remote sensing

MDPI

Article

Performance Comparison of Machine Learning Algorithms for Estimating the Soil Salinity of Salt-Affected Soil Using Field Spectral Data

Sijia Wang [1,2], Yunhao Chen [1,2,*], Mingguo Wang [3] and Jing Li [1,2]

[1] State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China
[2] Beijing Key Laboratory of Environmental Remote Sensing and Digital City, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China
[3] NingXia Agriculture Technology Extension Service Centre, Yinchuan 750001, China
* Correspondence: cyh@bnu.edu.cn; Tel.: +86-010-5880-6098

Received: 29 August 2019; Accepted: 5 November 2019; Published: 6 November 2019

Abstract: Salt-affected soil is a prominent ecological and environmental problem in dry farming areas throughout the world. China has nearly 9.9 million km^2 of salt-affected land. The identification, monitoring, and utilization of soil salinization have become important research topics for promoting sustainable progress. In this paper, using field-measured spectral data and soil salinity parameter data, through analysis and transformation of spectral data, five machine learning models, namely, random forest regression (RFR), support vector regression (SVR), gradient-boosted regression tree (GBRT), multilayer perceptron regression (MLPR), and least angle regression (Lars) are compared. The following performance measures of each model were evaluated: the collinear problems, handling data noise, stability, and the accuracy. In terms of these four aspects, the performance of each model on estimating soil salinity is evaluated. The results demonstrate that among the five models, RFR has the best performance in dealing with collinearity, RFR and MLPR have the best performance in dealing with data noise, and the SVR model is the most stable. The Lars model has the highest accuracy, with a determination coefficient (R^2) of 0.87, ratio of performance to deviation (RPD) of 2.67, root mean square error (RMSE) of 0.18, and mean absolute percentage error (MAPE) of 0.11. Then, the comprehensive comparison and analysis of the five models are carried out, and it is found that the comprehensive performance of RFR model is the best; hence, this method is most suitable for estimating soil salinity using hyperspectral data. This study can provide a reference for the selection of regression methods in subsequent studies on estimating soil salinity using hyperspectral data.

Keywords: RFR; SVR; GBRT; MLPR; Lars; soil salt content; hyperspectral

1. Introduction

Salt-affected soil is a general term that refers to saline soil and alkaline soil. The content of soluble salt substances in saline soil typically exceeds 2 g/kg, which affects the normal development of crops. Alkaline soil is classified according to the alkalization degree and has a pH that exceeds 8 [1]. Typically, salt soil and alkaline soil are mixed; hence, they are collectively referred to as salt-affected soil. Salt-affected soil is a prominent ecological and environmental problem in the world's dry farming areas [2–4]. The total area of the salt-affected soil resources in China is approximately 9.9 million km^2, which are mainly distributed in the northeast plain, the arid and semi-arid areas in the northwest, the Huang-Huai-Hai plain, and the eastern coastal areas. Among them, the arid area in the northwest is China's largest salt-affected soil distribution area, with a total area of approximately 1.3 million km^2, which includes Qinghai, Xinjiang, western Inner Mongolia, Gansu Hexi Corridor, and northern Ningxia.

The second-largest is a coastal salt-affected soil region with an area of approximately 0.8 million km^2. This region is mainly distributed along the coasts of the Yellow Sea, the Bohai Sea, and the East China Sea [5]. The identification, monitoring, prevention, development, and utilization of soil salinization have become important research topics for social and economic development and for promoting sustainable progress. In recent years, with the development of technologies such as geographic information systems, remote sensing technology and global positioning systems, increasingly many remote sensing technologies have been applied to the research and application of monitoring and measurement of salt-affected land [6]. Hyperspectral remote sensing plays an increasingly important role in the identification and monitoring of salt-affected land because of its richer spectral information.

There are three main approaches for studying soil salinization with hyperspectral data: One is to use hyperspectral image data to construct a distribution map of salt-affected land and to analyze the changes and driving factors of salt-affected land [7–10]. The second is to study the growth of crops under salt stress based on hyperspectral data and to infer the soil salt content from the crop growth [11,12]. The third is to invert the degree of soil salinization by using hyperspectral data to study the relationship between the field hyperspectral data and the measured parameters that are related to the soil salinity [13–16]. Here, we focus on the use of spectral data to estimate or invert the extent of soil salinization. The methods that are commonly used in this research include spectral decomposition methods and regression analysis methods [17–21]. Spectral decomposition methods are often used to estimate the degree of soil salinization using hyperspectral images. Typically, the pixels of hyperspectral images are mixed pixels. It is necessary to obtain pure endmembers via spectral demixing and other methods and to use the endmembers to estimate soil salinity [18,19]. Regression analysis methods are commonly used to estimate soil salinity based on near-end measured spectra [13,21,22].

The regression methods that are used for soil salinity estimation are mainly divided into traditional regression analysis methods and machine learning methods. Traditional regression analysis methods include least squares regression and partial least squares regression (PLSR) [21,23]. The least squares method identifies the best-matching function to the data by minimizing the sum of the squared errors. However, this method is highly sensitive to outliers, especially when the soil salt content is unevenly distributed and there is a maximum or minimum value [24,25]. The partial least squares regression method can overcome the multicollinearity problem of independent variables in regression analysis by combining principal component analysis and canonical correlation analysis on the basis of ordinary multiple regression. This method shows satisfactory applicability in the face of hyperspectral multidimensional spectral data [26–28]. Peng et al. (2019) used PLSR method in regression electrical conductivity (EC) to estimate soil salinity, and the results showed that the accuracy and stability of the model were good [29].

Machine learning methods use algorithms to parse data, to learn from data and to make decisions and predictions about events in the real world. Unlike traditional methods for solving specified tasks, machine learning uses a large amount of data to "train" and learns how to accomplish tasks from the data via various algorithms [30,31]. Commonly used machine learning methods include decision trees, support vector machines, multilayer perceptron, and regularization [30]. Random tree and support vector machines show strong robustness when facing high dimensional data. The hidden layer in the multilayer perceptron can reduce the dependence of the algorithm on the data. The regularization can reduce the influence of collinearity by adding offset to the optimization function [30]. With the deep research of machine learning methods, the results are uneven and many studies in which machine learning methods are used to estimate soil salinity have been reported [4,21,24,32–34]. Jiang et al. monitored soil salinity by integrating multiple biophysical indicators with support vector machine (SVM) and artificial neural network (ANN) regression algorithms. The results demonstrate that the SVM regression algorithm outperforms the ANN algorithm in monitoring soil salinity [4]. Farifteh used the ANN algorithm and the PLSR algorithm to estimate the soil salinity. When using field data for the estimation, the R^2 value that was obtained by the ANN algorithm is $0.42 < R^2 < 0.69$ and the R^2 value that was obtained by the PLSR algorithm is 0.8. However, on the experimental data, the R^2

value that was obtained by the ANN algorithm exceeds 0.92 and the R^2 value that was obtained by PLSR is 0.81 [21]. Wang et al. used the PLSR algorithm and the random forest (RF) algorithm to measure soil salinity. According to the validation accuracies, the RF models outperformed the PLSR models [34]. These studies were conducted in different data and different environments, and the results obtained cannot be put together for comprehensive evaluation of these models. It is also unable to comprehensively judge the accuracy of each model and the applicable data of each model. Therefore, it is necessary to analyze and compare the models from multiple perspectives with the same data. In common machine models, random forest regression (RFR) and gradient-boosted regression tree (GBRT) in random tree, support vector regression (SVR) in support vector machine, multilayer perceptron regression (MLPR) in multilayer perceptron, and least angle regression (Lars) in regularization are selected for comparison.

This paper makes a comparative analysis of machine learning model (random forest regression (RFR), support vector regression (SVR), gradient-boosted regression tree (GBRT), multilayer perceptron regression (MLPR), and least angle regression (Lars)) from four perspectives. The performance of each model to deal with collinearity problem was compared by inputting different number of bands, adding different degree of Gaussian white noise to compare the performance of each model to deal with data noise, changing the number of training data of the model to compare the stability of each model, and comparing the accuracy of each model by leave-one-out cross-validation method. Then, the performance of each model to estimate soil salt content with hyperspectral data was evaluated comprehensively.

2. Materials

2.1. Study Area

The data collection area is Shizuishan City in Ningxia hui autonomous region in northwestern China. Shizuishan City is located between 105°58′~106°39′ east longitude and 38°21′~39°25′ north latitude, in the upper reaches of the middle reaches of the Yellow River. Shizuishan City is approximately 88.8 km wide from east to west and 119.5 km long from north to south and at an altitude of 1090~3475.9 m (Figure 1) [35]. This area has a typical temperate continental climate, with an average temperature of 8.4 to 9.9 °C and an average precipitation of 167.5–188.8 mm. The soil in this area is mainly viscous soil. Because of the abundant sunshine throughout the year, concentrated rainfall, strong evaporation, and high groundwater level, coupled with the irrigation with Yellow River water, the large amount of cultivated land in this area exhibits various degrees of salinization [36].

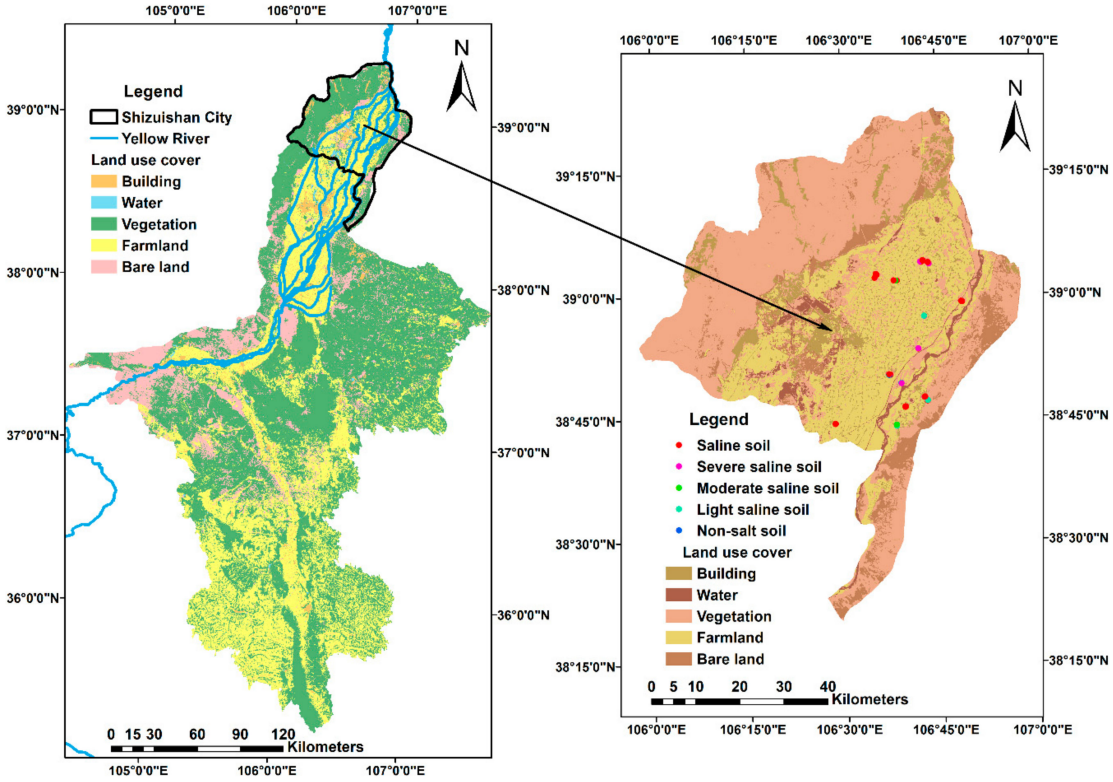

Figure 1. Study area. The left figure shows the geographical location of Shizuishan City. The figure on the right shows the distribution of the field data sampling points in Shizuishan City. The color corresponds to the degree of salinization according to the soil salt content at the sampling point. Red indicates the highest degree of salinization and blue indicates that the soil is non-salt alkali soil. The land use cover that are presented in the figure are derived from the land use survey in 2010.

