feature bands in the interval of 4600–4000 cm^{-1}. It is seen that CARS selects the bands related to O-H chemical bonds well and filters the invalid information greatly. The number of overlapping feature waves between GA and CARS is 25, which is the highest among the overlapping wave statistics between CARS and SPA or UVE. However, GA has a large number of feature variables, which is 970, and the number of overlapping waves with CARS is only 2.5%. This suggests that GA contains a significant amount of invalid information. GA increases the complexity of spectral feature search, and its mutation and crossover information composition is not sufficient to traverse the solution space of large spectral data with complex information.

For the full-spectrum data and the spectral data of the characteristic bands selected by the five algorithms (GA, SPA, UVE, CARS, and LARS), the prediction accuracy of the stacking integrated model consisting of SVR, AdaBoost, and PLSR proposed in this paper is improved to different degrees compared with each single model. This is because the stacking integrated model first trains several weak learners in parallel to combine the advantages of multiple models, and then combines the different weak learners by training a meta-learner to output a final prediction result. This effectively overcomes the shortcomings of a single model in the training process, leading to improved prediction accuracy. The prediction results of the full-spectrum data and the spectral data of the feature bands selected by the five algorithms of GA (970), SPA (23), UVE (569), CARS (46), and LARS (500) in the stacking integrated model are shown in Figure 7.



(**a**)

(**b**)

(**c**)

(**d**)

(**e**)

(**f**)

Figure 7. Model prediction effect under optimal combination. <Preprocessing> − <Feature Selection> − <Model> = Model results after preprocessing and feature extraction methods: (**a**) MSC-stacking; (**b**) MSC-GA-stacking; (**c**) MSC-SPA-stacking; (**d**) MSC-UVE-stacking; (**e**) MSC-CARS-stacking; and (**f**) MSC-LARS-stacking.

4. Conclusions

This study aims to develop a predictive model for the leaf moisture content of *P. massoniana* seedlings using near-infrared spectroscopy and multiple chemometric methods. Five preprocessing methods, including Nirmaf, L2-normalize, MSC, SG, and SNV, were employed to process the spectral data. A stacking learning framework was then introduced, and six models, namely AdaBoost, ExtraTree, RF, SVR, PLSR, and Ridge, were analyzed based on their predictive results. PLSR and AdaBoost were chosen as the candidate base learners for the stacking learning model, and the remaining four models were allocated to the base learners and meta-learners. The optimal combination of learners was determined by searching for all possible combinations. Furthermore, the SPA, CARS, UVE, GA, and LARS algorithms were utilized to select the wavelength variables, and the MSC full spectrum and MSC characteristic wavelength spectrum were used to establish a quantitative analysis model for the *P. massoniana* seedling moisture content based on the selected stacking learning model. The results show that the SVR-AdaBoost-PLSR+AdaBoost stacking learning model in full-spectrum near-infrared spectroscopy can accurately quantify the moisture content of *P. massoniana* seedlings. As there are many spectral variables, the commonly used multivariate calibration methods in chemometrics are no longer applicable. The SVR-AdaBoost-PLSR+AdaBoost model still shows stable predictive performance in feature variable selection, indicating that the stacking learning model has good applicability and predictive performance in near-infrared spectroscopy quantitative analysis. As a result, this model holds significant potential for further research in the field of spectroscopy analysis. The modeling methods and procedures used in this study are also applicable to other forest seedlings. This provides a reference for the precise cultivation technology of *P. massoniana* seedlings and an effective and accurate modeling method for quantitatively analyzing seedling moisture content.

Author Contributions: Conceptualization, Y.L. (Yurong Li) and H.X.; methodology, Y.L. (Yurong Li) and H.X.; software, Y.L. (Yurong Li) and H.X.; validation, L.H. and B.G.; formal analysis, Y.L. (Yurong Li) and H.X.; investigation, B.G.; resources, C.N.; data curation, Y.L. (Yurong Li) and H.X.; writing—original draft preparation, Y.L. (Yurong Li) and H.X.; writing—review and editing, Y.L. (Ying Liu); visualization, Y.L. (Yurong Li), H.X. and C.N.; supervision, Y.L. (Ying Liu); project administration, Y.L. (Ying Liu); funding acquisition, Y.L. (Ying Liu). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (NSFC: 31570714).

