Hindawi
Journal of Spectroscopy
Volume 2024, Article ID 8180765, 16 pages
https://doi.org/10.1155/2024/8180765

Research Article

Feature Variable Selection Based on VIS-NIR Spectra and Soil Moisture Content Prediction Model Construction

Nan Zhou,[1,2,3] **Jin Hong**,[1,2,3] **Bo Song**,[1,3] **Shichao Wu**,[1,3] **Yichen Wei**,[1,2,3] **and Tao Wang**[4]

[1]*Anhui Institute of Optics and Fine Mechanics, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China*
[2]*University of Science and Technology of China, Hefei 230026, China*
[3]*Key Laboratory of General Optical Calibration and Characterization Technology, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China*
[4]*Key Laboratory of Radiometric Calibration and Validation for Environmental Satellites, Haidian, Beijing 100081, China*

Correspondence should be addressed to Jin Hong; hongjin@aiofm.ac.cn

Received 6 September 2023; Revised 2 April 2024; Accepted 10 April 2024; Published 18 May 2024

Academic Editor: Daniel Cozzolino

The hydrological cycle, surface energy balance, and the management of water resources are all significantly impacted by soil moisture. Because it governs the physical processes of evapotranspiration and rainfall penetration, surface soil moisture is a significant climatic variable. In this work, visible-near infrared (VIS-NIR) bands were used to compare and analyze the spectra of loess samples with varying moisture concentrations. The investigation looked at how changes in the soil moisture content impacted the response of the soil spectra. The researchers used a genetic algorithm (GA), interval combination optimization (ICO), and competitive adaptive reweighted sampling (CARS) to filter feature variables from full-band spectral data. To forecast the moisture content of loess on the soil surface, models like partial least squares regression (PLSR), support vector machine (SVM), and random forest (RF) were created. The findings indicate that: (1) the most reliable spectrum preprocessing technique is the first derivative (FD), which can significantly enhance the model's prediction power and spectral characteristic information. (2) The feature band selection method's prediction effect of soil moisture content is typically superior to that of full-spectrum data. (3) The random forest (RF) prediction model for soil moisture content with the highest accuracy was built by combining the genetic algorithm (GA) with the FD preprocessed spectra. The results may provide a new understanding on how to use VIS-NIR to measure soil moisture content.

1. Introduction

Soil moisture content (SMC) plays a crucial role in determining the physical characteristics of soil [1, 2]. The soil moisture content impacts the physical and chemical processes within the soil, as well as the overall ecological environment, hydrology, and patterns of climate change [3]. Soil moisture content monitoring plays a crucial role in ensuring the protection of crop growth, mitigating geological calamities, and averting soil desertification [4, 5].

Remote sensing is widely regarded as a valuable technique for monitoring soil characteristics across extensive regions, proving to be a more economical alternative to in situ measurements [6]. Remote sensing relies on the physical correlation between the soil moisture content and the particular reflectance spectrum.

Traditional SMC monitoring is based on the drying method, neutron meter determination method, γ-ray method, etc. [7], which have high accuracy in single-point determination but require a large amount of human and

material resources and time resources, and are unable to provide continuous spatial surface information of SMC. Due to the characteristics of real-time, nondestructive, and noncontact, visible-near-infrared remote sensing [8] and microwave remote sensing [9] provide effective means for SMC monitoring. Hyperspectral data are rich in band information and can provide more detailed spectral information to reflect geophysical properties, but at the same time there are problems such as noise, high redundancy of information, and overlapping absorption features. Investigating whether selecting feature variables or sensitive bands from the original spectrum can replace the full band is highly important. This can lead to improved prediction accuracy, reduced model workload, and enhanced model efficiency [10]. Scholars have examined different algorithms to enhance the effectiveness and precision of soil property prediction through VIS-NIR spectroscopy by eliminating noise and extracting bands with distinctive features. The accuracy of the model is affected by the fact that various algorithms for selecting variables do not choose the same feature variables [11]. Many prior research studies have utilized linear models in conjunction with multiple variable selection techniques, whereas the utilization of support vector machine (SVM), random forest (RF), and other models in combination has been observed in a few less studies [12–14]. Combining different variable selection methods with regression methods to predict soil properties can provide a theoretical basis for remote sensing means to obtain information about the target object.

Spectral technology has improved models for specific soil properties. This greatly improves the accuracy and efficiency of soil moisture content detection, which significantly enhances the effectiveness of soil property evaluation. The current methods for establishing soil moisture detection models based on soil spectral features mainly include univariate linear models, multivariate linear models, and other nonlinear models. The linear model is efficient and simple, with high accuracy, while the nonlinear model is more complex but has better generalization performance. Appropriate models can accurately predict various soil parameters, such as soil organic matter [15], salinity [16], and texture [17], as indicated by previous research. Researchers have obtained favorable forecast outcomes by employing various models, including stepwise multiple linear regression [18], multivariate adaptive regression splines [19], memory-based learning [20], partial least squares regression (PLSR) [21], cubist [22], support vector machines (SVMs) [23], and random forests (RFs) [24, 25]. Nevertheless, there is a scarcity of research that has integrated diverse feature selection algorithms with various machine-learning models to identify SMCs. Hence, in this research, we merged three feature selection techniques (CARS, ICO, and GA) with three common machine-learning models (PLSR, SVM, and RF) to identify the best prediction model for SMC in the designated region.

2. Methods and Materials

2.1. Soil Sampling and Laboratory Sample Preparation. The research site is situated in the Qinzhou District of Tianshui City, in the southeastern part of Gansu Province. It is in a loess hilly area with complex geological structures and a poor geological environment. Moreover, it is part of the transition zone between the Qinling Mountains and the Longshan Mountains. Adjacent to the loess plateau, the area is situated within the latitude range of 34°05′–34°40′N and the longitude range of 105°13′–106°01′E, as depicted in Figure 1. The elevation ranges from 1,107 to 2,707 meters. Situated in the southern vicinity of the Wei River and to the west of the Jialing River, this region serves as the dividing line for the water systems of both the Yellow River and the Yangtze River [26]. In terms of climate, it falls within the warm temperate humid and semihumid climate region, experiencing an average yearly temperature ranging from 9 to 13 degrees Celsius and an annual rainfall of 420 to 660 mm. The distribution of annual rainfall is not uniform, with June through September making up 81.6% of the total precipitation. The terrain consists primarily of mountainous brown soil, while the plant life predominantly consists of warm temperate mixed coniferous broad-leaved forest [27].

In July of 2022, a total of five sampling locations were chosen at random within the designated region to collect undisturbed soil samples from a depth of 0-20 cm. Each black dot in Figure 1 represents a sampling area. The sampling technique used in each sampling area was the "five-point" method [28, 29]. In our study, we considered the effect of soil particle size on spectral reflectance found in previous studies, and in order to exclude this interference, we decided to standardise the size of the soil particles after mixing the soil from all the collected areas. Taking into account the fact that temperature, humidity, and degree of weathering all impact the soil formation process, soil samples were gathered from various locations within the target area during the search so as to preserve the natural environment's similarity and reduce the experimental impact of variables [30, 31]. In each black dot area, a square area (1 m^2) was randomly selected and equal amounts of soil were collected at five locations (the square's four corners and the centre) and the five samples were combined to create a representative sample. The laboratory received all the samples for the purpose of air-drying, eliminating impurities, grinding, and sieving (less than 2 mm). Before rehydrating the samples, the soil samples were subjected to a constant temperature of 105°C in an oven for 24 hours to fully eliminate the moisture content [32, 33]. Transparent petri dishes measuring 15 cm in diameter and 2 cm in depth were filled excessively with soil samples and make its surface even and flat through treatment. Each sample was saturated by slowly adding pure water using a spray bottle, and then immediately sealed with plastic wrap for 24 hours to prevent evaporation. This technique ensured that the water in the samples was spread uniformly. The identical process was used to prepare 70 samples.

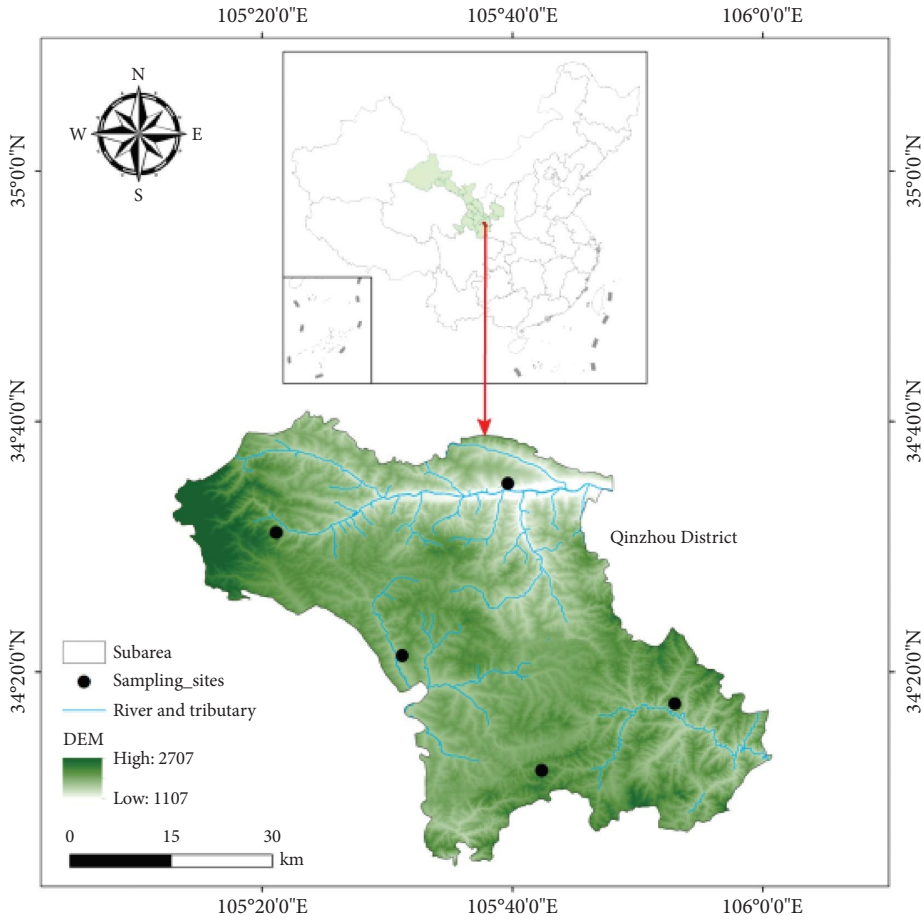
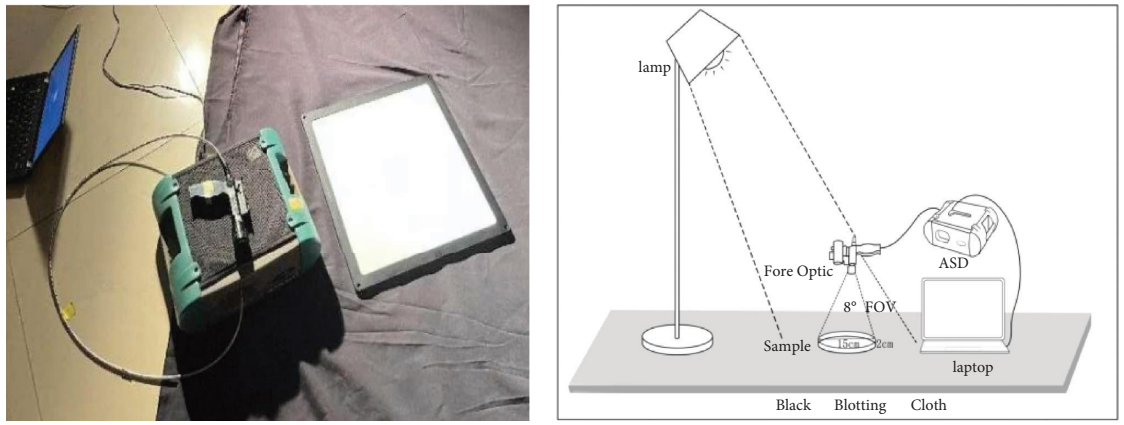