2.2. Data Collection

From April 7 to April 10, 2018, we collected a total of 60 sets of data in Shizuishan. Each set of data was acquired via a five-point sampling method. First, we used a Spectra Vista Corporation (SVC) spectrometer to measure the spectral data without artificially disturbing the land. The related parameter information of the SVC spectrometer is listed in Table 1. The spectral measurement time was between 10 am and 2 pm. To reduce the error during measurement, a whiteboard was used for calibration prior to each measurement. During measurement, the probe is vertically downward and about 60 cm away from the ground. Then, a surface soil sample was collected at the place where the spectrum was measured, placed in an aluminum cassette, returned to the laboratory for processing, and sent to the Ningxia Agricultural Technology Extension Station and the Analytical Testing Center of Beijing Normal University for measurement of the soil salt content and the main salt segregant content. Finally, the latitude and longitude information of the data collection point was obtained by using a handheld GPS instrument at the data collection location [37]. Figure 2 presents the processed field spectra and photos of field sampling points.

Several obvious absorption zones can be seen in Figure 2, among which the water absorption zone is near 1400 nm and 1900 nm, and the clay mineral absorption zone is near 2200 nm [12]. Soil moisture is often a major interfering factor in the inversion of soil substances by using hyperspectral. Numerous studies have shown that the absorption zone of moisture is around 1450 nm and 1940 nm [38–40]. Therefore, in the subsequent analysis, the two absorption bands of water will be removed.

Table 1. Main performance parameters of the instrument.

Attributes	SVC HR-1024i
Spectral range	350 nm–2500 nm
Spectral resolution	Silicon array ≤ 3.5 nm InGaAs array ≤ 9.5 nm Extended InGaAs array ≤ 6.5 nm
Number of bands	1024
Field of view	25°

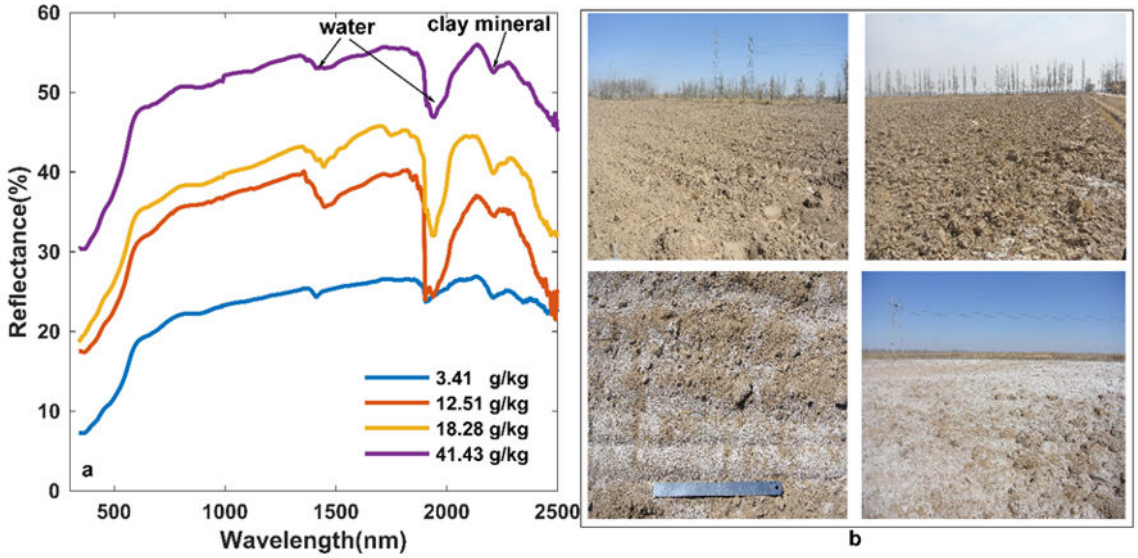

Figure 2. (**a**) shows the field-measured spectra after pretreatment and (**b**) shows photos of field sampling points. The spectra in a correspond one to one with the photos in (**b**). (**b**) shows the surface characteristics as soil salt content increases.

2.3. Soil Parameters

The pH, soil salinity, and sodium, potassium, magnesium, calcium, chloride, nitrate, sulfate, carbonate, and bicarbonate contents of the soil samples that were collected in the field were measured in the laboratory. The minimum, maximum, and average values (Table 2) of these quantities for 60 samples were calculated and a data distribution histogram and a normal probability map (Figure 3a,b) was plotted. The pH value of the sampled soil was at least 7.53; hence, all samples were alkaline. The salt content ranged from 2919.76 to 290857.70 mg/kg. The contents of sodium ion, chloride ion, and sulfate ion in the soil salt were high, which are the main constituent ions of the soil salt. According to Figure 3a,b, the soil salt content distribution in the collected samples is not a normal distribution. In most samples, the soil salt content is low and only a small part of the sample has a very high soil salt content. Such sample data may produce high R^2 values; however, the RMSE values could be proportionally high. To avoid this scenario, we transformed the soil salt content data and calculated the base 10 logarithm so that the converted soil salt content data follows a normal positive distribution (Figure 3c,d).

Table 2. Soil parameters and related conditions.

	PH	Salt Content	Na^+	K^+	Mg^{2+}	Ca^{2+}	Cl^-	NO_3^-	SO_4^{2-}	CO_3^{2-}	HCO_3^-
Unit	1	mg/kg	mg/kg	mg/kg	mg/kg	mg/kg	mg/kg	mg/kg	mg/kg	mg/kg	mg/kg
Min	7.53	2919.76	327.71	28.40	43.75	187.15	178.39	22.94	14.64	0.00	145.69
Max	9.98	290857.70	87551.62	370.24	7508.08	9416.35	89678.66	2701.86	98199.80	3133.70	891.37
Mean	8.56	43827.40	12840.01	101.25	1263.77	3291.60	12103.64	502.78	13715.98	63.81	321.44

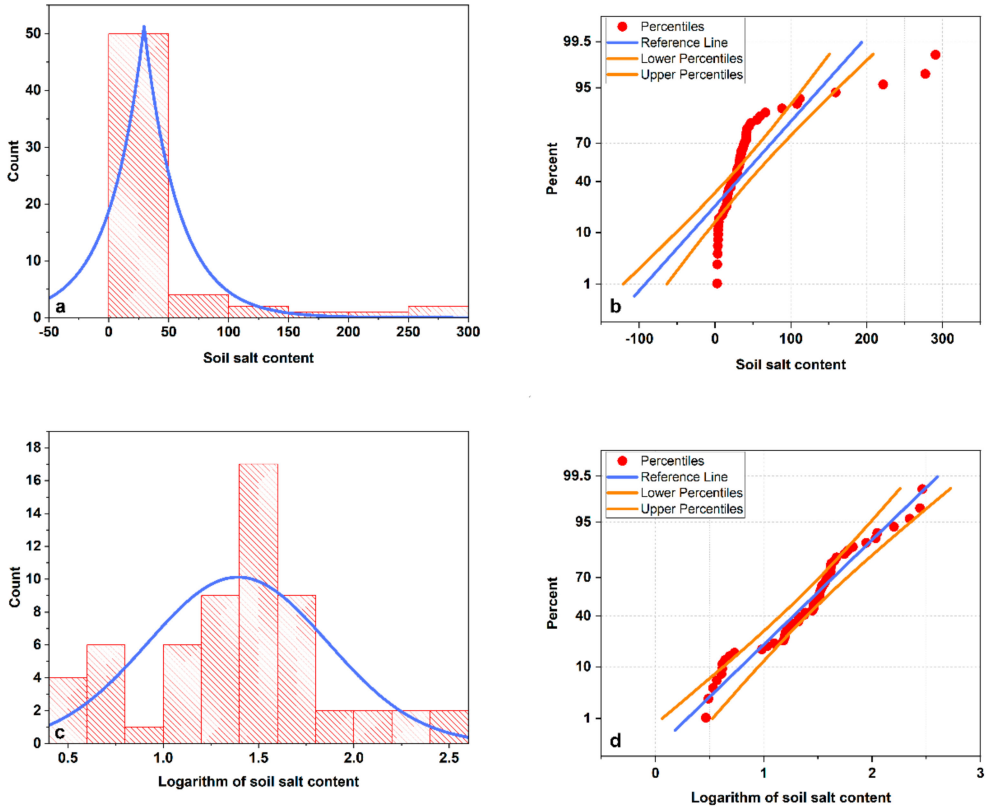

Figure 3. (**a**) presents a histogram of the soil salt content. (**b**) shows the normal probability distributions of the soil salt content. (**c**) presents a histogram of the logarithm of the soil salt content. (**d**) shows the normal probability distributions of the logarithm of the soil salt content. The solid blue line in (**a**) and (**c**) is the distribution curve that is fitted according to the data. The blue line in (**b**) and (**d**) is the reference line, where the data obey a normal distribution, and the yellow line is the reference line for the upper and lower limits of the case in which the data obey a normal distribution. The red dot indicates the distribution of the current data.

3. Methods

The overall technical approach of this paper is divided into three main parts (Figure 4): The first part is data acquisition and preprocessing, which mainly involves the collection of field spectral data, spectral data de-noising, resampling and smoothing, and field soil sample collection processing and parameter measurement. The de-noising and resampling of spectral data are carried out by using the software (SVC HR-1024i, Version 1.17.14) provided by the measuring instrument. Without affecting the overall variation trend of spectral curve, the noise of the data is removed and the data is interpolated into more than 2000 bands. The smoothing of spectral data is carried out by using the five-point method. The second part is the data analysis stage, which mainly transforms the spectral data and analyses the correlations between various forms of spectral data and soil parameter data. The third part is the core part of this paper, namely, the comparative analysis of the model. First, according to the characteristics of the parameters of each model and the data characteristics, after repeated testing, the optimal parameters of each model are determined. Then, 60 groups of sample data were divided into training group data and test group data. Specifically, 50 groups of samples were used for training and the remaining 10 groups of samples were used for testing. In the model comparison, 50 sets of trained data were used for modeling, and the remaining 10 sets of data were tested and compared. Finally, five models are comprehensively compared and analyzed from the four aspects: performance in dealing with collinear problems, performance in dealing with noisy problems, stability and model accuracy.

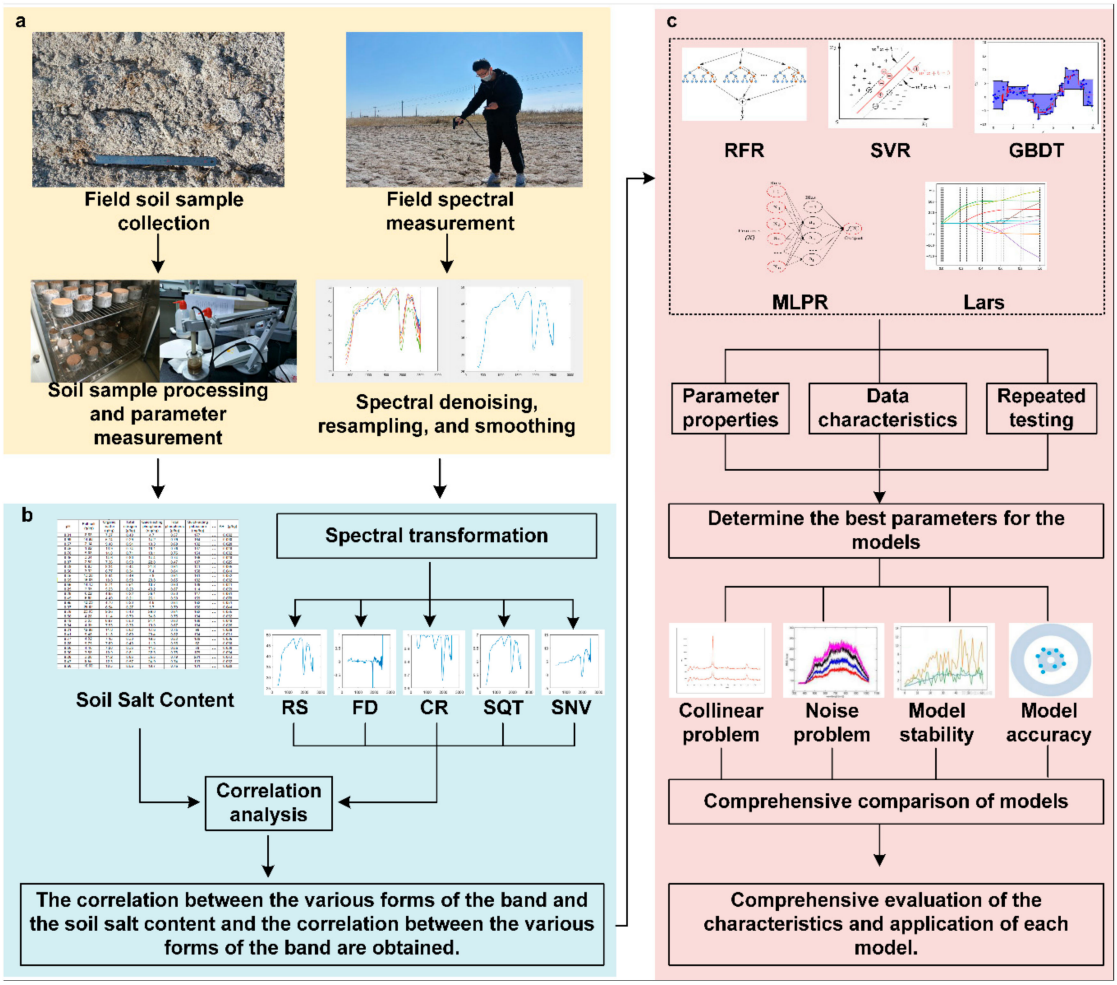

Figure 4. Technical flow chart. (**a**) represents the data collection and pre-processing process; (**b**) represents the data analysis process; and (**c**) represents the model comparison analysis process.