Data Availability Statement: The data are not publicly available because this study is still in progress.

Acknowledgments: We thank Zilong Zhuang, Wei Yu, Chenxi Zhou, Kaihao Ji, Qi Sun and Yinxi Shen for their help with field work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yang, R.; Meng, J. Using Advanced Machine-Learning Algorithms to Estimate the Site Index of Masson Pine Plantations. *Forests* **2022**, *13*, 1976. [CrossRef]
2. Chen, F.; Yuan, Y.-j.; Yu, S.-l.; Zhang, T.-w. Influence of climate warming and resin collection on the growth of Masson pine (*Pinus massoniana*) in a subtropical forest, southern China. *Trees-Struct. Funct.* **2016**, *30*, 1017. [CrossRef]
3. Dungey, H.S.; Dash, J.P.; Pont, D.; Clinton, P.W.; Watt, M.S.; Telferl, E.J. Phenotyping Whole Forests Will Help to Track Genetic Performance. *Trends Plant Sci.* **2018**, *23*, 854–864. [CrossRef] [PubMed]
4. Scharwies, J.D.; Dinneny, J.R. Water transport, perception, and response in plants. *J. Plant Res.* **2019**, *132*, 311–324. [CrossRef]
5. Wang, J.; Li, X.; Wang, W.; Wang, F.; Liu, Q.; Yan, L. Research on Rapid and Low-Cost Spectral Device for the Estimation of the Quality Attributes of Tea Tree Leaves. *Sensors* **2023**, *23*, 571. [CrossRef]
6. Aboulwafa, M.M.; Youssef, F.S.; Gad, H.A.; Sarker, S.D.; Nahar, L.; Al-Azizi, M.M.; Ashour, M.L. Authentication and discrimination of green tea samples using UV-vis, FTIR and HPLC techniques coupled with chemometrics analysis. *J. Pharm. Biomed. Anal.* **2019**, *164*, 653–658. [CrossRef]
7. Rebufa, C.; Pany, I.; Bombarda, I. NIR spectroscopy for the quality control of *Moringa oleifera* (Lam.) leaf powders: Prediction of minerals, protein and moisture contents. *Food Chem.* **2018**, *261*, 311–321. [CrossRef]
8. Zhang, H.; Ge, Y.; Xie, X.; Atefi, A.; Wijewardane, N.K.; Thapa, S. High throughput analysis of leaf chlorophyll content in sorghum using RGB, hyperspectral, and fluorescence imaging and sensor fusion. *Plant Methods* **2022**, *18*, 60. [CrossRef]
9. Shi, G.; Cao, J.; Li, C.; Liang, Y. Compression strength prediction of Xylosma racemosum using a transfer learning system based on near-infrared spectral data. *J. For. Res.* **2020**, *31*, 1061–1069. [CrossRef]
10. Falcioni, R.; Moriwaki, T.; Antunes, W.C.; Nanni, M.R. Rapid Quantification Method for Yield, Calorimetric Energy and Chlorophyll a Fluorescence Parameters in Nicotiana tabacum L. Using Vis-NIR-SWIR Hyperspectroscopy. *Plants* **2022**, *11*, 2406. [CrossRef] [PubMed]
11. Zhang, K.; Jiang, H.; Zhang, H.; Zhao, Z.; Yang, Y.; Guo, S.; Wang, W. Online Detection and Classification of Moldy Core Apples by Vis-NIR Transmittance Spectroscopy. *Agriculture* **2022**, *12*, 489. [CrossRef]
12. Zhang, M.; Guo, J.; Ma, C.; Qiu, G.; Ren, J.; Zeng, F.; Lu, E. An Effective Prediction Approach for Moisture Content of Tea Leaves Based on Discrete Wavelet Transforms and Bootstrap Soft Shrinkage Algorithm. *Appl. Sci.* **2020**, *10*, 4839. [CrossRef]
13. Li, C.; Zhao, J.; Li, Y.; Meng, Y.; Zhang, Z. Modeling and Prediction of Soil Organic Matter Content Based on Visible-Near-Infrared Spectroscopy. *Forests* **2021**, *12*, 1809. [CrossRef]