Figure 1: Sampling area.

2.2. Soil Spectral Acquisition. We utilized a FieldSpec Pro 4 spectroradiometer to measure the soil samples' spectra (Analytical Spectral Devices, Inc. USA). In a dark room, the spectral reflectance of every soil sample was measured while being positioned on a black absorbent fabric. A 200 W tungsten halogen lamp, positioned at a zenith angle of 40°, served as the light source. The gun head of the analytical spectral devices was securely positioned on the holder, with a zenith angle of 0°, at a distance of 15 cm from the soil sample's surface. The probe had a significantly narrower field of vision compared to the size of the Petri dish, as depicted in Figure 2(a) and 2(b). To enhance measurement accuracy and reduce instrument noise, the soil's spectral reflectance was measured at a perpendicular angle to the soil sample. After performing arithmetic averaging, the soil samples in the area yielded average spectral reflectance. Prior to conducting the soil spectral measurement, the instrument underwent a 15-minute preheating process. Spectral measurements involved sampling the spectral range of the soil sample at its centre. A total of 10 spectral curves were collected for each soil sample, with a sampling interval of 1 nm. The spectral value of the soil sample was considered as the average spectral reflectance. Reflectance was calibrated by a standard whiteboard with a reflectance of 99% prior to each soil spectrum measurement [34]. The reflectance measurement of one sample is completed every hour according to the uniform procedure. The precise values of the soil spectral reflectance that were collected are depicted in Figure 3.

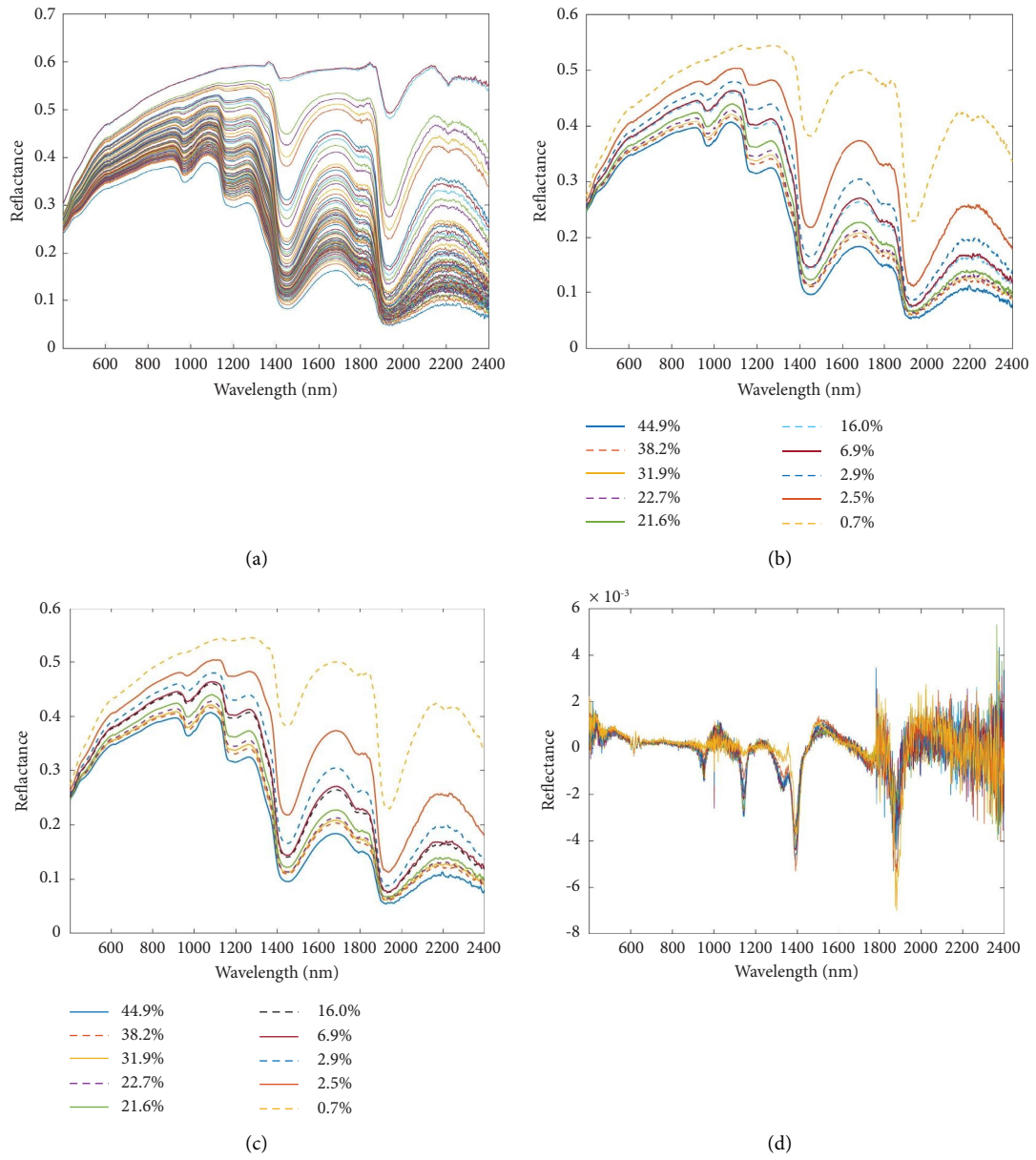
2.3. Partitioning of the Training and Validation Sets. Methods for partitioning the datasets include the method of concentration gradient, method of random sampling, Kennard-Stone method (KS), and partitioning of sample sets based on joint x-$\mathcal{Y}$ distances (SPXY). The SPXY algorithm was used to divide 70 soil samples that had been prepared beforehand into the training and validation sets. Training utilized 70% of the soil moisture content measurement data, with the remaining 30% allocated for validation. There were 50 samples in the training set and 20 samples in the validation set, with a ratio of 7 to 3. The training dataset should include a diverse and extensive range of moisture content for the samples, ensuring it is both a broad and evenly spread [35]. Alternatively, there will be the emergence of systematic forecasting mistakes.

The SPXY technique originated from the Kennard-Stone approach. The Kennard-Stone technique divides the dataset by considering the Euclidean distances of various samples in the x-vector direction (also referred to as the feature dimension direction of the dataset). Euclidean distances were



(a) (b)

Figure 2: Spectral acquisition equipment. (a) shows the actual instruments required for soil spectrum acquisition, and (b) is a plane diagram of the instruments required for this experiment.



(a) (b)

(c) (d)

Figure 3: Spectral curves of soil with varying levels of soil moisture contents. (a) Original reflectance curve; (b) randomly selected original reflectance curve; (c) Savitzky–Golay (SG); (d) first derivative (FD).

calculated for various samples in the $\mathcal{Y}$-vector direction, which corresponds to the true value dimension of the dataset. To achieve a more thorough assessment and division of the dataset, regularization was employed to merge the distances in both the x and $\mathcal{Y}$ directions. In order to give equal significance to the allocation of samples in both x and $\mathcal{Y}$ spaces, the normalized $x\mathcal{Y}$ distance can be computed [36]. $d_{x\mathcal{Y}}$ can be done in a manner similar to the Kennard-Stone algorithm, rather than using the stepwise selection process alone. Instead of solely relying on K-S, this approach utilizes both independent and response variables to choose a representative sample. It combines the independent variable (x) and the dependent variable ($\mathcal{Y}$). The target composition distance $D_{x\mathcal{Y}}(m,n)$ can be calculated for every pair of (m,n) samples. The distances between $D_x(m,n)$ and $D_{\mathcal{Y}}(m,n)$ are normalized by dividing them by the highest value in the dataset they belong to. Equation (1) can be used to calculate the normalized $x\mathcal{Y}$ distance:

$$D_{x\mathcal{Y}}(m,n) = \frac{D_x(m,n)}{\max_{m,n\in[1,N]} D_x(m,n)} + \frac{D_{\mathcal{Y}}(m,n)}{\max_{m,n\in[1,N]} D_{\mathcal{Y}}(m,n)}, \tag{1}$$

where $D_x(m,n)$ and $D_{\mathcal{Y}}(m,n)$ are the Euclidean distances of samples m and n, $\max_{m,n\in[1,N]} D_x(m,n)$ and $\max_{m,n\in[1,N]} D_{\mathcal{Y}}(m,n)$ denote the maximum distances in the $\mathcal{X}$ and $\mathcal{Y}$ directions, respectively. The total number of samples is denoted by the symbol N.

3. Key Techniques in Predicting Soil Moisture Content by Spectral Analysis

It is necessary to understand the key techniques in spectral soil moisture detection and to explore in depth the feasibility of spectral technology in soil moisture content detection. First, the original spectral data were preprocessed. Second, the spectral features sensitive to soil moisture were extracted. Third, a soil moisture content detection model was constructed and then validated.