3.1. Data Preprocessing Method

The spectral data that were measured in the field were subjected to a series of denoising and smoothing operations, which were conducted using the software that comes with the instrument. Then, since the sampling method uses a five-point sampling method, the five spectral curves that are measured at five points are averaged to obtain the field spectral data of each sampling point. Finally, the raw spectral (RS) data were transformed into four types, namely, first derivative (FD) [41], continuum removal (CR) [41], square root (SQT) [41], and standard normal variate (SNV) [42], and five types (RS, FD, CR, SQT, SNV) of spectral data were obtained for each sampling point.

3.2. Machine Learning Method

3.2.1. Random Forest Regression (RFR)

In the random forest method, a random approach is used to build a forest. There are many decision trees in the forest and there are no correlations among the decision trees. When a new sample is input, each decision tree in the forest evaluates it separately. In the regression problem, the random forest outputs the average of all decision tree outputs [43]. The random forest algorithm is a bagging algorithm and bagging is an integrated learning method. The main strategy is to train multiple weak models to form a strong model. The performance of the strong model is far superior to that of a single weak model [44]. In a random forest, each decision tree "plants" and "grows" in four main steps:

1. Suppose the training set size is N. The N samples are obtained via repeated multiple sampling with reset. The sampling results will be used as the training set for our decision tree;
2. If there are M input variables, each node will randomly select m (m < M) variables and use these m variables to determine the best split point. During the generation of the decision tree, the value of m remains unchanged;
3. Each decision tree grows as much as possible without pruning;
4. New data are predicted by summing all decision trees (using majority voting in classification and averaging in regression) [43,45].

The RFR predictor is:

$$\hat{f}_{rf}^{K}(x) = \frac{1}{K}\sum_{k=1}^{K} T(x) \quad (1)$$

where x represents an input vector that is composed of various evidential features, k represents the number of regression trees that are constructed in RFR, and T(x) represents each constructed tree [46].

In this paper, RFR is implemented in Python (computer programming language) based on classification and regression tree (CART). CART is a model that is implemented internally by sklearn and spark [47]. Multiple binary decision trees are packaged into RFRs. When using RFR for regression, the mean square error is minimized. The main parameters of the RFR call in Python are "n_estimators" and "max_features." "N_estimators" represents the number of trees in the forest. The larger the value, the higher the performance but also the longer the calculation time. "Max_features" represents the size of the random feature subset that is considered when splitting nodes.

3.2.2. Support Vector Regression (SVR)

SVR is an application of support vector machine (SVM) to the regression problem, namely, to find a regression plane that minimizes the distance of all the data of a set from the plane. For nonlinear models, the data must be projected into the feature space. Then, a linear classifier is used in the feature space. To avoid huge computational complexity when mapping to the feature space, a kernel function is introduced when performing support vector machine regression. Kernel function is the transformation of feature from low dimension to high dimension, but it is first calculated on the low dimension, and the actual classification effect is shown in the high dimension, which avoids the complex calculation of high dimension and can get the same result. [48]. In this article, SVR is implemented in Python. The common parameters when calling the function are kernel and C, where kernel indicates the type of kernel to be used in the algorithm and C indicates the penalty parameter for the error, namely, a regularization parameter, which is a trade-off between the adjustment function complexity and the tolerance of the empirical material [47,49]. More time is required for training if C is larger; however, the prediction results stop improving after a threshold. The SVR function can be expressed as:

$$f(x) = \sum_{i=1}^{N}\left(\alpha_i - \alpha_i^*\right)\cdot k(x, x_i) + b \quad (2)$$

where N is the number of samples, $\left(\alpha_i - \alpha_i^*\right)$ is the Lagrange multiplier, $k(x, x_i)$ is the kernel function, and b is the bias term [50,51].

3.2.3. Gradient-Boosted Regression Tree (GBRT)

The GBRT is an iterative decision tree algorithm that consists of multiple decision trees and the conclusions of all the trees are summed to obtain the final answer. The GBRT was considered together with the SVM to be a highly generalized algorithm when it was proposed. The core strategy is that each tree learns the residuals of all previous tree conclusions. This residual is the cumulative amount that can be obtained after adding the predicted value [52,53]. The typical parameters that GBRT uses

in Python are "learning_rate," "n_estimators," "subsample," "max_depth," and "loss": "learning_rate" represents the learning rate, on which the contribution of each tree depends; "n_estimators" represents the number of boosting stages to be executed; "subsample" represents the sample score of each basic learner that is used for fitting; "max_depth" represents the maximum depth of the regression estimator, where the maximum depth limits the number of nodes in the tree; and "loss" represents the loss function optimization method. It is necessary to balance "learning_rate" with "n_estimators" and "n_estimators" with "subsample" when the model is called [47].

$$f(x) = \text{predict}(0) + \sum_{m=1}^{M} s * \text{predict}(T_m) \tag{3}$$

where $\text{predict}(0)$ is the initial predicted value, M is the number of trees, s is the scaling factor, and $\text{predict}(T_m)$ is the predicted value of each tree [54].

3.2.4. Multilayer Perceptron Regression (MLPR)

Multilayer perceptron regression is also called artificial neural network. The perceptron is a simple neuron model, which is a precursor to large neural networks, and typically consists of an input layer, an output layer, and multiple hidden layers. The multilayer perceptron layer is fully connected to the next layer, namely, each neuron in the upper layer is connected to all neurons in the next layer. The layers are articulated by weights and the neurons are arranged in the layers. In each layer, the nodes receive input only from the nodes in the previous layer and only pass their output to the nodes of the next layer [55,56]. Common parameters for calling MLPR in Python are "hidden_layer_sizes," which specifies the number of neurons in the hidden layer; "activation," which indicates the function that is used in the hidden layer [47]. This model optimizes the squared loss via stochastic gradient descent.

$$f(x) = G(b^{(2)} + W^{(2)}(s(b^{(1)} + W^{(1)}x))) \tag{4}$$

where $b^{(1)}$ and $W^{(1)}$ are the offset vector and the weight matrix, respectively, of the output layer to the hidden layer; s is the activation function of the layer; $b^{(2)}$ and $W^{(2)}$ are the offset vector and the weight matrix, respectively, of the output layer to the hidden layer; and G is the activation function of the layer [57].

3.2.5. Least Angle Regression (Lars)

Least angle regression is a regression algorithm for high-dimensional data. Similar to forward stepwise regression, at each step, it identifies the features that are most relevant to the target. When there are multiple features with equal correlation, instead of continuing along a single feature, it proceeds along an isometric direction between the features [58]. The parameter that is typically specified when LARS is called in Python is "n_nonzero_coefs," which represents the number of targets with non-zero coefficients [47].

$$\beta_k = \beta_k + \delta_k * \text{sign}\left(f_k^T r\right) \tag{5}$$

where r is the initial state; k is the feature that is most relevant to r; $\text{sign}\left(f_k^T r\right)$ is the forward direction, namely, β_k is updated in the direction of $\text{sign}\left(f_k^T r\right)$; and δ_k is the step size [59].

3.2.6. Ordinary Least Squares (OLS)

Ordinary least squares regression is a linear regression model that identifies the best-matching function for data by minimizing the sum of the squared errors. OLS is implemented in Python. The general form of ordinary least squares is as follows:

$$\hat{\lambda_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad (6)$$

$$\hat{\lambda_0} = \overline{y} - \hat{\lambda_1}\overline{x} \quad (7)$$

where x_i is the input variable, y_i is the measured value, $\overline{x}$ is the average of the input variables, and $\overline{y}$ is the average of the measured values [24].

3.3. Model Comparison Method

3.3.1. Collinear Problems

In using field hyperspectral data to evaluate the performance of each model in estimating soil salinity, the first evaluation criterion is the performance of each model in dealing with collinear problems. Collinearity problem means that one variable can be represented linearly by several other variables [60]. The field hyperspectral data that we use contains more than 2000 bands, namely, more than 2000 features (variables), which easily causes collinearity among variables. If no variable selection has been conducted or if a suitable regression model has not been selected, false regression can easily occur. In this paper, to evaluate the performance of each model in dealing with collinear problems, each model uses the five variations of the spectrum to conduct modelling regression analyses with 5, 10, 15, 20, 30, 40, 50, and 100 bands. These bands are sorted according to the correlation with the soil salt content from high to low prior to being input into the model and the corresponding number of bands are input in order. The performance of each model in dealing with collinear problems is evaluated by comparing the regression results.

3.3.2. Data Noise Problems

When measuring a spectrum, many errors are generated and the noise is expressed by the spectral curve. Reflections, scattering, and light from other objects in the measurement will enter the probe together with the reflected light of the salt-affected soil, thereby affecting the spectral information of the soil. In addition, the sensor that receives the information will introduce system noise into the spectral information. If the regression model is subsequently applied to satellite imagery, noise from atmospheric radiation will also be introduced [61]. A satisfactory model is necessary for performing regression analysis stably even if there is noise in the data. Gaussian white noise is an ideal model for analyzing channel additive noise, where Gaussian indicates that the probability distribution is a normal function and white noise indicates that its second-order moment is irrelevant and the first-order moment is constant. To evaluate the performance of each model in dealing with noise, the original data are added to Gaussian white noise in MATLAB according to the following signal-to-noise ratios: 5, 10, 15, 20, 30, 40, and 50. The data with various degrees of noise will be added to the regression analysis and the performance of each model in dealing with noise will be evaluated by comparison with the results of the regression analysis.

3.3.3. Stability

When using the model for regression analysis, a large change in the regression result will occur in response to a change in the amount of input data. Hence, the constructed model is only suitable for limited scenarios and the regression results will be unstable under changes in the amount of data [62]. To evaluate the stability of each model, the input data, namely, the training data, are set to 3/4, 2/4,

and 1/4 of the original data. The stability performance of each model is evaluated according to the magnitude of the change of each evaluation index as the number of input data changes.

3.3.4. Leave-One-Out Cross-Validation Method

In leave-one-out cross-validation, if the size of the data set D is N, then N-1 data items are used for training and the remaining data item is used for verification. The main disadvantage of using a single data item for verification is that there may be a large difference between the true value and the predicted value. Therefore, in leave-one-out cross-validation, a group is removed from D as a verification set in each round until all the samples have been evaluated, which requires a total of N calculations, and the verification error is averaged [63].

3.3.5. Evaluation Index

To evaluate the performance of the model in various aspects, we use determination coefficient (R^2), the ratio of the performance to the deviation (RPD), the root mean square error (RMSE), and the mean absolute percentage error (MAPE) evaluation indicators. Their calculation formulas are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{N}\left(Y_i - \overline{Y}_i\right)^2} \tag{8}$$

$$RPD = \frac{SD_s}{RMSE} \tag{9}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}{N}} \tag{10}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \tag{11}$$

where N is the sample size, Y_i is the measured value of soil sample i, $\hat{Y}_i$ is the salt content of soil sample i that is predicted by the models, $\overline{Y}_i$ is the average salt content of soil sample i, SD_s is the standard deviation of the measured salt content, and RMSE is the root mean square error of the predicted salt content. Larger values of R^2 and RPD and smaller values of RMSE and MAPE correspond to higher model performance [21,22,64].

4. Results

4.1. Spectrum Analysis

After the transformation of the raw spectral curve in four forms, four spectral curves were drawn for each spectral form, each curve representing different soil salt content (Figure 5). When the FD spectrum curve is near 600 nm, the higher the soil salt content, the larger the value. The CR spectrum curve shows patterns that differ according to the salt content at 400 nm and 1900 nm. The higher the SQT curve at a fixed salt content, the larger the value of the curve. When the SNV curve is near 1500 nm and 2000 nm, different absorption characteristics and curve trends are observed as the salt content is varied.