14. Neto, A.J.S.; Lopes, D.d.C.; Silva, T.G.F.d.; Ferreira, S.O.; Grossi, J.A.S. Estimation of leaf water content in sunflower under drought conditions by means of spectral reflectance %J Engineering in Agriculture, Environment and Food. *Eng. Agric. Environ. Food* **2016**, *10*, 104–108. [CrossRef]
15. Rabanera, J.D.; Guzman, J.D.; Yaptenco, K.F. Rapid and Non-destructive measurement of moisture content of peanut (*Arachis hypogaea* L.) kernel using a near-infrared hyperspectral imaging technique. *J. Food Meas. Charact.* **2021**, *15*, 3069–3078. [CrossRef]
16. Sun, Z.; Zhou, J. L-1-PLS Based on Incremental Extreme Learning Machine. In Proceedings of the 9th IEEE Data Driven Control and Learning Systems Conference (DDCLS), Liuzhou, China, 20–22 November 2020; pp. 947–952. [CrossRef]
17. Hu, H.; He, Z.; Ling, Y.; Li, J.; Sun, L.; Li, B.; Liu, J.; Chen, W. A SOM-RBFnn-Based Calibration Algorithm of Modeled Significant Wave Height for Nearshore Areas. *J. Mar. Sci. Eng.* **2022**, *10*, 706. [CrossRef]
18. Yang, J.; Chen, Y. Tender Leaf Identification for Early-Spring Green Tea Based on Semi-Supervised Learning and Image Processing. *Agronomy* **2022**, *12*, 1958. [CrossRef]
19. Xie, C.; Zhu, H.Y.; Fei, Y.Q. Deep coordinate attention network for single image super-resolution. *IET Image Process.* **2022**, *16*, 273–284. [CrossRef]
20. Lian, X.-q.; Chen, Q.; Tang, S.-m.; Wu, J.-z.; Wu, Y.-l.; Gao, C. Quantitative Analysis Method of Key Nutrients in Lanzhou Lily Based on NIR and SOM-RBF. *Spectrosc. Spectr. Anal.* **2022**, *42*, 2025–2032. [CrossRef]
21. He, W.; Li, Y.; Wang, J.; Yao, Y.; Yu, L.; Gu, D.; Ni, L. Using Field Spectroradiometer to Estimate the Leaf N/P Ratio of Mixed Forest in a Karst Area of Southern China: A Combined Model to Overcome Overfitting. *Remote Sens.* **2021**, *13*, 3368. [CrossRef]
22. Wang, Y.; Wang, D.; Geng, N.; Wang, Y.; Yin, Y.; Jin, Y. Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection. *Appl. Soft Comput.* **2019**, *77*, 188–204. [CrossRef]
23. Ni, C.; Wang, D.; Tao, Y. Variable weighted convolutional neural network for the nitrogen content quantization of Masson pine seedling leaves with near-infrared spectroscopy. *Spectrochim. Acta Part A-Mol. Biomol. Spectrosc.* **2019**, *209*, 32–39. [CrossRef] [PubMed]
24. Sun, J.; Cong, S.; Mao, H.; Wu, X.; Zhang, X.; Wang, P. CARS-ABC-SVR model for predicting leaf moisture of leaf-used lettuce based on hyperspectral. *Trans. Chin. Soc. Agric. Eng.* **2017**, *33*, 178–184.
25. Yu, L.; Liang, Y.; Zhang, Y.; Cao, J. Mechanical properties of wood materials using near-infrared spectroscopy based on correlation local embedding and partial least-squares. *J. For. Res.* **2020**, *31*, 1053–1060. [CrossRef]