3.1. Spectral Data Preprocessing Method. Experimental errors are caused by different factors in the process of spectral data acquisition and modeling. To establish a more stable and accurate spectral detection model of soil moisture, it is necessary to preprocess the data, eliminate interference information, highlight the absorption and reflectance peaks of the soil moisture spectra, and extract effective information [37].

The process of acquiring spectral data is vulnerable to disruption caused by noise from the instrument, uneven distribution of soil particles, and random measurement errors. Consequently, the measured sample spectra include spectral noise, which ultimately impacts the precision of the prediction model [38]. In order to enhance the relationship between spectral reflectance and soil elements, it is necessary to perform data preprocessing to extract valuable information from the spectra in nearly all assessment procedures. For this research, we utilized well-established preprocessing techniques, specifically Savitzky–Golay (SG) smoothing and first derivative (FD), to enhance the spectral profile and remove any potential baseline offset and background noise interference in the spectrum [39, 40].

The Savitzky–Golay smoothing algorithm is a popular choice for data preprocessing because of its straightforward, rapid, and user-friendly nature. The fundamental concept involves selecting a window with an odd number of points in width initially, and employing the least-squares technique to fit through the window's translation. The data are smoothed by replacing the original value with the midpoint of the window [41, 42]. In this study, the spectra were smoothed using a filter window length of 15 and a polynomial order of 3.

The first derivative (FD) can be used to identify changes in the spectral slope, thus allowing for the identification of delicate fluctuations and the recognition of overlapping peaks. Incorporating first-order derivatives and Savitzky–Golay smoothing filters into spectral data analysis provides a powerful way to improve data interpretation, recognize subtle changes, and improve the visibility of spectral features [39].

3.2. Spectral Feature Band Selection Algorithm. The study utilized three methods (CARS, GA, and ICO) for selecting the soil moisture content feature bands from the full-band spectra, employing spectral feature band selection techniques.

3.2.1. Interval Combination Optimization (ICO). Interval combination optimization is different from the other wavelength selection algorithms used in this study. It replaces wavelength points with wavelength intervals as the optimization object and uses weighted bootstrap sampling (WBS) to gradually shrink and optimize the combination of wavelength intervals. Finally, combined with a local search strategy, the edge bands of each wavelength range are further optimized [43]. To begin an ICO, the initial action is to evenly partition the spectral variable region into M subintervals of equal length based on wavelength. Using the WBS method, combinations of various wavelength intervals are sampled from M intervals in the second step, with the initial weight of each interval set at 1. The weight determines the probability of being selected for each wavelength interval.

$$p_k = \frac{w_k}{\sum_1^n w_k} \times n. \tag{2}$$

In equation (2), the number of wavelengths is n; p_k denotes the probability of the kth wavelength being selected; wk denotes the sampling weight of the kth wavelength. The third step of ICO is to establish PLSR models based on N combinations of wavelength intervals and calculate RMSECV for each model. The cross-verified mean square error (RMSECV) of the corrected set of samples indicates that the quantitative model has better prediction accuracy

and stability when the RMSECV value is smaller. The model set with the minimum RMSECV is taken as the optimal model set, and the selection ratio is denoted as α.

3.2.2. Genetic Algorithms (GA). Genetic algorithm as the heuristics, an automatic wavelength selecting process is proposed for constructing a multiparameter calibration model with the capability of self-organized and robust optimization [44]. GAs mimic the biological evolution process by drawing inspiration from the theory of biological evolution to simulate the problem that needs to be solved. Operations such as replication, crossover, and mutation generate the subsequent iteration of solutions. Solutions with low fitness function values are gradually eliminated and solutions with high fitness function values are increased. After N generations of evolution, individuals with high fitness function values will emerge. By simulating the natural evolutionary process, a search is conducted for the most favorable outcome, utilizing strong adaptability and global optimization capability [45].

This approach is to search for optimal parameters from a set of possible solutions by choosing a representative feature variable, and to increase the precision at the same time [46].

3.2.3. Competitive Adaptive Reweighted Sampling (CARS). CARS evaluates the significance of each variable by analyzing the absolute values of regression coefficients in the partial least square model. The selection of N subsets of variables is achieved through iterative N sampling runs, drawing inspiration from Darwin's evolution theory [47]. Monte Carlo resampling is employed in a competitive and iterative manner to sequentially select subsets of variables with a fixed sample ratio [48]. In each resampling iteration, the approach is employed to identify the spectral variables that exhibit significant absolute regression coefficients in the training model. The number of selected variables (nVAR) is set by an exponentially decreasing function. Finally, cross-validation is employed to identify the most suitable subsets of variables [40].

Let Y represent the desired $m \times 1$ characteristic matrix of sample soil moisture content, X represent the $m \times n$ observed spectral matrix of the sample, m represent the number of samples, n represent the number of variables, w denote the combination coefficients, T represent the submatrix X (a linear combination of X and w), c represent the regression coefficient vector of the PLSR model constructed by T and T, b represent the n-dimensional regression coefficient vector, and e represent the prediction residuals [49].

$$\begin{aligned} T &= w \cdot X. \\ Y &= c \cdot T + e = c \cdot w \cdot X + e = b \cdot X + e. \end{aligned} \tag{3}$$

The absolute value $|b_i|\,(1 \le i \ll p)$ of the i element of the regression coefficient vector $b = w \cdot c = [b_1, b_2, \cdots, b_n]$, b in equation (3) represents the contribution of the i th wavelength variable to Y. Then, the total contribution of all wavelengths to Y is $f = \sum_{i=1} \cdot |b_i|$. As a criterion for variable selection, the weight w_i was determined by calculating the proportion of $|b_i|$ each wavelength's contribution to the total (equation (4)) in order to assess its significance. The significance of the wavelength variable is indicated by higher values of $|b_i|$ and w_i.

$$w_i = |b_i| \cdot \frac{i}{f}. \tag{4}$$

Each evaluation of the importance of wavelength variables was the process of calculating w_i. The wavelength variables with smaller values of $|b_i|$ were then removed. To calculate the RMSECV values, the PLSR correction model was reconstructed using the new set of variables obtained through the adaptive reweighted sampling (ARS) technique from the retained variables. The process mentioned above was iterated N times (the predetermined number of Monte Carlo sampling) until the completion of sampling. By comparing, the model obtained the optimal subset of variables that had the smallest RMSECV value.

3.3. Prediction Model and Accuracy Evaluation. Three distinct models were employed to forecast the local soil moisture content in order to ascertain the optimal model for soil moisture content prediction. For the soil moisture content's hyperspectral inversion, three models (RF, SVM, and PLSR) were chosen in this research.

The SVM is a kernel-based approach introduced by Vapnik [50]. It is a kind of nonlinear modeling approach which is based on the theory of statistics. Based on the SVM, we can make the best decision by making use of the support vector in the training data. It is capable of dealing with both linear and nonlinear problems, and can also be used to solve the problem of regression modeling. By using SVM in the remote sensing field, it is possible to effectively manage a small training set with fewer samples, thereby decreasing the model's generalization error and minimizing the sample error, thus enhancing the model's generalization capability and achieving a high level of precision [51].

In 2001, Breiman combined random forest (RF) with classification trees to create an integrated learning algorithm. This algorithm is advantageous due to its capacity for nonlinear mining, its antinoise capabilities, its inability to meet any assumptions in terms of data distribution, its ability to quickly adapt to datasets, and its rapid training speed [52].

In spectroscopic applications, the conventional view is that partial least squares regression (PLSR) is highly resistant to interference and can be involved in building full-wavelength calibration models [53]. Wold et al. introduced the concept of partial least squares regression (PLSR). The PLSR model, which merges the features of a principal component analysis with a multiple linear regression, successfully resolves the issue of multicollinearity and notably forecasts a set of response variables from

multiple independent variables. This technique has been recognized as a beneficial instrument for gauging soil moisture [54]. During the construction of the model, it takes into consideration the correlation between spectral data sources and is able to accurately explain the spectral signal. Furthermore, it has a fairly consistent level of precision when it comes to modeling and testing.

Initially, the calibration set was used to train the model in this study, followed by verifying the accuracy using the validation set. The model accuracy test is based on the standard regression and error-index evaluation; we evaluated the model's precision and consistency by utilizing the coefficient of determination (R^2), root mean square error (RMSE), and relative analysis error (RPD). The calculation formulae are as follows:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\widehat{\mathcal{Y}}_i - \mathcal{Y}_i\right)^2}{\sum_{i=1}^{n}\left(\mathcal{Y}_i - \overline{\mathcal{Y}}_i\right)^2}, \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(\widehat{\mathcal{Y}}_i - \mathcal{Y}_i\right)^2}{n}}, \tag{6}$$

$$\mathrm{RPD} = \frac{SD_{\mathcal{Y}_i}}{\mathrm{RMSE}}, \tag{7}$$

where $\widehat{\mathcal{Y}}_i$ represents the measured values, $\mathcal{Y}_i$ represents the predicted values, and $\overline{\mathcal{Y}}_i$ denotes the average of the measured values. The standard deviation of predicted values is represented by $\mathrm{SD}_{\mathcal{Y}_i}$ and n denotes the number of soil samples.

A higher accuracy of model estimation is indicated by larger RPD and R^2 values, as well as smaller RMSE values. However, the model estimation's precision is low [55]. RPD, which is also referred to as relative prediction deviation, indicates the predictive ability of the calibration model for the data. If the RPD value is below 1.4, it signifies that the application requirements cannot be fulfilled, and the model is unable to predict the sample. A model with a relative analysis error value ranging from 1.4 to 2.0 suggests a moderate ability to predict the sample. A value of 2.0 or higher for relative analysis error signifies the model's excellent predictive capability and its ability to accurately estimate the sample. R^2 represents the relative alteration in dependent variables accounted for by the predictors. The assessment of model accuracy is provided by evaluating the degree to which the model aligns with the observed outcomes and accurately predicts future results, offering a fairly reliable measure of its performance. The root mean square error (RMSE) represents the dispersion of data points from the regression line, indicating the concentration of data around the line of best fit.

4. Results

4.1. Spectral Curve Preprocessing. This study examined two preprocessing techniques for model performance, using the original spectra as the control set. For model optimization, the original VIS-NIR spectra were compared with the results to choose the most suitable preprocessing technique for VIS-NIR-based soil moisture content modeling and prediction. The original soil spectra at various moisture levels are displayed in Figure 3. Each color corresponds to a spectrum. In order to see the reflectance spectral curves of soils with different water contents more clearly, we randomly selected ten curves from the original spectra for Savitzky–Golay smoothing and first derivative, and the results are shown in Figure 3(b). Figure 3(c) shows Savitzky–Golay smoothing of randomly selected original reflectance curves. Figure 3(d) shows the first derivative processing of randomly selected original reflectance curves.