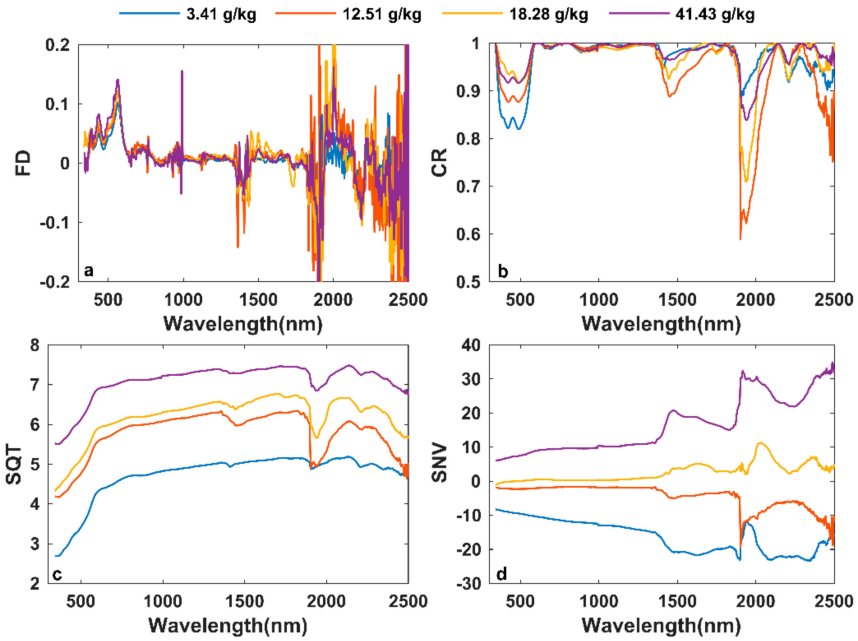

Figure 5. Various forms of spectral curves, namely, (**a**) first derivative (FD), (**b**) continuum removal (CR), (**c**) square root (SQT), and (**d**) standard normal variate (SNV).

To evaluate the relationship between the band and the soil salinity, various forms of spectra were correlated with the soil salt content (Figure 6). The results shown in Figure 6 are those that pass the significance test. The correlations between RS, SQT, and SNV and the soil salt content are strong in the wavelength range of 350–1400 nm, with correlation coefficients that exceed 0.6, and the correlation between the band in the range of 1400–1900 nm and soil salt content is from 0.4–0.6. The bands in which FD is highly correlated with the soil salinity are mainly concentrated from 1400 to 1800 nm and from 2000 to 2100 nm. The bands in which CR and FD are highly correlated with the soil salinity are similar and are mainly concentrated from 400 to 600 nm, from 1300 to 1800 nm, and from 1900 to 2100 nm.

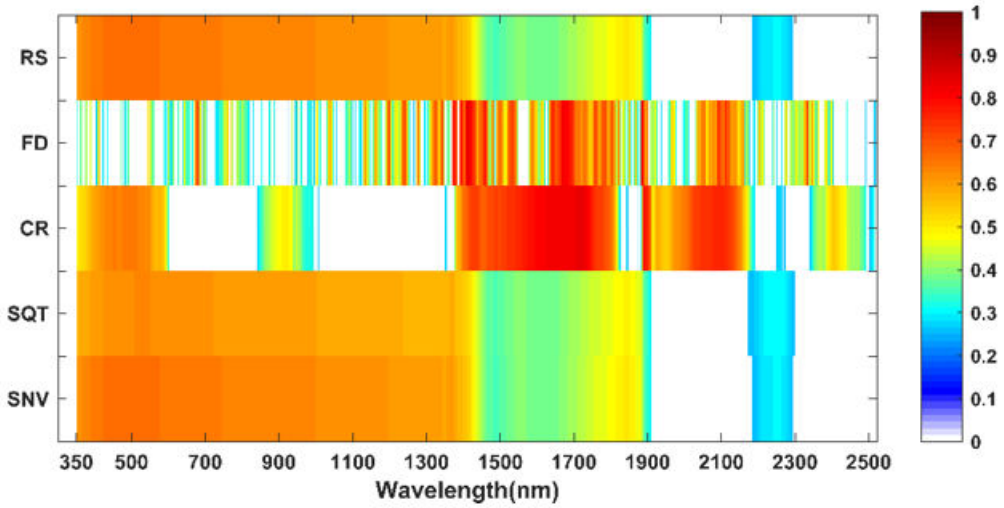

Figure 6. Correlations between various bands and the soil salt content.

To evaluate the relationship among the bands with high correlations with the soil salt content, the curves of the various bands were subjected to autocorrelation analysis (Figure 7a–e). RS, SQT, and SNV have high autocorrelations in the ranges of 300–1900 nm and 2100–2300 nm and the correlation coefficients exceed 0.8. The correlation coefficients of FD between 350 and 500 nm, between 1100 and

1900 nm, and between 2000 and 2200 nm exceed 0.8. The correlation coefficients of CR between 350 and 600 nm, between 1000 and 1400 nm, and between 1400 and 2200 nm exceed 0.8. By comparison, there is also a strong correlation among the bands with high correlations with the soil salt content.

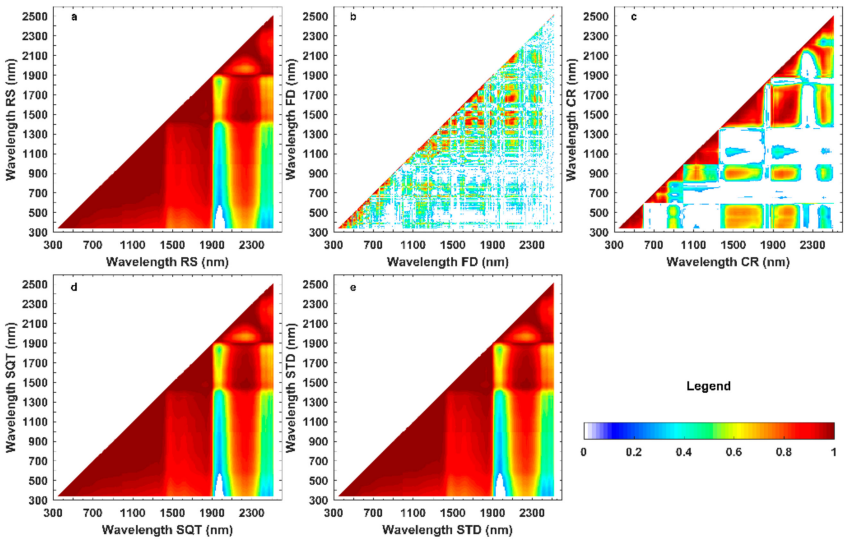

Figure 7. Correlation analysis. (**a**) presents the autocorrelation plot of RS, (**b**) presents the autocorrelation plot of FD, (**c**) presents the autocorrelation plot of CR, (**d**) presents the autocorrelation plot of SQT, and (**e**) presents the autocorrelation plot of SNV. The results of a 0.01 significance test are also presented in the figure.

Therefore, based on the correlation analysis results in this study, some characteristic wavelengths were selected for modeling while avoiding water absorption characteristics. The bands selected by RS are 455–554 nm, the bands selected by FD are 1372–1397 nm, 1640–1697 nm, 2090–2113 nm, the bands selected by CR are 1635–1734 nm, the bands selected by SQT are 464–563 nm, and the bands selected by SNV are 455–554 nm.

4.2. Model Parameter Determination

After considering the characteristics of each model and the characteristics of the data and after conducting repeated tests, the main parameters of each model are determined. The parameters are listed in Table 3.

Table 3. Parameters set by each model.

Model	Parameters
RFR	n_estimators=2500,min_samples_split=13,max_features="sqrt",oob_score=True
SVR	kernel='linear', C=15
GBRT	n_estimators=500, learning_rate=0.1, subsample=0.5, min_samples_split=20, max_depth=4, random_state=0, loss='huber'
MLPR	hidden_layer_sizes=(100,),activation='tanh', random_state=1
Lars	n_nonzero_coefs=1

4.3. Analysis on Collinear Problems

The results of the regression analysis of each model with five spectral forms according to various bands are presented in Figures 8 and 9. Figure 8 plots R^2 and RPD between the model estimated value and the measured value. Figure 9 plots RMSE and MAPE between the estimated and measured results. By comparison, it is found that the RFR model performs the best on the collinear problem, except that the regression effect of the FD is abrupt when the input band is 20, whereas the regression results of

all other variations do not exhibit sharp changes with the increase of the number of bands. The RFR model is followed by Lars, for which slight fluctuations (in RS and FD) are observed among input bands 5-30. When the number of bands exceeds 30, the regression performance is slightly reduced. Lars is followed by GBRT, which is similar to Lars, except that there are more spectral forms that exhibit fluctuations. In SVR, when the number of input RS bands is greater than 10, the regression result suddenly drops sharply. In MLPR, the regression results in RS, FD, and SQT fluctuated drastically with the number of bands. Therefore, the order of the models from strongest to weakest performance on collinear problems is RFR > Lars > GBRT > SVR > MLPR.

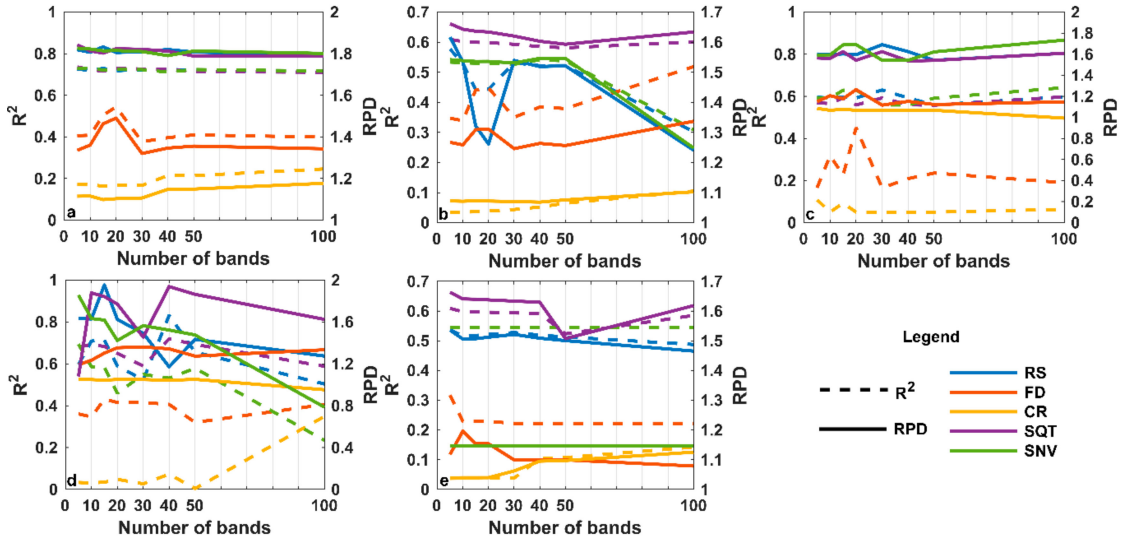

Figure 8. R^2 and ratio of performance to deviation (RPD) for each model regression result. (**a**) corresponds to random forest regression (RFR), (**b**) to support vector regression (SVR), (**c**) to gradient-boosted regression tree (GBRT), (**d**) to multilayer perceptron regression (MLPR), and (**e**) to Lars.

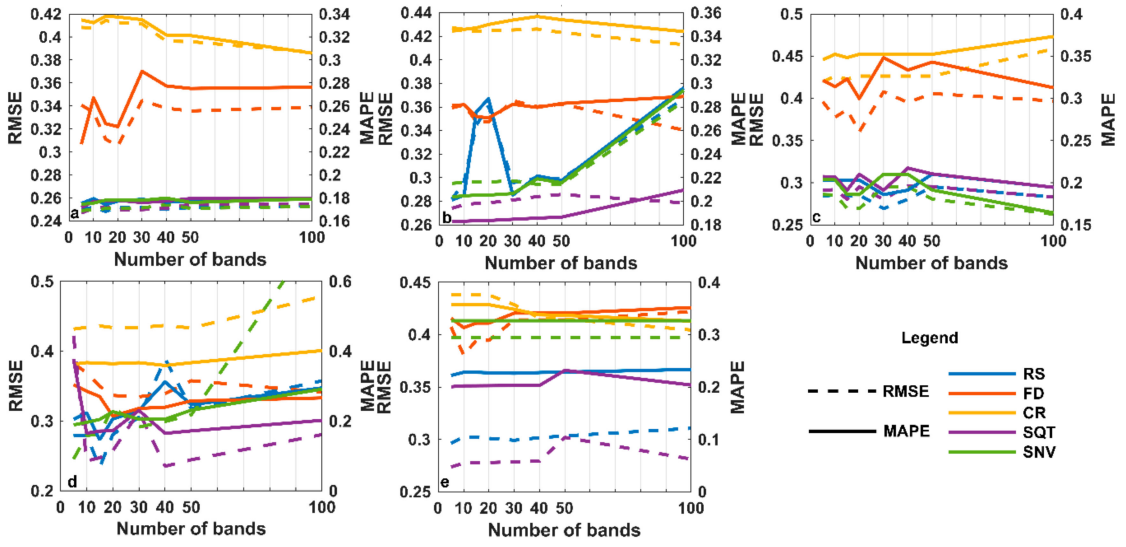

Figure 9. Root mean square error (RMSE) and MAPE for each model regression result. (**a**) corresponds to RFR, (**b**) to SVR, (**c**) to GBRT, (**d**) to MLPR, and (**e**) to Lars.