26. Ren, G.; Zhang, X.; Wu, R.; Yin, L.; Hu, W.; Zhang, Z. Rapid Characterization of Black Tea Taste Quality Using Miniature NIR Spectroscopy and Electronic Tongue Sensors. *Biosensors* **2023**, *13*, 92. [CrossRef]
27. Ma, L.; Zhang, Y.; Zhang, Y.; Wang, J.; Li, J.; Gao, Y.; Wang, X.; Wu, L. Rapid Nondestructive Detection of Chlorophyll Content in Muskmelon Leaves under Different Light Quality Treatments. *Agronomy* **2022**, *12*, 3223. [CrossRef]
28. Zhu, R.; Jiang, D. Frequency modulation analysis of solar array using genetic algorithm. *Proc. Inst. Mech. Eng. Part G-J. Aerosp. Eng.* **2022**, *10*, 19. [CrossRef]
29. Wang, Z.; Zhang, Y.; Fan, S.; Jiang, Y.; Li, J. Determination of Moisture Content of Single Maize Seed by Using Long-Wave Near-Infrared Hyperspectral Imaging (LWNIR) Coupled with UVE-SPA Combination Variable Selection Method. *IEEE Access* **2020**, *8*, 195229–195239. [CrossRef]
30. Wang, Z.; Chen, J.; Zhang, J.; Tan, X.; Raza, M.A.; Ma, J.; Zhu, Y.; Yang, F.; Yang, W. Assessing canopy nitrogen and carbon content in maize by canopy spectral reflectance and uninformative variable elimination. *Crop. J.* **2022**, *10*, 1224–1238. [CrossRef]
31. Tabarangao, J.T.; Slepkov, A.D. Mimicking Multimodal Contrast with Vertex Component Analysis of Hyperspectral CARS Images. *J. Spectrosc.* **2015**, *2015*, 575807. [CrossRef]
32. Elrewainy, A.; Sherif, S.S. Kronecker least angle regression for unsupervised unmixing of hyperspectral imaging data. *Signal Image Video Process.* **2020**, *14*, 359–367. [CrossRef]
33. Cheng, J.; Sun, J.; Yao, K.; Xu, M.; Wang, S.; Fu, L. Hyperspectral technique combined with stacking and blending ensemble learning method for detection of cadmium content in oilseed rape leaves. *J. Sci. Food Agric.* **2022**, *103*, 2690–2699. [CrossRef]
34. Mantanus, J.; Ziemons, E.; Lebrun, P.; Rozet, E.; Klinkenberg, R.; Streel, B.; Evrard, B.; Hubert, P. Moisture content determination of pharmaceutical pellets by near infrared spectroscopy: Method development and validation. *Anal. Chim. Acta* **2009**, *642*, 186–192. [CrossRef] [PubMed]
35. Ishikawa, H.; Boukar, O.; Fatokun, C.; Shono, M.; Muranaka, S. Development of calibration model to predict nitrogen content in single seeds of cowpea (*Vigna unguiculata*) using near infrared spectroscopy. *J. Near Infrared Spectrosc.* **2017**, *25*, 211–214. [CrossRef]
36. Breiman, L. Stacked Regressions. *Mach. Learn.* **1996**, *24*, 49–64. [CrossRef]
37. Divina, F.; Gilson, A.; Gomez-Vela, F.; Torres, M.G.; Torres, J.E. Stacking Ensemble Learning for Short-Term Electricity Consumption Forecasting. *Energies* **2018**, *11*, 949. [CrossRef]
38. Frost, R.L.; Erickson, K.L. Near-infrared spectroscopic study of selected hydrated hydroxylated phosphates. *Spectrochim. Acta Part A-Mol. Biomol. Spectrosc.* **2005**, *61*, 45–50. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.