Figure 3 displays the outcomes of every spectral preprocessing technique. In general, the VIS-NIR reflectance spectra of the 70 soil samples exhibited similar overall trends. Changes in moisture content were observed to affect the spectral reflectance of soil. Typically, the reflectivity decreased as the moisture content increased [56]. The outcome aligned with the conclusions of prior research and was attributed to the oscillations of O-H clusters and water particles. The reflectivity patterns of every soil sample exhibited three primary absorption peaks around 1,400, 1,900, and 2,200 nm. The absorption band observed at approximately 1,400 nm can be attributed to the first overtone of O-H stretching, which signifies the presence of water molecules absorbed onto the clay surface. On the other hand, the band observed at around 1,900 nm indicates a combination of O-H stretching and H-O-H bending, indicating the presence of water molecules contained within the lattice. It is conceivable that organic molecules, including CH2, CH3, and NH3, as well as Si-OH bonds and cation-OH bonds found in phyllosilicate rocks, such as kaolinite and montmorillonite, may have a potential association with wavelengths around 2,200 nm. The O-H groups, including water, exhibited the highest absorption properties in the vicinity of 1,400 nm and 1,900 nm [57], but these two bands are not practically useful in inverting soil moisture using remote sensing due to the interference of water vapour in the atmosphere [58]. Sun et al. [59] showed in field spectral data collection on soil samples that the spectral curves obtained had significant spectral data noise around 1,400 and 1,900 nm and could not be used for moisture inversion.

4.2. Selecting Feature Variables. CARS successfully achieved variable optimization by iteratively adjusting the number of sampling processes and evaluating the RMSECV values. The iterative procedure was used to determine the ideal variable subset, which had the lowest RMSECV value. Using this iterative procedure, the ideal variable subset with the lowest RMSECV value was identified. As the sample methods increased, Figure 4 illustrates a progressive decline in the number of selected wavelength variables. The RMSECV values decreased continuously during the 1st to 38th sampling processes, indicating that the variables removed during the selection process were not related to moisture content. Meanwhile, after the 38th sampling process, the RMSECV values showed a rebounding trend, indicating that important variables related to moisture content began to be removed,

which led to an increase in RMSECV values. It can also be seen that the minimum RMSECV value was obtained after the 40th sampling, with Savitzky–Golay smoothing, while the minimum RMSECV value was obtained at the fourth iteration of sampling for the spectra with first derivative preprocessing and then continuously increased.

At the 38th sampling, the original spectra R exhibited the lowest RMSECV value. The subset of spectral variables corresponding to the position of the blue asterisk "∗" in a vertical column in the figure is considered to be the best [60]. This subset contained 141 spectral variables, accounting for 7.05% of the original spectral variables. At this point, for the calibration model, $Rc^2 = 0.9858$ and RMSEC = 1.757; for the validation model, $Rv^2 = 0.944$, RMSEV = 2.7477, and RPD = 4.2511. The SG-preprocessed spectra reached the minimum RMSECV value at the 40th sampling, and 123 spectral variables were preferentially selected, accounting for 6.15% of the original spectral variables. At this point, for the calibration model, $Rc^2 = 0.9835$ and RMSEC = 1.8935; for the validation model, $Rv^2 = 0.9501$, RMSEV = 2.6093, and RPD = 4.4765. The first derivative preprocessed spectra reached the minimum RMSECV value at the fourth sampling, and 428 spectral variables were preferentially selected, accounting for 21.4% of the original spectral variables. At this point, for the calibration model, $Rc^2 = 0.9999$ and RMSEC = 0.0330; for the validation model, $Rv^2 = 0.673$, RMSEV = 2.5370, and RPD = 5.5335. The distribution of the feature wavelength points is depicted in Figure 5.

Typically, the absorption bands of peaks are hydroxyl (-OH) spectral bands, H2O spectral bands, and combined spectral bands representing the hydroxyl stretching vibrations and AL-OH vibrations [61]. The first derivative processing performed limit correction on the original spectra, making significant changes to the spectra. Thus, it can decompose overlapped mixed spectra, expand the spectral feature differences between samples, and increase spectral sensitivity bands.

The number of equal fractions of wavelength subintervals was set to 30, the number of WBS sampling processes per round was taken as 1,000, and the optimal model set ratio α was set to 0.05 when optimal band selection was performed using ICO. Figure 6 shows the ICO screening variable process. Figure 6(a)–6(c) show the weight changes of each interval for each round of sampling for the original, SG-preprocessed, and FD-preprocessed spectra, respectively. The color gradient from dark blue to dark yellow represent the increasing weight values as the iterative process proceeded. The difference between the weighted bootstrap sampling (WBS) method utilized by ICO and the weighted binary matrix sampling (WBMS) method was that even if a band changed its weight to 1 in the previous round of sampling by chance, it may still be excluded in subsequent iterations. As shown in Figure 6(a), the sampling weight of the sixth subinterval was initially 1. However, with continuous iteration, its importance gradually decreased. The weight value returned to zero in the last iteration; thus, it was not selected in the optimal subset. This indicated that ICO had high fault tolerance in wavelength selection. Figure 6(d)–6(f) show the preserved variables for different preprocessing spectra. Figure 6(d) illustrates that the ICO algorithm preserved 500 out of the 2,000 variables, with selected intervals ranging from 400 to 600 nm, 700 to 800 nm, 1,100 to , nm, and 1,400 to 15,00 nm. In the same way, (e) kept 400 variables while (f) retained 300 variables.

Figure 7 illustrates the frequency at which all variables are selected during the genetic algorithm feature variable selection process. The selection frequency thresholds are represented by two horizontal lines in the figure. Increasing the frequency threshold led to a decrease in the number of selected variables. A larger frequency threshold resulted in a smaller number of variables selected [62]. The selection frequency variable in the figure has two horizontal lines representing different numbers of modelled feature bands. The feature bands whose selection frequency is greater than the corresponding frequency of the horizontal line are selected for modelling. As to which horizontal line is taken as the selection criterion, it can be based on the requirement of model accuracy; the higher the position of the horizontal line, the lower the number of eigenwavelengths used for modelling [63]. The retained feature variables, which are above the horizontal line, were used for model construction, and those below the horizontal line are unselected variables that were not used for modeling analysis.

The preprocessed spectral values were calculated by GA and repeated five times to screen the characteristic spectral bands. Figure 7 shows the results with soil spectral bands, for example, the spectral measurement range is between 401 and 2,400 nm, with 2,000 bands in total, and the horizontal axis represents the sequence number of bands 1-1,999. As for which horizontal line is chosen as the selection criterion, the result shows that the upper horizontal line has a higher number of wavelengths and the lower horizontal line has a lower number of wavelengths. It was found that the higher the position of the horizontal line, the lower the number of feature wavelengths modelled. When a small number of feature wavelengths are modelled, a lower model prediction error can be obtained. Nineteen feature variables were selected from the original spectra using genetic algorithm in this study, accounting for 0.95% of all the variables in the VIS-NIR spectra. Eighteen feature variables were selected from the Savitzky–Golay preprocessed spectra, accounting for 0.9% of all variables in the VIS-NIR spectra. From the results of model processing, 67 feature variables were selected from the first derivative preprocessed spectra, accounting for 3.35% of all the variables in the VIS-NIR spectra.

4.3. Accuracy and Validation of Different Prediction Models. The accuracy of the PLSR, SVM, and RF models constructed using SMC, along with the full-band spectral data and feature bands filtered by various methods, is presented in Table 1. The models' estimation accuracy was enhanced to varying extents after Savitzky–Golay and first derivative preprocessing, in comparison to R (original spectra). The models constructed using the feature bands had higher estimation accuracy compared to the calibration and

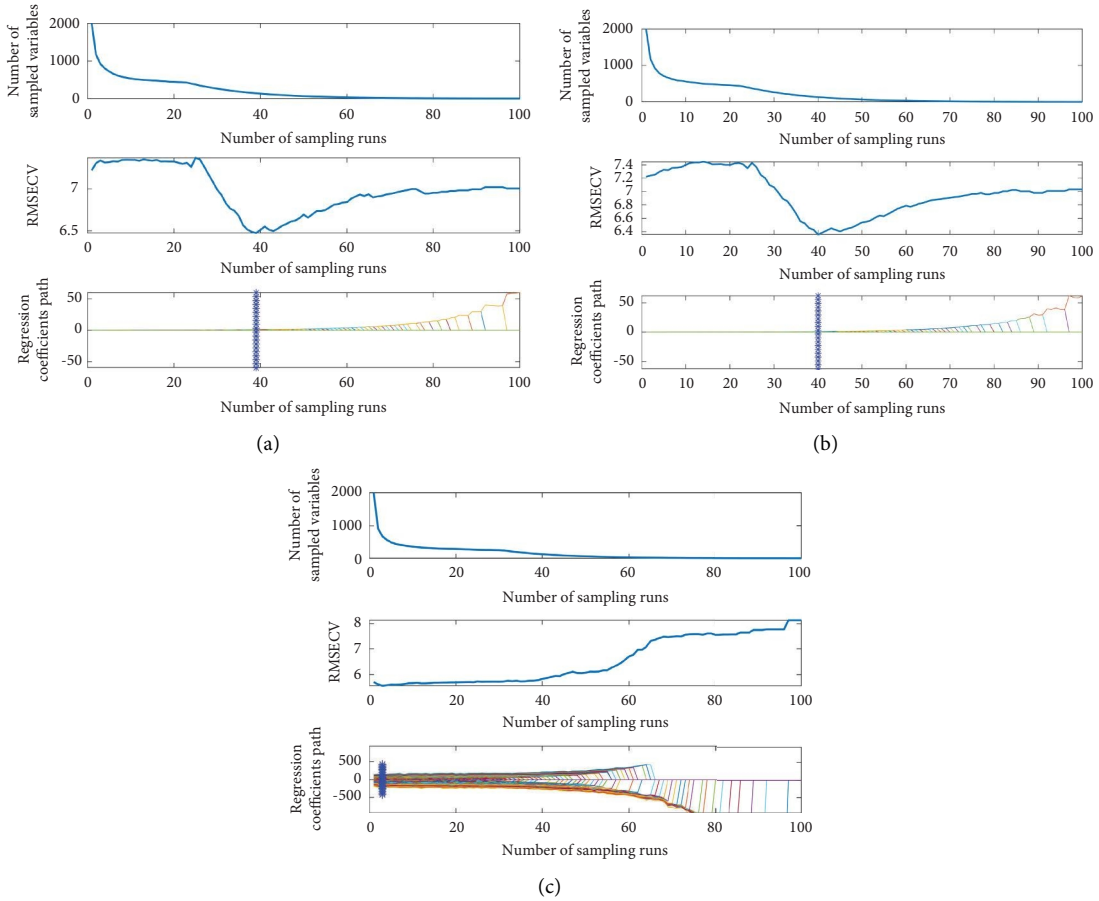


Figure 4: The outcomes of variable selection for soil moisture content (SMC) in the visible (VIS)-near infrared (NIR) band using competitive adaptive reweighted sampling (CARS) are shown in the figure. (a) Effect of CARS on the original spectra; (b) effect of CARS on SG preprocessed spectra; and (c) effect of CARS on FD preprocessed spectra.