4.4. Noise Processing Performance Analysis

The regression results in terms of R^2, RPD, RMSE, and MAPE of each model as the noise is gradually decreased, namely, as the signal-to-noise ratio is increased, are plotted (Figure 10). When

the signal-to-noise ratio is less than 20, the fluctuation of the estimation results of GBRT is the largest; hence, the regression of this model is not highly robust to noise. GBRT is followed by SVR and Lars, which exhibit slight fluctuations in the signal-to-noise ratio. On noisy problem, models RFR and MLPR are the most stable. Therefore, the order of the models from the strongest to weakest performance on noisy problems is MLPR > RFR > Lars > SVR > GBRT.

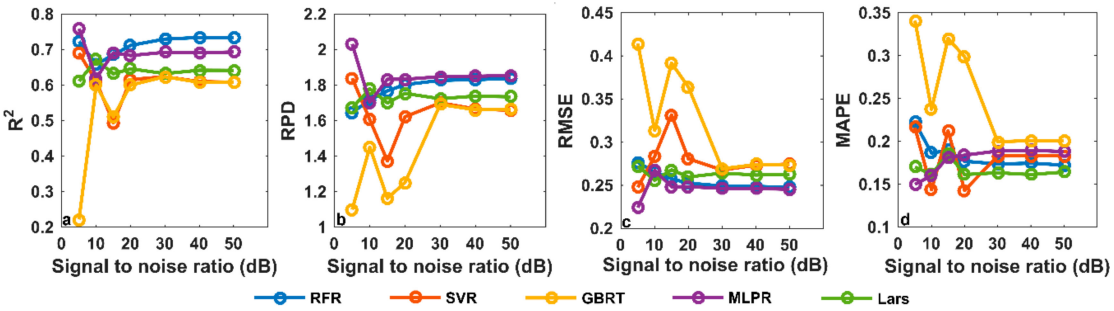

Figure 10. Results of each model on processing noise. (**a**) represents R^2, (**b**) represents RPD, (**c**) represents RMSE, and (**d**) represents MAPE.

4.5. Stability Analysis

When the training data and the test data change during modelling, the results of each model are presented in Figure 11. For quantitatively analyzing the changes in the estimation results of each model, the statistics of the changes in the indicators of the model data are presented in Table 4. When the training data are 3/4 of the total data, the smallest change occurs with the RFR model. When the test data are 2/4 and 1/4 of the total data, the smallest change occurs with the SVR model. According of the magnitude of the change, the SVC model is the most stable, followed by the MLPR, RFR, and Lars models. The GBRT model is the least stable. The order of the models from highest to lowest stability under three changes in the data volume is as follows: SVR > MLPR > RFR > Lars > GBRT.

Table 4. Changes in the regression results of each model.

Data	Δ	RFR	SVC	GBRT	MLPR	Lars
3/4	ΔRMSE	0.028	0.049	0.037	0.154	0.101
	ΔMAPE	0.045	0.107	0.074	0.111	0.111
	ΔRPD	0.099	0.008	0.054	1.108	0.238
	ΔR^2	0.050	0.096	0.043	0.021	0.102
2/4	ΔRMSE	0.107	0.073	0.202	0.071	0.079
	ΔMAPE	0.143	0.135	0.191	0.028	0.116
	ΔRPD	0.462	0.250	0.682	0.369	0.277
	ΔR^2	0.177	0.071	0.437	0.207	0.089
1/4	ΔRMSE	0.282	0.085	0.265	0.026	0.111
	ΔMAPE	0.196	0.090	0.203	0.039	0.077
	ΔRPD	0.894	0.265	0.783	0.039	0.361
	ΔR^2	0.296	0.082	0.642	0.502	0.166

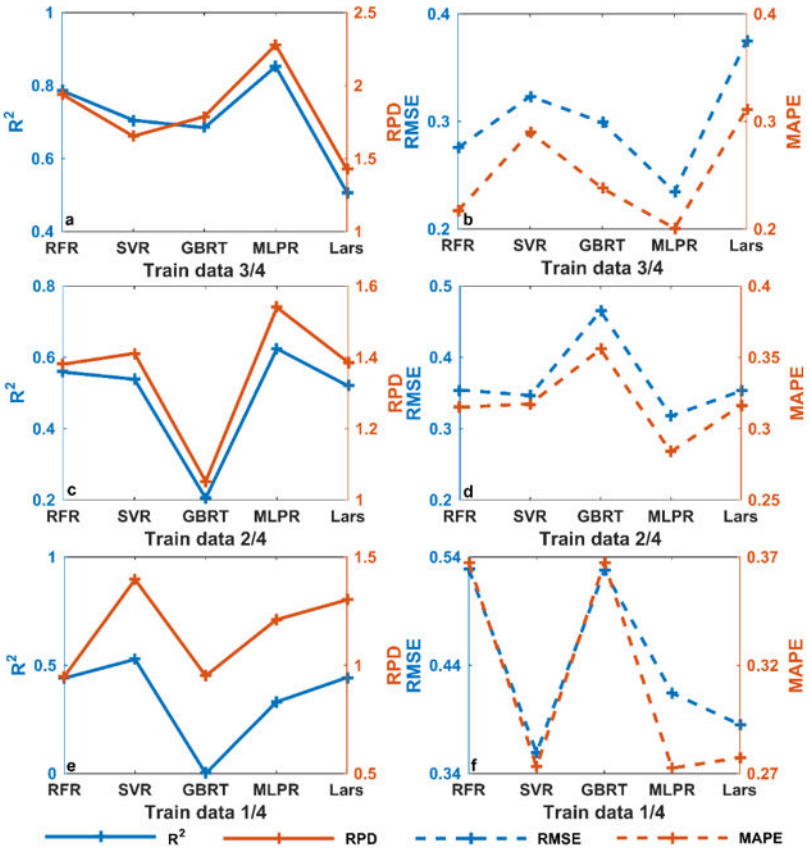

Figure 11. Regression results for each model with various amounts of data. (**a**) and (**b**) show the results of each model when the training data is 3/4 of the total data. (**c**) and (**d**) show the results of each model when the training data is 2/4 of the total data. (**e**) and (**f**) show the results of each model when the training data is 1/4 of the total data.

4.6. Precision Analysis

The six models were evaluated using the leave-one-out cross-validation method. The results are presented in Figure 12. The Lars model yields the best prediction result. The obtained R^2 is 0.87, the RPD is 2.67, the RMSE is 0.18, and the MAPE is 0.11. The Lars model is followed by the RFR, SVR, and MLPR models. The obtained results are both greater than 0.8. GBRT yield the worst modelling results. The order of the models from highest to lowest precision is as follows: Lars > SVR>RFR > MLPR > GBRT.

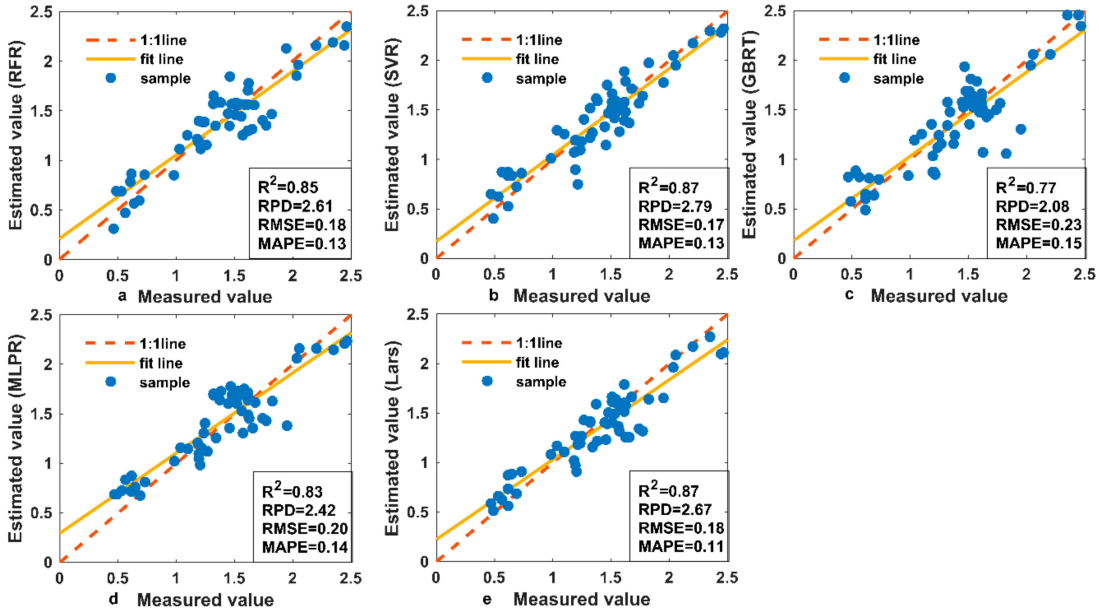

Figure 12. Cross-validation results graph for each model. (**a**) represents RFR, (**b**) represents SVR, (**c**) represents GBRT, (**d**) represents MLPR, and (**e**) represents Lars.

5. Discussion

5.1. Band Analysis

In this study, some bands were selected according to the sequence from high to low according to the correlation coefficients based on the correlation analysis results while removing the water absorption band. The selected bands mainly focus on 455–563 nm, 1372–1397 nm, 1640–1734 nm, and 2090–2113 nm. The selected bands are compared with the previous studies (Figure 13). Csillag et al. (1993) studied the spectral reflectivity of surface soil in the spectral range of 495–2395 nm, and found that the key bands for identifying salt-affected soil were mainly 550–770 nm, 900–1030 nm, 1270–1520 nm, 1940–2150 nm, 2150–2310 nm, and 2330–2400 nm [65]. These bands are similar to the bands used in this study: 455–563 nm, 1372–1397 nm, and 2090–2113 nm. Dehaan et al. (2002) pointed out in the study on the field-derived spectra of salinized soils that the absorption features of spectra were 505, 920, 1415, 1915, and 2205 nm [12]. The absorption features of the spectrum at 505 nm are in the bands used in this study. Zhou et al. (2006) used laboratory hyperspectral data to estimate the physicochemical of reclaimed saline soils and selected six spectral band, namely 448, 530, 670, 880, 1400, and 1900 nm, to discriminate the four saline land and groups [66]. Among them, 448 nm, 530 nm, and 1400 nm are similar to the bands used in this study. Weng et al. (2008) used reflectance spectroscopy to estimate the soil salt content in soils, and pointed out that the reflectance at 1931–2123 nm and 2153–2254 nm was highly correlated with soil salt content [14]. It is the same as the 2090–2113 nm band used in this study according to the correlation analysis results. Wang et al. (2012) modeling and inversion of the effect of salinity on soil reflectance under various moisture conditions were carried out in the laboratory, and some sensitive bands of salt types were obtained. Sensitive band for Na_2SO_4 type of salt affected soils were identified as from 1920–2230 nm, and 1970–2450 nm for NaCl, 350–400 nm for Na_2CO_3 [67]. Sidike et al. (2014) used image and spectra to estimate the soil salinity, and the statistical analysis showed that the sensitive bands of soil salinity were 350–436 nm, 516–814 nm, 1445–1506 nm, 1667–1699 nm, 1882–2096 nm, and 2160–2393 nm [68]. Srivastava et al. (2017) used the visible-near infrared reflectance spectroscopy to rapidly identify salt-affected soil and pointed out that the spectral range of 1390–2400 nm was highly sensitive to salinization [69]. The bands used in this study are similar to those in previous studies. Though, this study chooses the bands that cannot fully cover the soil salinity sensitive bands used in previous study, by comparison, this study selects the bands that

also can represent the features of the soil salinity. At the same time, through the regression results also can indicated, the bands used in this study can be used to estimate the soil salt content.

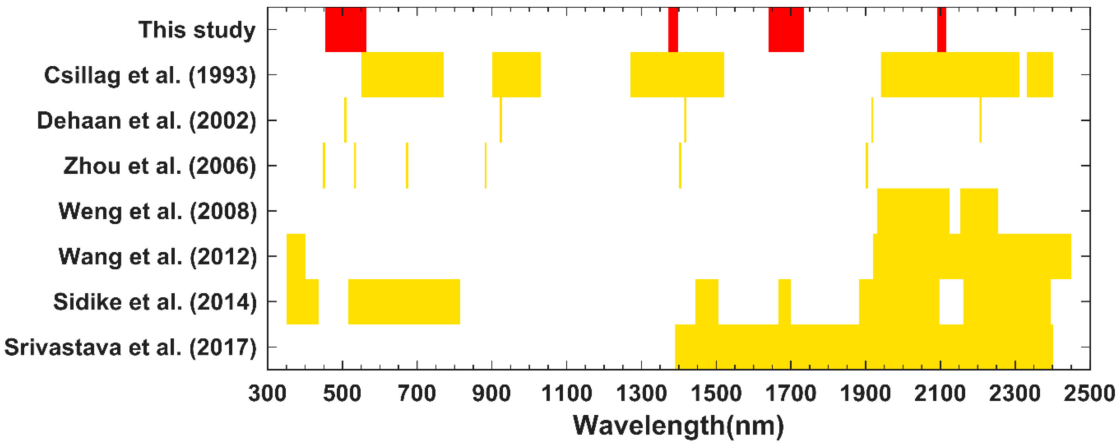

Figure 13. A comparison between the bands used in this study and those used in previous studies. The red represents the bands used in this study. The yellow represents the bands used in previous studies.

5.2. Comprehensive Comparison of Models

According to the comparative analysis, the RFR model performs the best on collinear problems and it also realizes satisfactory performance in terms of noise processing performance, model stability, and model accuracy. The SVR model has the highest stability. GBRT is the best at dealing with collinearity in terms of four aspects, but it is poor at dealing with data noise, model stability, and model accuracy. The MLPR model performs the best on noisy data. The OLS model performs the worst on collinear problems and its accuracy is the lowest. The Lars model is the most accurate. To comprehensively consider the performance of each model, each performance is assigned to one of the six levels: the best performance is assigned to 6 and the worst to 1. Then, the scores of each model are summed to comprehensively evaluate the performance of each model (Table 5). The overall performance of each model in terms of the four aspects is presented in Figure 14. The RFR model has the best comprehensive performance, followed by SVR, MLPR, and Lars, and GBRT and OLS perform the worst.

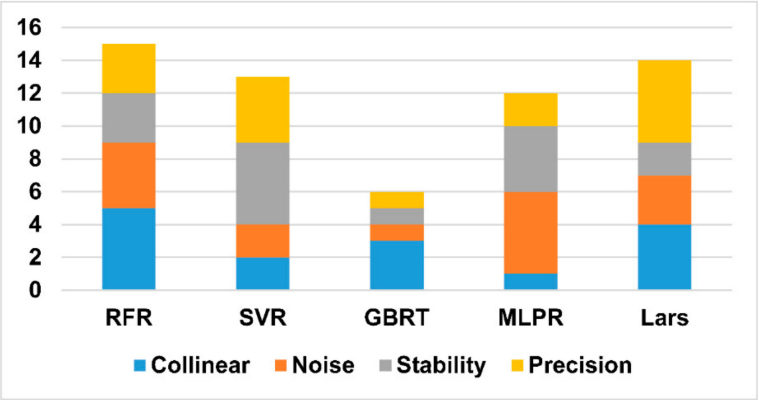

Figure 14. Performance comparison among the models.

Table 5. Performance comparison of each model.

Model	Collinear	Noise	Stability	Precision	Comprehensive
RFR	5	4	3	3	15
SVR	2	2	5	4	13
GBRT	3	1	1	1	6
MLPR	1	5	4	2	12
Lars	4	3	2	5	14

5.3. The Best Precision Model—Lars Model Depth Analysis

An in-depth analysis was conducted on the most accurate model, namely, Lars. First, the ability of the model on co-linear problems is explored. The number of bands input into the model is incremented from 1 to 100 in steps of 1 unit. By analyzing the regression results of each model, it is found that the regression results of Lars model change little with the number of bands. It shows that this model can process high-dimensional data, and the number of input bands has little influence on the results of the model (Figure 15).

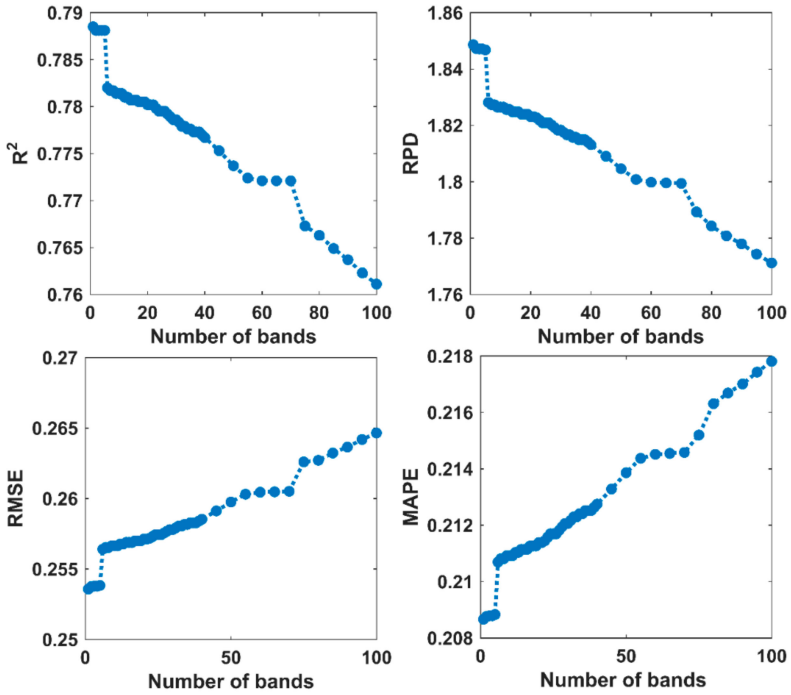

Figure 15. Explore Lars' ability to deal with collinearity. The regression results of the model are compared when the input band is gradually increased by 1 unit.

Then, we explore the limitations of the model in dealing with noisy problems. In the signal-to-noise ratio range of 1–50, the performance of the model on noisy problems is explored with a step size of 1. The results are presented in Figure 16. After the analysis, it is found that when the signal-to-noise ratio is less than 23, the Lars model results in severe fluctuations. When the signal-to-noise ratio is greater than 23, the results of the Lars model tend to be stable. Therefore, for the Lars model, when the signal-to-noise ratio of the input data is greater than 23, the results of the model are stable and acceptable.

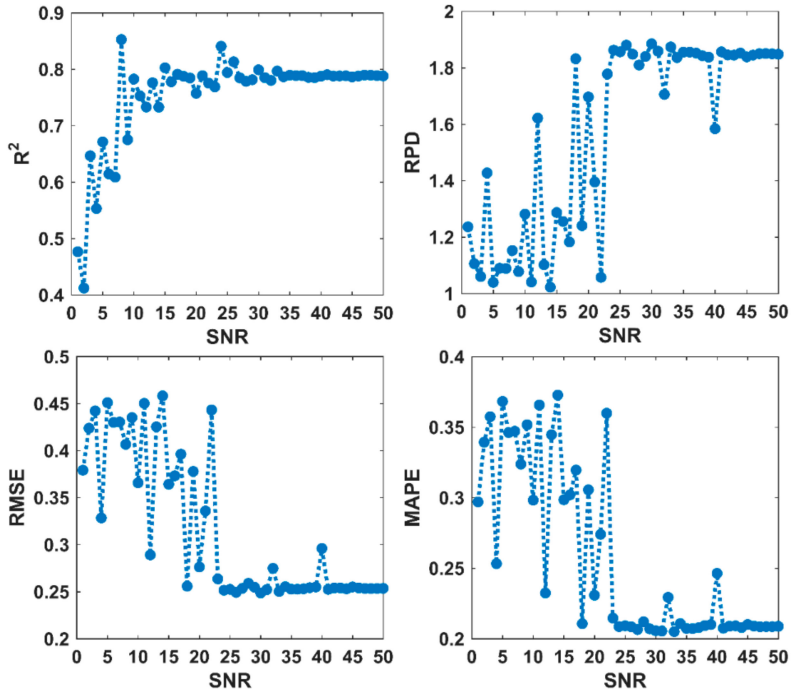

Figure 16. Results of the Lars model on noisy problems. In the signal-to-noise ratio range of 1–50, the analysis is performed with a step size of 1.

5.4. Comparison of Traditional Methods and Machine Learning Methods

The above research shows that most machine learning methods show good performance in soil salinization estimation. However, in the study, if the traditional regression method can show good performance, it is not necessary to choose complex machine learning method. Therefore, a comparison needs to be made between the traditional methods and machine learning methods. OLS and PLSR models were selected for comparison in the traditional method, while RFR model with the best comprehensive performance and Lars model with the highest accuracy were selected for comparison in the machine learning method.

The four models were compared from four aspects (Figure 17). In dealing with the collinearity problem, the regression results of the OLS model decreased sharply with the increase of the number of bands, indicating that the model cannot cope with the collinear phenomenon in the data. With the increase of the number of bands in the PLSR model, the regression results of the model are stable with slight fluctuations. In dealing with data noise, the regression results of the OLS model and the PLSR model vary greatly with the change of the data signal-to-noise ratio, while the machine learning model is more stable, explaining that these two traditional methods are not as good at processing data noise as machine learning methods. In terms of model stability, the four models showed similar stability. In terms of model accuracy, machine learning model is obviously better than traditional methods. Ma et al. (2019) also compared the accuracy of the machine learning method with the traditional method in their study, and reached the same conclusion, that is, the accuracy of the machine learning algorithm is better than the traditional method [70]. Therefore, by comparison, machine learning methods are more suitable for soil salinization research than traditional methods.

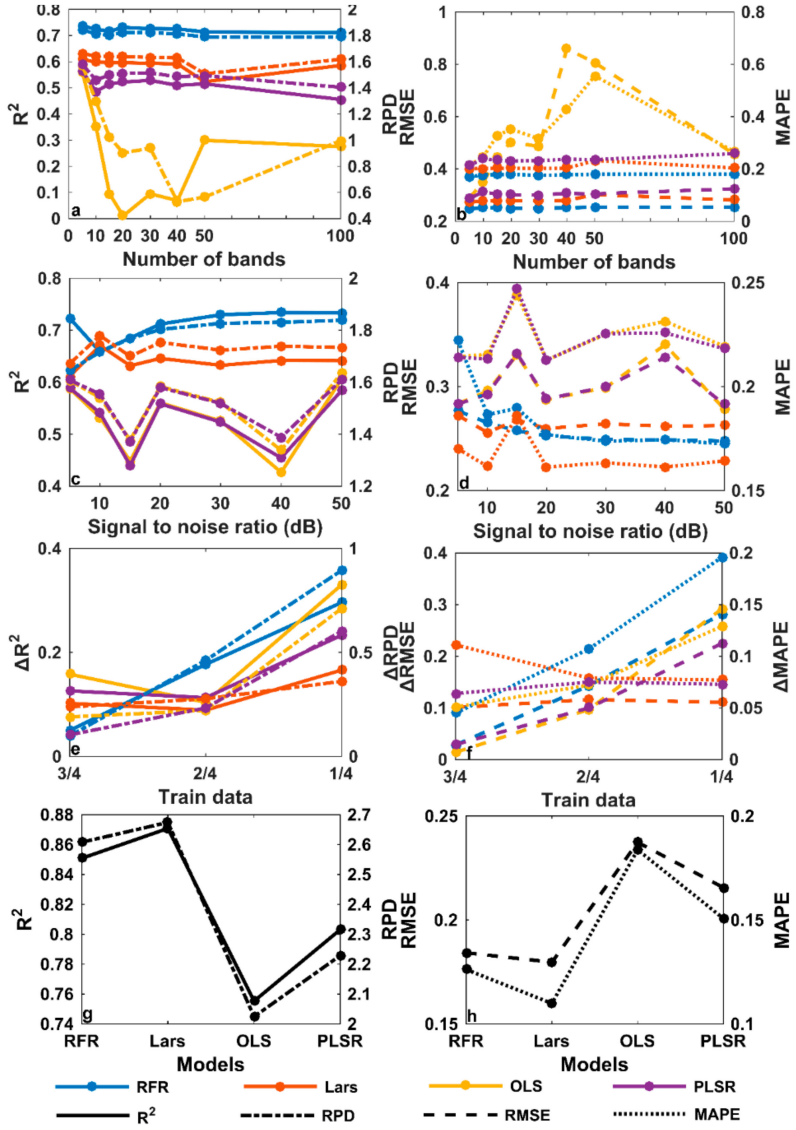

Figure 17. The results of the four models are compared from four aspects. (**a**) and (**b**) represent the performance of the four models in dealing with collinearity; (**c**) and (**d**) represent the performance of the four models to deal with data noise; (**e**) and (**f**) represent the stability of the four models. The figure shows the variation value of model regression results on the four evaluation indicators when the amount of training data changes. (**g**) and (**h**) represent the accuracy of the four models.