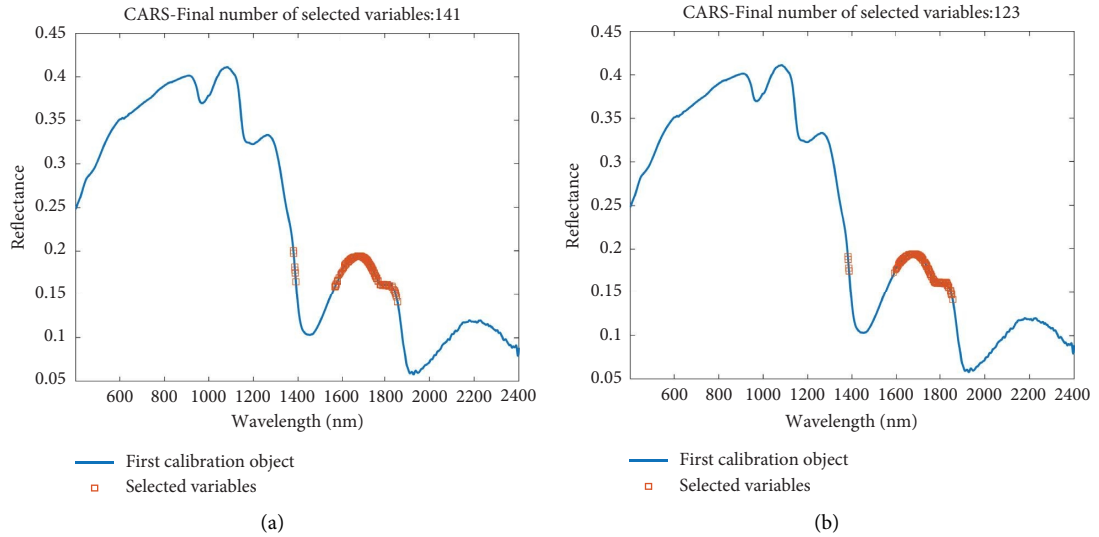


Figure 5: Continued.

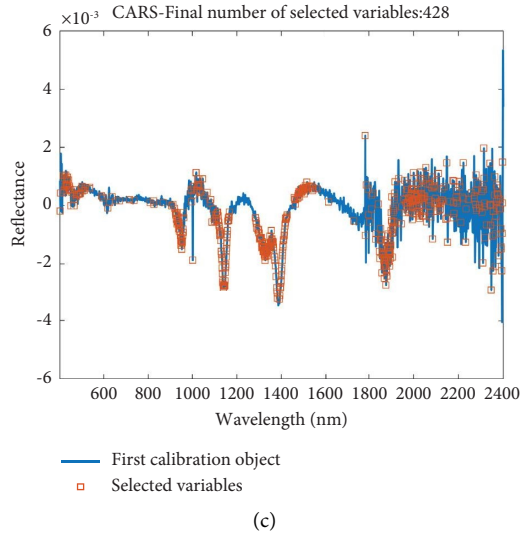


Figure 5: In the VIS-NIR band, the figure displays the count of variables chosen by the CARS-based SMC. (a) Effect of CARS on the original spectra; (b) effect of CARS on SG preprocessed spectra; and (c) effect of CARS on FD derivative preprocessed spectra.

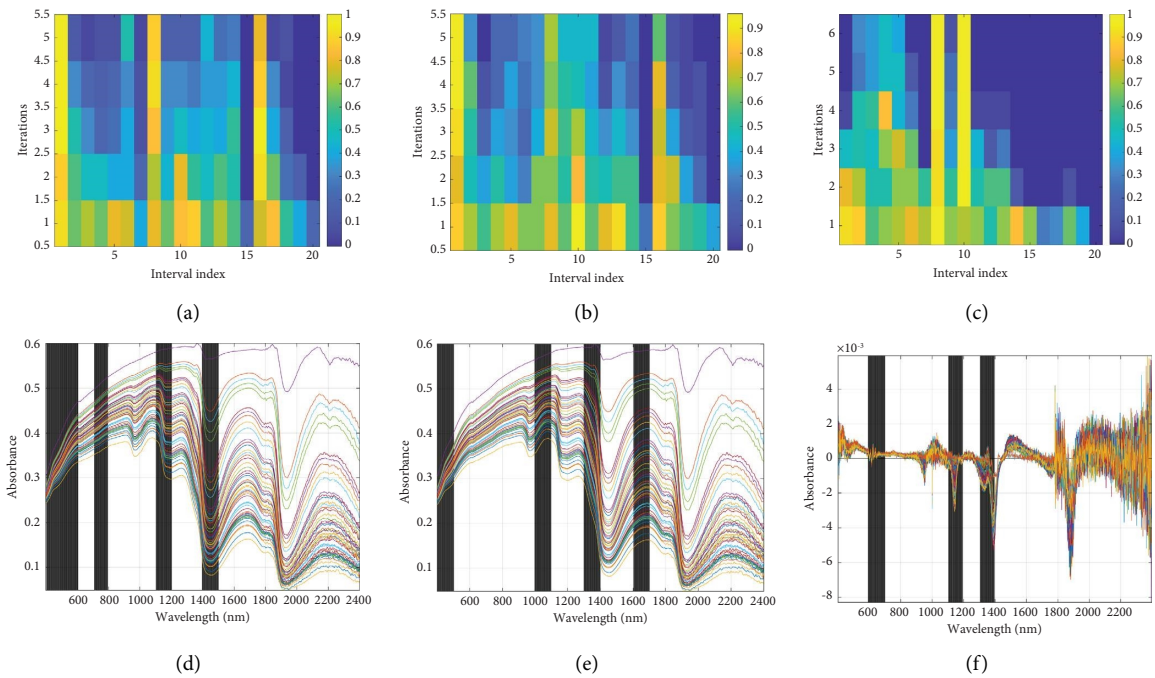


Figure 6: Interval combination optimization (ICO) algorithm iteration process. Weight changes of each wavelength interval during the iterative process: (a) original spectra; (b) SG-preprocessed spectra; (c) FD-preprocessed spectra; algorithm-selected preserve variables: (d) original spectra; (e) SG-preprocessed spectra; (f) FD-preprocessed spectra.

validation sets' modeling of the full-band spectral data (401-2,400 nm) for all three feature variable selection methods. Specifically, the Savitzky–Golay (SG)-genetic algorithm (GA) technique achieved feature band modeling by utilizing 18 modeling variables, accounting for a mere 0.9% of the complete spectra. The ranking of the three feature screening methods in terms of their ability to enhance model accuracy was determined as GA > CARS > ICO. Based on the evaluation results in Table 1, we used the 19 sensitive wavelengths selected using R-GA and the 18 sensitive wavelengths preferred by FD-GA as input variables for modeling soil water content using PLSR, RF, and SVM methods, respectively.

A comparison of the data presented in Table 1 leads to the use of the FD-GA-RF combination, when $Rc^2 = 0.9957$ and RMSEC = 0.9838; the validation model $Rv^2 = 0.9962$ and RMSEV = 0.8643, and RPD = 16.2421. Similarly, for both PLSR and SVM, the corresponding optimal combinations of

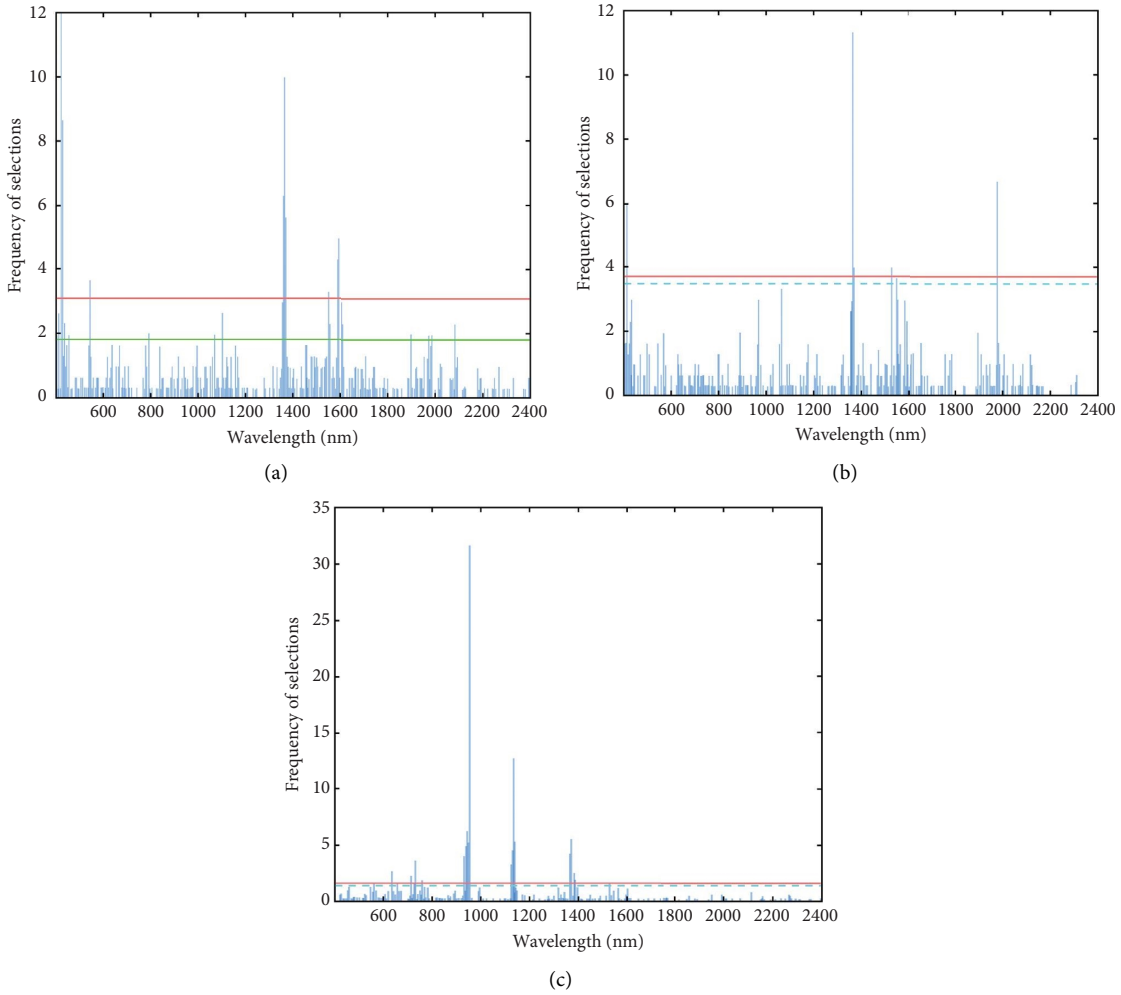


Figure 7: Genetic algorithm (GA) feature variable screening process. (a) Effect of the GA on the original spectra; (b) effect of the GA on SG-preprocessed spectra; and (c) effect of the GA on FD-preprocessed spectra.

Table 1: Cross-validation of the spectral dataset and the PLSR, SVM, and RF calibration models of SMC.

Spectral preprocessing	Feature variable selection	Model	Calibration models		Validation models		
			R^2	RMSE	R^2	RMSE	RPD
R	CARS	PLSR	0.9858	1.7570	0.9447	2.7477	4.2511
		SVM	0.9440	1.1139	0.9652	2.1799	5.3583
		RF	0.9960	0.9380	0.9923	1.0227	11.4218
	GA	PLSR	0.9504	3.2853	0.8591	4.3838	2.6645
		SVM	0.9964	0.8912	0.9900	1.1684	9.9968
		RF	0.9956	0.9821	0.9897	1.1843	9.8631
	ICO	PLSR	0.9890	1.5505	0.9889	1.2329	9.4738
		SVM	0.9956	0.9812	0.9864	1.3618	8.5770
		RF	0.9961	0.9211	0.9908	1.1179	10.4485
SG	CARS	PLSR	0.9835	1.8935	0.9501	2.6093	4.4765
		SVM	0.9940	1.1415	0.9657	2.1647	5.3959
		RF	0.9959	0.9500	0.9912	1.0948	10.6693
	GA	PLSR	0.9583	3.0143	0.8803	4.0415	2.8902
		SVM	0.9956	0.9795	0.9800	1.2786	9.1357
		RF	0.9960	0.9305	0.9901	1.1609	10.0612
	ICO	PLSR	0.9915	1.3639	0.9884	1.2559	9.3002
		SVM	0.9953	1.0075	0.9856	1.4018	8.3328
		RF	0.9952	1.0233	0.9903	1.1488	10.1676

Table 1: Continued.

Spectral preprocessing	Feature variable selection	Model	Calibration models		Validation models		
			R^2	RMSE	R^2	RMSE	RPD
FD	CARS	PLSR	0.9999	0.0030	0.9673	2.5370	5.5335
		SVM	0.9996	0.2267	0.9697	2.4433	5.7458
		RF	0.9891	1.5736	0.9888	1.4843	9.4583
	GA	PLSR	0.9908	1.4448	0.9904	1.372	10.2275
		SVM	0.9866	1.7434	0.9828	1.8420	7.6215
		RF	0.9957	0.9838	0.9962	0.8643	16.2421
	ICO	PLSR	0.9998	0.0884	0.9744	2.1090	6.6564
		SVM	0.9997	0.2379	0.9833	1.1827	7.7447
		RF	0.9955	1.0124	0.9868	1.6146	8.6949

CARS (competitive adaptive reweighted sampling); GA (genetic algorithm); ICO (interval combination optimization); R (original spectra); SG (Savitzky–Golay); FD (first derivative); PLSR (partial least squares regression); SVM (support vector machine); RF (random forest).

soil moisture content prediction models can be obtained. In addition, the RPD value is calculated to be 16.2421. Likewise, the optimal combinations of soil moisture content prediction models can be acquired for both PLSR and SVM. Figure 8 displays the outcomes of the modelling. The visual representation accurately depicts the relationship between the observed and projected soil moisture content values. The calibration model's trend line closely matched the target trend line, with a value close to 1 : 1, suggesting that the FD-GA-RF model successfully predicted various soil moisture contents. A high degree of fit of the validation model trend line to the target trend line indicated that the FD-GA-RF model ensured good prediction robustness.

The RF model demonstrated superior forecasting compared to the SVM and PLSR models. The relationship between soil moisture content and the spectrum was more complex. PLSR is a linear method that performs poorly at solving nonlinear problems, whereas SVM and RF can better solve complex nonlinear relationships between independent and dependent variables. However, the SVM model is prone to severe bias estimation caused by high spectral noise, which reduces the model's accuracy. The RF model incorporates two machine-learning technologies of random feature selection and the Bagging algorithm. Compared with the traditional classifier algorithm, the RF model can better tolerate outliers and noise, so that the established model has higher accuracy and better robustness. Moreover, it can handle continuous and discrete data simultaneously [64].

5. Discussion

The study's findings demonstrate that choosing appropriate spectral feature bands is an essential first step in creating reliable models. There is a lot of redundant information in the full-band data, which can be somewhat eliminated by the variable filtering [65]. In this study, the feature wavelengths of CARS, GA, and ICO were selected. The 401-2,400 nm band of the original VIS-NIR spectra was used as the full wavelength for feature band selection. Different preprocessing conditions were considered when selecting the feature wavelengths using the three methods. The new datasets were constructed by utilizing these wavelengths to create the three prediction models. Both GA and CARS algorithms possess identical traits as they are both global optimization methods. Both algorithms explore the entire solution space and detect confidential data within their search collections. The ICO algorithm retains more useful information than the GA and CARS algorithms.

In order to guarantee the full representation of the calibration set samples and the even distribution of all samples within each set, the dataset was partitioned using SPXY [66]. The study compared the effects of pretreatment using two different methods. The findings indicated that certain preprocessing techniques were not successful in eliminating noise and minimizing errors in the spectral data of intricate sample systems. The first derivative preprocessing technique was employed to enhance the calibration model's performance in contrast to the initial spectral modeling. In order to develop the model, the results were also compared to determine the most effective method. The significance of spectral preprocessing in forecasting SOM content in Vis-NIR spectroscopy lies in its ability to diminish or nullify spectral noise, thereby enhancing the model's predictive precision [67]. Nonetheless, not every spectral preprocessing technique is capable of yielding favorable outcomes. As a result, choosing the right spectral pretreatment technique is crucial. Research has demonstrated that first derivative (FD), Savitzky–Golay (SG) smoothing, and other approaches all have significant spectral preprocessing effects; nevertheless, the effects of FD spectral preprocessing are superior and more stable than those of the other techniques [35].

When the feature selection algorithm was executed in isolation, CARS excelled in identifying highly informative variables, whereas GA excelled in determining optimal band combinations. However, the computation time of GA was the longest and its capability to simplify the model inputs was weaker. The ICO exhibited the poorest ability to extract efficient wavelengths, which ultimately did not enhance the model's accuracy. When employing feature variable selection techniques, certain methods enhance the model's processing speed by refining the variable selection outcomes, whereas others enhance the model's predictive capability and improve its quality [15]. Typically, simplifying the model coincided with a boost in the model's predictive capability. Furthermore, the selection of characteristic variables using

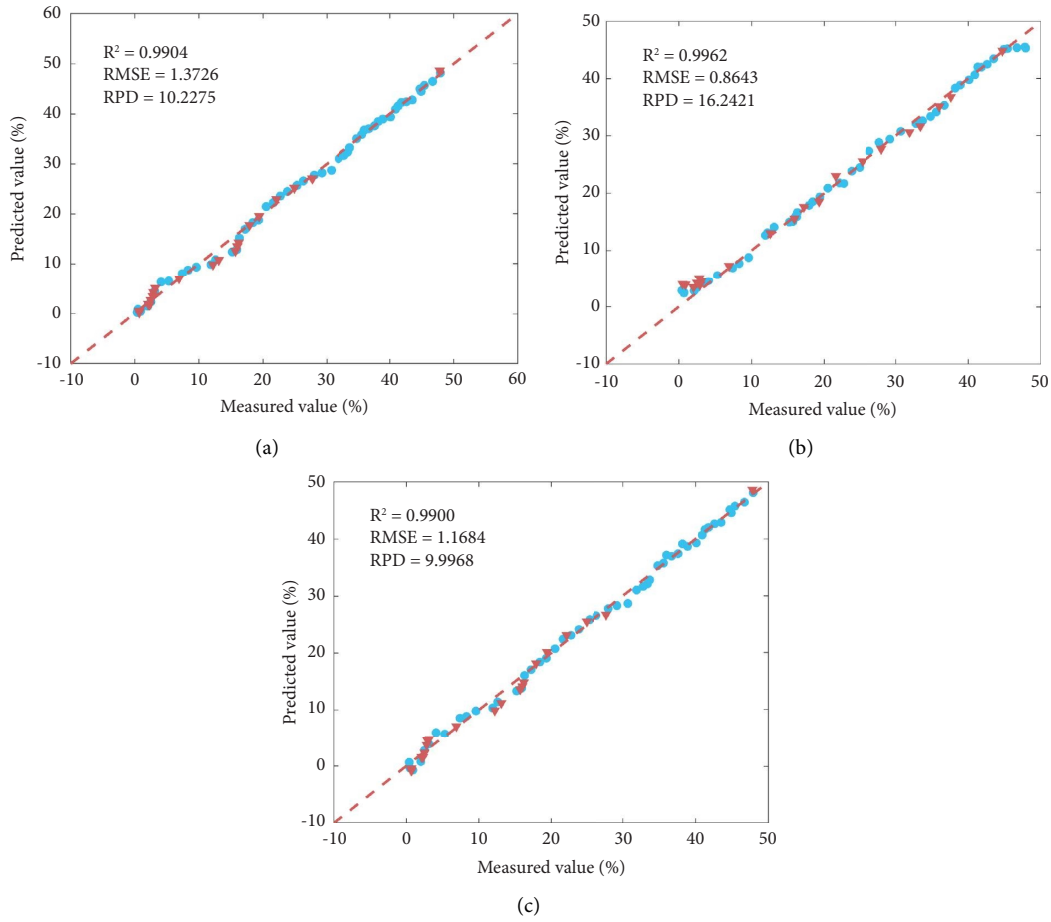


Figure 8: Scatter plot of soil moisture content with optimal estimation based on (a) R-GA-PLSR, (b) FD-GA-RF, and (c) R-GA-SVM.

various spectral preprocessing techniques can also influence the generated models. Hence, it is essential to examine and deliberate on the preprocessing technique that is more advantageous for modeling when combined with the process of selecting wavelength variables.