5.5. Spectral Difference Analysis Before and After Soil Disturbance

Field measurements of soil spectra were used in the study, and the soil was undisturbed at the time of the measurements. But when the soil was chemically analyzed, it was disturbed, causing the spectrometer to "see" a different salinity than the chemical analysis. Thus, spectral differences between disturbed and undisturbed soils were investigated. Before the soil samples collected from the field were sent to the laboratory for analysis, we air-dried the soil samples, removed the gravel, dead branches, and passed them through a 1-mm sieve, and measured the laboratory spectra of each group of samples in a dark room. We compared the field spectra with laboratory spectra (Figure 18).

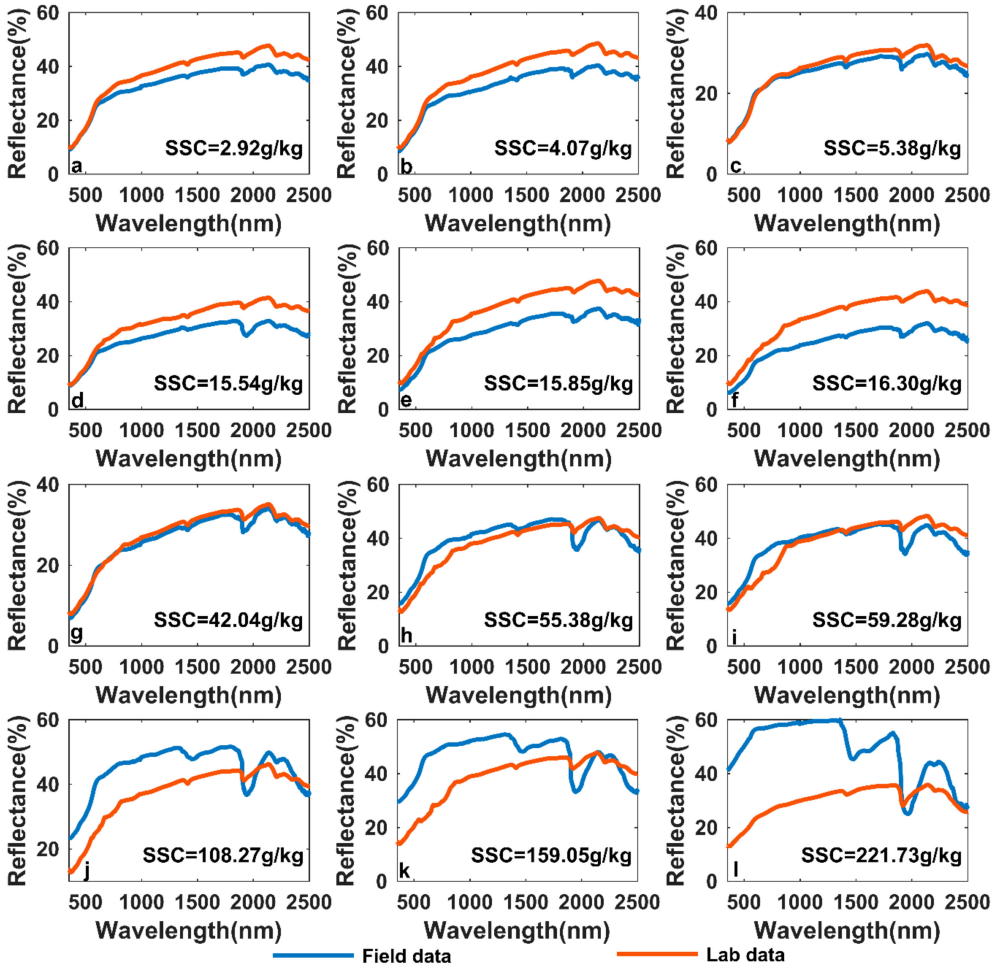

Figure 18. Field spectra versus laboratory spectra. SSC stands for soil salt content. The figure shows the changes between field and laboratory spectra as soil salt content increases. (**a**) shows the spectra for soil salt content of 2.92 g/kg. (**b**) shows the spectra for soil salt content of 4.07 g/kg. (**c**) shows the spectra for soil salt content of 5.38 g/kg. (**d**) shows the spectra for soil salt content of 15.54 g/kg. (**e**) shows the spectra for soil salt content of 15.85 g/kg. (**f**) shows the spectra for soil salt content of 16.30 g/kg. (**g**) shows the spectra for soil salt content of 42.04 g/kg. (**h**) shows the spectra for soil salt content of 55.38 g/kg. (**i**) shows the spectra for soil salt content of 59.28 g/kg. (**j**) shows the spectra for soil salt content of 108.27 g/kg. (**k**) shows the spectra for soil salt content of 159.05 g/kg. (**i**) shows the spectra for soil salt content of 221.73 g/kg.

When the soil salt content is low, the change trend of field spectrum curve is consistent with that of laboratory spectrum curve. Only when the wavelength is greater than 800 nm, the reflectivity of the laboratory spectrum is higher than that of the field spectrum. This is due to the fact that the soil measured in the field contains moisture, which makes the spectral reflectivity of the soil decreased as a whole. With the increase of soil salt content, soil surface crust phenomenon is serious and the field spectral reflectivity gradually increases. At 1900 nm, there is a strong water absorption valley in the field spectrum. In this study, the selected wavelength was mainly concentrated at 400–600 nm, within which the spectral difference before and after soil disturbance was small and had little influence on the research results. However, the spectra before and after the soil disturbance are different to a certain extent. The existence of such difference needs to be considered in the subsequent studies, and the reasons for such difference can be analyzed in depth.

6. Conclusions

This paper comprehensively analyses the performances of five machine learning models (RFR, SVR, GBRT, MLPR, and Lars) in estimating the soil salinity using field-measured spectral data. In dealing with collinearity problem, RFR model is the best, which has strong processing performance for high-dimensional data. The MLPR and RFR models perform well on data noise problems; thus, these two models can be considered if the data noise is high. The stability of the SVR models is satisfactory and changes in the amount of data have less of an impact on the model. The Lars model has the highest accuracy in estimating the soil salinity. When applying the Lars model for optimal modelling results, the signal-to-noise ratio of the input data should exceed 23.

Through comprehensive analysis of each model, the ranking of the model comprehensive performance from strong to weak is RFR>Lars>SVR>MLPR>GBRT. Through comparison it is found that the comprehensive performance of machine learning model is superior to the traditional regression learning method. Therefore, RFR model can be used as the preferred model for subsequent studies on hyperspectral estimation of soil salt content.

Author Contributions: Data curation, S.W., Y.C., M.W. and J.L.; Funding acquisition, Y.C. and J.L.; Methodology, S.W.; Visualization, S.W.; Writing—original draft, S.W.; Writing—review and editing, S.W. and Y.C.

Funding: This research was funded by The Beijing Key Laboratory of Environmental Remote Sensing and Digital City. Ningxia Agriculture, and Animal Husbandry Department East-West Cooperation Project "Study on Remote Sensing Monitoring of Agronomic Improvement of Saline-alkali Land in Yinbei Area," Ningxia Academy of Agriculture and Forestry Science and Technology Innovation Guide Project (NKYG-18-01). Natural Science Foundation of China (41571342,51579135). The Beijing Laboratory of Water Resources Security.

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

References

1. Zhu, J.K. Plant salt tolerance. *Trends Plant Sci.* **2001**, *6*, 66–71. [CrossRef]
2. Nurmemet, I.; Sagan, V.; Ding, J.L.; Halik, U.; Abliz, A.; Yakup, Z. A WFS-SVM Model for Soil Salinity Mapping in Keriya Oasis, Northwestern China Using Polarimetric Decomposition and Fully PolSAR Data. *Remote Sens.* **2018**, *10*, 598. [CrossRef]
3. Gorji, T.; Sertel, E.; Tanik, A. Monitoring soil salinity via remote sensing technology under data scarce conditions: A case study from Turkey. *Ecol. Indic.* **2017**, *74*, 384–391. [CrossRef]
4. Jiang, H.; Rusuli, Y.; Amuti, T.; He, Q. Quantitative assessment of soil salinity using multi-source remote sensing data based on the support vector machine and artificial neural network. *Int. J. Remote Sens.* **2019**, *40*, 284–306. [CrossRef]
5. Zhang, J.F.; Li, W.Y.; Hu, H.; Chen, W.Z.; Wang, X.L. Current status and prospects of research on saline-alkali land improvement. *Jiangsu Agric. Sci.* **2007**, *45*, 7–10.
6. Fan, X.W.; Liu, Y.B.; Tao, J.M.; Weng, Y.L. Soil Salinity Retrieval from Advanced Multi-Spectral Sensor with Partial Least Square Regression. *Remote Sens.* **2015**, *7*, 488–511. [CrossRef]
7. Abbas, A.; Khan, S.; Hussain, N.; Hanjra, M.A.; Akbar, S. Characterizing soil salinity in irrigated agriculture using a remote sensing approach. *Phys. Chem. Earth* **2013**, *55*, 43–52. [CrossRef]
8. Nawar, S.; Buddenbaum, H.; Hill, J.; Kozak, K. Modeling and Mapping of Soil Salinity with Reflectance Spectroscopy and Landsat Data Using Two Quantitative Methods (PLSR and MARS). *Remote Sens.* **2014**, *6*, 10813–10834. [CrossRef]
9. Dehni, A.; Lounis, M. Remote Sensing Techniques for Salt Affected Soil Mapping: Application to the Oran Region of Algeria. *Procedia Eng.* **2012**, *33*, 188–198. [CrossRef]
10. Verma, K.S.; Saxena, R.K.; Barthwal, A.K.; Deshmukh, S.N. Remote sensing technique for mapping salt affected soils. *Int. J. Remote Sens.* **1994**, *15*, 1901–1914. [CrossRef]
11. Zhang, T.T.; Zeng, S.L.; Gao, Y.; Ouyang, Z.T.; Li, B.; Fang, C.M.; Zhang, B. Using hyperspectral vegetation indices as a proxy to monitor soil salinity. *Ecol. Indic.* **2010**, *11*, 1552–1562. [CrossRef]
12. Dehaan, R.L.; Taylor, G.R. Field-derived spectra of salinized soils and vegetation as indicators of irrigation-induced soil salinization. *Remote Sens. Environ.* **2002**, *80*, 406–417. [CrossRef]