After Savitzky–Golay smoothing and first derivative preprocessing of the original reflectance of soil samples at different soil moisture contents, the feature absorption bands were more obvious. The characteristics of the absorption peaks were particularly prominent near 450; 1,400; 1,900; and 2,200 nm. This laid the foundation for variable optimization. By selecting CARS variables for soil samples at different soil moisture contents, the optimal variable set for predicting the soil moisture content was obtained. The GA algorithm had high prediction accuracy and strong prediction capability and effectively reduced the number of modelling wavelength variables. In order to more accurately estimate the amount of organic matter, Sun et al. [68] selected the band using a genetic algorithm (GA) and hyperspectral satellite data and the PLSR model.

The ideal combination of prediction models was obtained by combining different preprocessing methods with feature variable selection algorithms. The PLSR model achieved the best combination of R-GA-PLSR with $R^2 = 0.9904$, RMSE = 1.3726, and RPD = 10.2275. For the RF model, the optimal FD-GA-RF combination was $R^2 = 0.9962$, RMSE = 0.8643, and RPD = 16.2421. For the SVM model, the optimal R-GA-SVM combination was $R^2 = 0.9900$, RMSE = 1.1684, and RPD = 9.9968. The SVM model's prediction performance on SMC was not as good as that of RF and PLSR. This might be explained by the restricted anti-interference ability of SVM and limitations resulting from parameter choices, such as kernel functions and penalty factors [69]. However, the performance of the PLSR model is subject to negative influence from other factors, such as texture and color [70]. The RF method is especially good at solving nonlinear problems because of its robustness, which is demonstrated by its strong anti-interference and antioverfitting characteristics as well as its high tolerance to background noise and outliers [71]. This outcome is consistent with the SMC inversion investigation conducted by Eyo et al. [72]. The random forest (RF) classifier employs several decision trees for the purpose of training and predicting samples. RF requires low hyperparameter settings [73]. When the sample features are high-dimensional, it arbitrarily chooses features to place at the decision tree's nodes in order to efficiently train the model for any dataset. RF is not sensitive to the lack of several

essential features and is comparatively easy to build. In summary, FD-GA-RF was the best combined model for estimating the soil moisture content.

6. Conclusions

After gathering soil samples from the target area, the spectral reflectance of soils with varying moisture contents was measured indoors using controlled variables. More accurate results were obtained by employing a range of spectral treatments, feature band selection, and prediction techniques in the subsequent work. Nevertheless, because this was a study carried out in a particular area, variations in soil texture, soil moisture content, and environmental factors may make the model less appropriate in some areas than others. Both feature bands and full bands were used in these models. The RF model outperformed the SVM and PLSR models in terms of accuracy. The coefficient of determination (R^2) for the calibration and validation sets of the RF-established soil moisture content prediction model was 0.9957 and 0.9962, respectively. The validation set had a relative prediction deviation (RPD) of 16.2421. The combination of feature variable selection and regression methods significantly enhanced modeling efficiency and maintained accuracy, in comparison to full-band modeling. Therefore, the results of this study can provide research ideas for examining soil properties in various contexts; further research is needed to see whether the model developed in this study can be applied to other places. In order to reduce the effects of various environmental factors, such as light, temperature, and moisture, on spectral reflectance, the soil samples in this investigation were sieved and prepared in an ideal laboratory setting. Future studies will therefore focus on determining how adaptable the current soil moisture content prediction model is in various settings and on undisturbed soil.

Data Availability

The data that support the findings of this study are available upon request from the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

We are very grateful to Bo Song, Shichao Wu, and Yichen Wei for their assistance during the experimental process. This research was performed at the Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences (CAS), Hefei, China. This research was funded by the National Natural Science Foundation of China, Grant number 42205146.

References

[1] M. Sadeghi, E. Babaeian, M. Tuller, and S. B. Jones, "The optical trapezoid model: a novel approach to remote sensing of soil moisture applied to Sentinel-2 and Landsat-8 observations," *Remote Sensing of Environment*, vol. 198, pp. 52–68, 2017.

[2] F. Hu, J. Liu, C. Xu et al., "Soil internal forces initiate aggregate breakdown and splash erosion," *Geoderma*, vol. 320, pp. 43–51, 2018.

[3] L. Samaniego, S. Thober, R. Kumar et al., "Anthropogenic warming exacerbates European soil moisture droughts," *Nature Climate Change*, vol. 8, no. 5, pp. 421–426, 2018.

[4] X. Zhang, X. Yuan, H. Liu, H. Gao, and X. Wang, "Soil moisture estimation for winter-wheat waterlogging monitoring by assimilating remote sensing inversion data into the distributed hydrology soil vegetation model," *Remote Sensing*, vol. 14, no. 3, p. 792, 2022.

[5] S. Chen, Q. Yan, S. Jin et al., "Soil moisture retrieval from the CyGNSS data based on a bilinear regression," *Remote Sensing*, vol. 14, no. 9, p. 1961, 2022.

[6] N. O. Gholami Bidkhani and M. R. Mobasheri, "Influence of soil texture on the estimation of bare soil moisture content using MODIS images," *European Journal of Remote Sensing*, vol. 51, no. 1, pp. 911–920, 2018.

[7] S. L. Su, D. N. Singh, and M. Shojaei Baghini, "A critical review of soil moisture measurement," *Measurement*, vol. 54, pp. 92–105, 2014.

[8] J. Yuan, X. Wang, C. X. Yan, S. R. Wang, X. P. Ju, and Y. Li, "Soil moisture retrieval model for remote sensing using reflected hyperspectral information," *Remote Sensing*, vol. 11, no. 3, p. 366, 2019.

[9] X. Wang, Q. Liu, Z. Qu, L. Wang, X. Li, and Y. Wang, "Inversion and verification of salinity soil moisture using microwave radar," *Transactions of the Chinese Society of Agricultural Engineering*, 2017.

[10] A. Longmire, T. A. S. Poblete, J. Hunt, D. Chen, and P. J. Zarco-Tejada, "Assessment of crop traits retrieved from airborne hyperspectral and thermal remote sensing imagery to predict wheat grain protein content," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 193, pp. 284–298, 2022.

[11] W. Sun and X. Zhang, "Estimating soil zinc concentrations using reflectance spectroscopy," *International Journal of Applied Earth Observation and Geoinformation*, vol. 58, pp. 126–133, 2017.

[12] X. Bao, C. Zhang, W. Li, M. Eisa, S. El-Gamal, and B. Benmokrane, "Monitoring the distributed impact wave on a concrete slab due to the traffic based on polarization dependence on stimulated Brillouin scattering," *Smart Materials and Structures*, vol. 17, no. 1, p. 015003, 2007.

[13] J. Liu, Z. Dong, J. Xia et al., "Estimation of soil organic matter content based on CARS algorithm coupled with random forest," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 258, 2021.

[14] Xiangyu, W. Jingzhe, W. Fei, C. Lianghong, and S. Huilan, "Estimation of soil moisture content based on competitive adaptive reweighted sampling algorithm coupled with machine learning," *Acta Optica Sinica*, vol. 38, no. 10, p. 1030001, 2018.

[15] W. Sun, S. Liu, X. Zhang, and Y. Li, "Estimation of soil organic matter content using selected spectral subset of hyperspectral data," *Geoderma*, vol. 409, 2022.

[16] M. A. Munnaf and A. M. Mouazen, "Spectra transfer based learning for predicting and classifying soil texture with short-ranged Vis-NIRS sensor," *Soil and Tillage Research*, vol. 225, 2023.

[17] M. Knadel, F. Castaldi, R. Barbetti, E. Ben-Dor, A. Gholizadeh, and R. Lorenzetti, "Mathematical techniques to remove moisture effects from visible–near-infrared–shortwave-infrared soil spectra—review," *Applied Spectroscopy Reviews*, vol. 58, no. 9, pp. 629–662, 2022.

[18] A. P Leone, R. A Viscarra-Rossel, P. Amenta, and A. Buondonno, "Prediction of soil properties with PLSR and vis-NIR spectroscopy: application to mediterranean soils from Southern Italy," *Current Analytical Chemistry*, vol. 8, no. 2, pp. 283–299, 2012.

[19] M. Bilgili, "Prediction of soil temperature using regression and artificial neural network models," *Meteorology and Atmospheric Physics*, vol. 110, no. 1-2, pp. 59–70, 2010.

[20] D. Rolnick, P. L. Donti, L. H. Kaack et al., "Tackling climate change with machine learning," *ACM Computing Surveys*, vol. 55, no. 2, pp. 1–96, 2022.

[21] M. Davari, S. A. Karimi, H. A. Bahrami, S. M. Taher Hossaini, and S. Fahmideh, "Simultaneous prediction of several soil properties related to engineering uses based on laboratory Vis-NIR reflectance spectroscopy," *Catena*, vol. 197, 2021.

[22] R. Viscarra Rossel, T. Behrens, E. Ben-Dor et al., "A global spectral library to characterize the world's soil," *Earth-Science Reviews*, vol. 155, pp. 198–230, 2016.

[23] G. Naibo, R. Ramon, G. Pesini et al., "Near-infrared spectroscopy to estimate the chemical element concentration in soils and sediments in a rural catchment," *Catena*, vol. 213, 2022.

[24] S. Dharumarajan and R. Hegde, "Digital mapping of soil texture classes using Random Forest classification algorithm," *Soil Use & Management*, vol. 38, no. 1, pp. 135–149, 2022.

[25] S. Nawar and A. Mouazen, "On-line vis-NIR spectroscopy prediction of soil organic carbon using machine learning," *Soil and Tillage Research*, vol. 190, pp. 120–127, 2019.

[26] N. Liu, "Taxonomy and species diversity of soil ciliates in Maijishan scenic spot", Master's thesis, Northwest Normal University, Lanzhou.

[27] J. R. Guo, *Preliminary Study on the Evaluation System of Wild Medicinal Plant Resources in Maijishan Scenic Spot*, Forestry Science and Technology, Newsletter, 2019.

[28] C. D. Patz, A. Blieke, R. Ristow, and H. Dietrich, "Application of FT-MIR spectrometry in wine analysis," *Analytica Chimica Acta*, vol. 513, no. 1, pp. 81–89, 2004.

[29] Y. Yuan, C. Xi, and Q. Jing, "Advance in agricultural drought monitoring using remote sensing data," *Spectroscopy and Spectral Analysis*, 2019.

[30] B. B. Guo, Y. L. Feng, C. Ma et al., "Suitability of different multivariate analysis methods for monitoring leaf N accumulation in winter wheat using in situ hyperspectral data," *Computers and Electronics in Agriculture*, vol. 198, 2022.

[31] W. Wu, X. Zhong, C. Lei et al., "Sampling survey method of wheat ear number based on UAV images and density map regression algorithm," *Remote Sensing*, vol. 15, no. 5, p. 1280, 2023.

[32] G. Liu, Q. Li, W. Quan, and A. Wang, "Effects of raising chickens under pinus massoniana forest on soil physicochemical properties and microbial community," *Polish Journal of Environmental Studies*, vol. 32, no. 3, pp. 2707–2718, 2023.

[33] C. T. Garten Jr., S. Kang, D. J. Brice, C. W. Schadt, and J. Zhou, "Variability in soil properties at different spatial scales (1m–1km) in a deciduous forest ecosystem," *Soil Biology and Biochemistry*, vol. 39, no. 10, pp. 2621–2627, 2007.

[34] W. Ji, V. I. Adamchuk, A. Biswas et al., "Assessment of soil properties in situ using a prototype portable MIR spectrometer in two agricultural fields," *Biosystems Engineering*, vol. 152, pp. 14–27, 2016.

[35] Y. Xiong, R. Zhang, F. Zhang et al., "A spectra partition algorithm based on spectral clustering for interval variable selection," *Infrared Physics & Technology*, vol. 105, 2020.

[36] T. Angelopoulou, A. Balafoutis, G. Zalidis, and D. Bochtis, "From laboratory to proximal sensing spectroscopy for soil organic carbon estimation—a review," *Sustainability*, vol. 12, no. 2, p. 443, 2020.

[37] T. Chen, Q. Chang, J. Liu, J. Clevers, and L. Kooistra, "Identification of soil heavy metal sources and improvement in spatial mapping based on soil spectral information: a case study in northwest China," *Science of the Total Environment*, vol. 565, pp. 155–164, 2016.

[38] G. Wang, J. Yang, and R. Li, "Imbalanced SVM-based anomaly detection algorithm for imbalanced training datasets," *ETRI Journal*, vol. 39, no. 5, pp. 621–631, 2017.

[39] A. C. Dotto, R. S. D. Dalmolin, A. Ten Caten, and S. Grunwald, "A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra," *Geoderma*, vol. 314, pp. 262–274, 2018.

[40] J. Sun, W. Yang, M. Zhang, M. Feng, L. Xiao, and G. Ding, "Estimation of water content in corn leaves using hyperspectral data based on fractional order Savitzky-Golay derivation coupled with wavelength selection," *Computers and Electronics in Agriculture*, vol. 182, 2021.

[41] A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[42] J. L. P. Calle, I. Punta-Sánchez, A. V. González-De-Peredo, A. Ruiz-Rodríguez, M. Ferreiro-González, and M. Palma, "Rapid and automated method for detecting and quantifying adulterations in high-quality honey using vis-NIRs in combination with machine learning," *Foods*, vol. 12, no. 13, p. 2491, 2023.

[43] X. Song, Y. Huang, H. Yan, Y. Xiong, and S. Min, "A novel algorithm for spectral interval combination optimization," *Analytica Chimica Acta*, vol. 948, pp. 19–29, 2016.

[44] J. Ning, M. Sheng, X. Yi et al., "Rapid evaluation of soil fertility in tea plantation based on near-infrared spectroscopy," *Spectroscopy Letters*, vol. 51, no. 9, pp. 463–471, 2018.

[45] Y. Liu, G. W. Zhang, and D. Liu, "Simultaneous measurement of chlorophyll and water content in navel orange leaves based on hyperspectral imaging," *Spectroscopy*, 2014.

[46] Y. Xu, F. Y. Kutsanedzie, H. Sun et al., "Rapid Pseudomonas species identification from chicken by integrating colorimetric sensors with near-infrared spectroscopy," *Food Analytical Methods*, vol. 11, no. 4, pp. 1199–1208, 2018.

[47] Y. H. Yun, H. D. Li, B. C. Deng, and D. S. Cao, "An overview of variable selection methods in multivariate analysis of near-infrared spectra," *TrAC, Trends in Analytical Chemistry*, vol. 113, pp. 102–115, 2019.

[48] L. Liang, L. Wei, G. Fang et al., "Prediction of holocellulose and lignin content of pulp wood feedstock using near infrared spectroscopy and variable selection," *Spectrochimica Acta*

Part A: Molecular and Biomolecular Spectroscopy, vol. 225, 2020.

[49] J. Alcaraz, M. Labbé, and M. Landete, "Support Vector Machine with feature selection: a multiobjective approach," *Expert Systems with Applications*, vol. 204, 2022.

[50] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[51] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.

[52] D. M. Haaland, K. L. Higgins, and D. R. Tallant, "Multivariate calibration of carbon Raman spectra for quantitative determination of peak temperature history," *Vibrational Spectroscopy*, vol. 1, no. 1, pp. 35–40, 1990.

[53] S. R. Araújo, J. Wetterlind, J. A. M. Demattê, and B. Stenberg, "Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from B razil by clustering into smaller subsets or use of data mining calibration techniques," *European Journal of Soil Science*, vol. 65, no. 5, pp. 718–729, 2014.

[54] X. Huang, Z. Shi, H. Zhu, H. Zhang, L. Ai, and W. Yin, "Soil moisture dynamics within soil profiles and associated environmental controls," *Catena*, vol. 136, pp. 189–196, 2016.

[55] K. Katarzyna, S. Justyna, S. Jakub, and S. Marcin, "Estimation of bare soil moisture from remote sensing indices in the 0.4-2.5 mm spectral range," *Transactions on Aerospace Research*, vol. 2021, no. 2, pp. 1–11, 2021.

[56] X. Y. Liu, L. Wang, R. J. Song, M. Liu, and Q. R. Chang, "Spectral prediction of soil moisture content during wind drying of yellow cotton soil," *Journal of Agricultural Machinery*, vol. 12, 2015.

[57] Y. Sun, X. Zheng, Q. Qin et al., "Modeling soil spectral reflectance with different mass moisture content," *Spectroscopy and Spectral Analysis*, vol. 35, no. 8, pp. 2236–2240, 2015.

[58] Y. Wang, F. Jiang, B. B. Gupta et al., "Variable selection and optimization in rapid detection of soybean straw biomass based on CARS," *IEEE Access*, vol. 6, pp. 5290–5299, 2018.

[59] S. Xu, Y. Zhao, M. Wang, and X. Shi, "Comparison of multivariate methods for estimating selected soil properties from intact soil cores of paddy fields by Vis–NIR spectroscopy," *Geoderma*, vol. 310, pp. 29–43, 2018.

[60] C. Malegori, J. Muncan, E. Mustorgi, R. Tsenkova, and P. Oliveri, "Analysing the water spectral pattern by near-infrared spectroscopy and chemometrics as a dynamic multidimensional biomarker in preservation: rice germ storage monitoring," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 265, 2022.

[61] H. Gao, H. Mao, and X. Zhang, "Measurement of nitrogen content in lettuce canopy using spectroscopy combined with BiPLS-GA-SPA and elm," *Guang Pu Xue Yu Guang Pu Fen Xi= Guang Pu*, vol. 36, no. 2, pp. 491–495, 2016.

[62] T. Qiao, C. Lv, and W. Xiao, "Hyperspectral prediction model of soil texture based on genetic algorithm," *Chinese Journal of Soil Science*, vol. 11, 2018.

[63] S. Nawar and A. M. Mouazen, "Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques," *Catena*, vol. 151, pp. 118–129, 2017.

[64] J. Wang, J. Ding, D. Yu et al., "Capability of Sentinel-2 MSI data for monitoring and mapping of soil salinity in dry and wet seasons in the Ebinur Lake region, Xinjiang, China," *Geoderma*, vol. 353, pp. 172–187, 2019.

[65] R. K. H. Galvao, M. C. U. Araujo, G. E. José, M. J. C. Pontes, E. C. Silva, and T. C. B. Saldanha, "A method for calibration and validation subset partitioning," *Talanta*, vol. 67, no. 4, pp. 736–740, 2005.

[66] A. Gholizadeh, L. Borůvka, M. M. Saberioon, J. Kozák, R. Vašát, and K. Němeček, "Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features," *Soil and Water Research*, vol. 10, no. 4, pp. 218–227, 2015.

[67] Z. Zhang, J. Ding, C. Zhu, and J. Wang, "Combination of efficient signal pre-processing and optimal band combination algorithm to predict soil organic matter through visible and near-infrared spectra," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 240, 2020.

[68] Y. Dai and P. Zhao, "A hybrid load forecasting model based on support vector machine with intelligent methods for feature selection and parameter optimization," *Applied Energy*, vol. 279, 2020.

[69] J. Zhang, Y. Tian, X. Yao, W. Cao, X. Ma, and Y. Zhu, "Estimating soil total nitrogen content based on hyperspectral analysis technology," *Journal of Natural Resources*, vol. 13, 2011.

[70] R. Dalal and R. Henry, "Simultaneous determination of moisture, organic carbon, and total nitrogen by near infrared reflectance spectrophotometry," *Soil Science Society of America Journal*, vol. 50, no. 1, pp. 120–123, 1986.

[71] Z. Fu, J. Jiang, Y. Gao et al., "Wheat growth monitoring and yield estimation based on multi-rotor unmanned aerial vehicle," *Remote Sensing*, vol. 12, no. 3, p. 508, 2020.

[72] E. Eyo, S. Abbey, T. Lawrence, and F. Tetteh, "Improved prediction of clay soil expansion using machine learning algorithms and meta-heuristic dichotomous ensemble classifiers," *Geoscience Frontiers*, vol. 13, no. 1, 2022.

[73] B. Hu, J. Xue, Y. Zhou et al., "Modelling bioaccumulation of heavy metals in soil-crop ecosystems and identifying its controlling factors using machine learning," *Environmental Pollution*, vol. 262, 2020.