13. An, D.Y.; Zhao, G.X.; Chang, C.Y.; Wang, Z.R.; Li, P.; Zhang, T.R.; Jia, J.C. Hyperspectral field estimation and remote-sensing inversion of salt content in coastal saline soils of the Yellow River Delta. *J. Remote Sens.* **2016**, *37*, 455–470. [CrossRef]
14. Weng, Y.L.; Gong, P.; Zhu, Z.L. Reflectance spectroscopy for the assessment of soil salt content in soils of the Yellow River Delta of China. *J. Remote Sens.* **2008**, *29*, 5511–5531. [CrossRef]
15. Scudiero, E.; Skaggs, T.H.; Corwin, D.L. Regional-scale soil salinity assessment using Landsat ETM + canopy reflectance. *Remote Sens. Environ.* **2015**, *169*, 335–343. [CrossRef]
16. Aldabaa, A.A.A.; Weindorf, D.C.; Chakraborty, S.; Sharma, A.; Li, B. Combination of proximal and remote sensing methods for rapid soil salinity quantification. *Geoderma* **2015**, *239*, 34–46. [CrossRef]
17. Bouaziz, M.; Matschullat, J.; Gloaguen, R. Improved remote sensing detection of soil salinity from a semi-arid climate in Northeast Brazil. *C. R. Geosci.* **2011**, *343*, 795–803. [CrossRef]
18. Dutkiewicz, A.; Lewis, M.; Ostendorf, B. Evaluation and comparison of hyperspectral imagery for mapping surface symptoms of dryland salinity. *J. Remote Sens.* **2009**, *30*, 693–719. [CrossRef]
19. Ghosh, G.; Kumar, S.; Saha, S.K. Hyperspectral Satellite Data in Mapping Salt-Affected Soils Using Linear Spectral Unmixing Analysis. *J. Indian Soc. Remote Sens.* **2012**, *40*, 129–136. [CrossRef]
20. Mashimbye, Z.E.; Cho, M.A.; Nell, J.P.; De Clercq, W.P.; Van Niekerk, A.; Turner, D.P. Model-Based Integrated Methods for Quantitative Estimation of Soil Salinity from Hyperspectral Remote Sensing Data: A Case Study of Selected South African Soils. *Pedosphere* **2012**, *22*, 640–649. [CrossRef]
21. Farifteh, J.; Der Meer, F.D.; Atzberger, C.; Carranza, E.J. Quantitative analysis of salt-affected soil reflectance spectra: A comparison of two adaptive methods (PLSR and ANN). *Remote Sens. Environ.* **2007**, *110*, 59–78. [CrossRef]
22. Abliz, A.; Tiyip, T.; Ding, J.; Sawut, M.; Hou, Y.J.; Nurmemet, I. Estimating Soil Salt Content in the Keriya Oasis Using Hyperspectral Slope Index. *Nat. Enviro. Pollut. Technol.* **2017**, *16*, 141–146.
23. Eldiery, A.; Garcia, R.M.; Reich, R.M. Estimating Soil Salinity from Remote Sensing Data in Corn Fields. *Hydrology* **2005**, *8*, 31–42.
24. Marabel, M.; Alvareztaboada, F. Spectroscopic Determination of Aboveground Biomass in Grasslands Using Spectral Transformations, Support Vector Machine and Partial Least Squares Regression. *Sensors* **2013**, *13*, 10027–10051. [CrossRef] [PubMed]
25. Eldeiry, A.A.; Garcia, L.A. Detecting Soil Salinity in Alfalfa Fields using Spatial Modeling and Remote Sensing. *Soil Sci. Soc. Am. J.* **2008**, *72*, 201–211. [CrossRef]
26. Gomez, C.; Rossel, R.A.; Mcbratney, A.B. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy an Australian case study. *Geoderma* **2008**, *146*, 403–411. [CrossRef]
27. Hansen, P.M.; Schjoerring, J.K. Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. *Remote Sens. Environ.* **2003**, *86*, 542–553. [CrossRef]
28. Shi, Z.; Ji, W.; Rossel, R.A.; Chen, S.; Zhou, Y. Prediction of soil organic matter using a spatially constrained local partial least squares regression and the Chinese vis–NIR spectral library. *Eur. J. Soil Sci.* **2015**, *66*, 679–687. [CrossRef]
29. Peng, J.; Biswas, A.; Jiang, Q.S.; Zhao, R.Y.; Hu, J.; Hu, B.F.; Shi, Z. Estimating soil salinity from remote sensing and terrain data in southern Xinjiang Province, China. *Geoderma* **2019**, *337*, 1309–1319. [CrossRef]
30. Alpaydin, E. *Introduction to Machine Learning*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2004; pp. 1–10.
31. Ahmad, S.; Kalra, A.; Stephen, H. Estimating soil moisture using remote sensing data: A machine learning approach. *Adv. Water Resour.* **2010**, *33*, 69–80. [CrossRef]
32. Hladik, C.M.; Schalles, J.F.; Alber, M. Salt Marsh Elevation and Habitat Mapping Using Hyperspectral and LIDAR Data. *Remote Sens. Environ.* **2013**, *139*, 318–330. [CrossRef]
33. Pang, G.J.; Wang, T.; Liao, J.; Li, S. Quantitative Model Based on Field-Derived Spectral Characteristics to Estimate Soil Salinity in Minqin County, China. *Soil Sci. Soc. Am. J.* **2014**, *78*, 546–555. [CrossRef]
34. Wang, J.Z.; Ding, J.L.; Abulimiti, A.; Cai, L.H. Quantitative estimation of soil salinity by means of different modeling methods and visible-near infrared (VIS–NIR) spectroscopy, Ebinur Lake Wetland, Northwest China. *PeerJ* **2018**, *6*, e4703. [CrossRef] [PubMed]
35. Wu, J.B.; Wang, X.Y. A target level detection method for saline land change in Shizuishan area, Ningxia. *J. Shanxi Norm. Univ. (Nat. Sci. Ed.)* **2018**, *46*, 104–109.

36. Shui, Y.; Xu, Z.H.; Liu, G.F. Effect of depth on bacterial diversity in saline-alkali soil in Shizuishan region in Ningxia. *Acta Ecol. Sin.* **2019**, *39*, 3597–3606.
37. Wang, S.J.; Chen, Y.H.; Wang, M.G.; Zhao, Y.F.; Li, J. SPA-based methods for the quantitative estimation of the soil salt content in saline-alkali land from field spectroscopy data: A case study from the Yellow River irrigation regions. *Remote Sens.* **2019**, *11*, 967. [CrossRef]
38. Pelta, R.; Carmon, N.; Bendor, E. A machine learning approach to detect crude oil contamination in a real scenario using hyperspectral remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2019**, *82*, 101901. [CrossRef]
39. Scafutto, R.D.M.S.; Filho, C.R.D.S.; Rivard, B. Characterization of mineral substrates impregnated with crude oils using proximal infrared hyperspectral imaging. *Remote Sens. Environ.* **2016**, *179*, 116–130. [CrossRef]
40. Yu, L.; Hong, Y.S.; Zhu, Y.X.; Huang, P.; He, Q.; Qi, F. Removing the effect of soil moisture content on hyperspectral reflectance for the estimation of soil organic matter content. *Spectrosc. Spectr. Anal.* **2017**, *37*, 2146–2151.
41. He, Y.; Liu, F.; Li, X.L.; Shao, Y.N. *Spectroscopy and Imaging Technology in Agriculture*; Science Press: Beijing, China, 2016; pp. 106–107.
42. Barnes, R.J.; Dhanoa, M.S.; Lister, S.J. Standard normal variate transformation and de-trending of near-infared diffuse reflectance spectra. *Appl. Spectrosc.* **1989**, *43*, 772–777. [CrossRef]
43. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
44. What Is the Difference between Bagging and Boosting? Available online: https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/ (accessed on 10 April 2019).
45. Mutanga, O.; Adam, E.; Cho, M.A. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *Int. J. Appl. Earth Obs. Geoinf.* **2012**, *18*, 399–406. [CrossRef]
46. Rodriguezgaliano, V.F.; Sanchezcastillo, M.; Chicaolmo, M.; Chicarivas, M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geol. Rev.* **2015**, *71*, 804–818. [CrossRef]
47. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.J.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
48. Smola, A.J.; Scholkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [CrossRef]
49. Pasolli, L.; Notarnicola, C.; Bruzzone, L. Estimating Soil Moisture with the Support Vector Regression Technique. *IEEE Geosci. Remote Sens. Lett.* **2011**, *8*, 1080–1084. [CrossRef]
50. Liang, D.; Guang, Q.S.; Huang, W.J.; Huang, L.S.; Yang, G.J. Remote sensing inversion of leaf area index based on support vector machine regression in winter wheat. *Trans. Chin. Soc. Agric. Eng. (Trans. CSAE)* **2013**, *29*, 117–123.
51. Zhai, Y.F.; Cui, L.J.; Zhou, X.; Gao, Y.; Fei, T.; Gao, W.X. Estimation of nitrogen, phosphorus, and potassium contents in the leaves of different plants using laboratory-based visible and near-infrared reflectance spectroscopy: Comparison of partial least-square regression and support vector machine regression methods. *Int. J. Remote Sens.* **2013**, *34*, 2502–2518.
52. Friedman, J.H. Greedy function approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
53. Mohan, A.; Chen, Z.; Weinberger, K. Web-Search Ranking with Initialized Gradient Boosted Regression Tress. *JMLR Workshop Conf. Proc.* **2011**, *14*, 77–89.
54. Greg, R. Generalized Boosted Models: A Guide to the GBM Package. *Compute* **2005**, *1*, 1–12.
55. Tomassetti, B.; Verdecchia, M.; Giorgi, F. NN5: A neural network based approach for the downscaling of precipitation fields -Model description and preliminary results. *J. Hydrol.* **2009**, *367*, 14–26. [CrossRef]
56. Choubin, B.; Khalighisigaroodi, S.; Malekian, A.; Kisi, O. Multiple linear regression, multi-layer perceptron network and adaptive neuro-fuzzy inference system for forecasting precipitation based on large-scale climate signals. *Hydrol. Sci. J.-J. Des. Sci. Hydrol.* **2016**, *61*, 1001–1009. [CrossRef]
57. Hu, X.F.; Weng, Q.H. Estimating impervious surfaces from medium spatial resolution imagery using the self-organizing map and multi-layer perceptron neural networks. *Remote Sens. Environ.* **2009**, *113*, 2089–2102. [CrossRef]
58. Augugliaro, L.; Mineo, A.M.; Wit, E. Differential geometric least angle regression: A differential geometric approach to sparse generalized linear models. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **2013**, *75*, 471–498. [CrossRef]

59. Feng, R.Y.; Wang, L.Z.; Zhong, Y.F. Least Angle Regression-Based Constrained Sparse Unmixing of Hyperspectral Remote Sensing Imagery. *Remote Sens.* **2018**, *10*, 1546. [CrossRef]
60. Coxeter, H.S.M. A problem of collinear points. *Am. Math. Mon.* **1948**, *55*, 26–28. [CrossRef]
61. Zhao, Y.S. *Principles and Methods of Remote Sensing Application Analysis*, 2nd ed.; Science Press: Beijing, China, 2013; pp. 19–33.
62. Yuan, H.H.; Yang, G.J.; Li, C.C.; Wang, Y.J.; Liu, J.G.; Yu, H.Y.; Feng, H.K.; Xu, B.; Zhao, X.Q.; Yang, X.D. Retrieving Soybean Leaf Area ndex from Unmanned Aerial Vehicle Hyperspectral Remote Sensing: Analysis of RF, ANN, and SVM Regression Models. *Remote Sens.* **2017**, *9*, 309. [CrossRef]
63. Cawley, G.C.; Talbot, N.L. Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers. *Pattern Recognit.* **2003**, *36*, 2585–2592. [CrossRef]
64. Paulus, S.; Schumann, H.; Kuhlmann, H.; Leon, J. High-precision laser scanning system for capturing 3D plant architecture and analysing growth of cereal plants. *Biosyst. Eng.* **2014**, *121*, 1–11. [CrossRef]
65. Csillag, F.; Pasztor, L.; Biehl, L.L. Spectral band selection for the characterization of salinity status of soil. *Remote Sens. Environ.* **1993**, *43*, 231–242. [CrossRef]
66. Shi, Z.; Cheng, J.L.; Huang, M.X.; Zhou, L.Q. Assessing reclamation levels of coastal saline lands with integrated stepwise discriminant analysis and laboratory hyperspectral data. *Pedosphere* **2006**, *16*, 154–160. [CrossRef]
67. Wang, Q.; Li, P.H.; Chen, X. Modeling salinity effects on soil reflectance under various moisture conditions and its inverse application: A laboratory experiment. *Geoderma* **2012**, *170*, 103–111. [CrossRef]
68. Sidike, A.; Zhao, S.; Wen, Y. Estimating soil salinity in Pingluo County of China using Quickbird data and soil reflectance spectra. *Int. J. Appl. Earth Obs. Geoinf.* **2014**, *26*, 156–175. [CrossRef]
69. Srivastava, R.; Sethi, M.; Yadav, R.K.; Bundela, D.S.; Singh, M.; Chattaraj, S.; Singh, S.K.; Nasre, R.A.; Bishnoi, S.R.; Dhale, S.; et al. Visible-near infrared reflectance spectroscopy for rapid characterization of salt-affected soil in the Indo-Gangetic Plains of Haryana, India. *J. Indian Soc. Remote Sens.* **2017**, *45*, 307–315. [CrossRef]
70. Ma, M.H.; Liu, C.J.; Zhao, G.; Xie, H.J.; Jia, P.F.; Wang, D.C.; Wang, H.X.; Hong, Y. Flash Flood Risk Analysis Based on MachineLearning Techniques in the Yunnan Province, China. *Remote Sens.* **2019**, *11*, 170. [CrossRef]