applied sciences

MDPI

Article

Enhanced Hyperspectral Forest Soil Organic Matter Prediction Using a Black-Winged Kite Algorithm-Optimized Convolutional Neural Network and Support Vector Machine

Yun Deng [1,2], Lifan Xiao [1,2,*] and Yuanyuan Shi [3]

[1] College of Computer Science and Engineering, Guilin University of Technology, Guilin 541006, China; 2002078@glut.edu.cn
[2] Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin 541006, China
[3] Key Laboratory of Central South Fast-Growing Timber Cultivation of Forestry Ministry of China, Guangxi Forestry Research Institute, Nanning 530002, China; syyfly@163.com
* Correspondence: 1020221061@glut.edu.cn

Abstract: Soil Organic Matter (SOM) is crucial for soil fertility, and effective detection methods are of great significance for the development of agriculture and forestry. This study uses 206 hyperspectral soil samples from the state-owned Yachang and Huangmian Forest Farms in Guangxi, using the SPXY algorithm to partition the dataset in a 4:1 ratio, to provide an effective spectral data preprocessing method and a novel SOM content prediction model for the study area and similar regions. Three denoising methods (no denoising, Savitzky–Golay filter denoising, and discrete wavelet transform denoising) were combined with nine mathematical transformations (original spectral reflectance (R), first-order differential (1DR), second-order differential (2DR), MSC, SNV, logR, (logR)′, 1/R, ((1/R)′) to form 27 combinations. Through Pearson heatmap analysis and modeling accuracy comparison, the SG-1DR preprocessing combination was found to effectively highlight spectral data features. A CNN-SVM model based on the Black Kite Algorithm (BKA) is proposed. This model leverages the powerful parameter tuning capabilities of BKA, uses CNN for feature extraction, and uses SVM for classification and regression, further improving the accuracy of SOM prediction. The model results are RMSE = 3.042, R^2 = 0.93, MAE = 4.601, MARE = 0.1, MBE = 0.89, and PRIQ = 1.436.

Keywords: optimization algorithm; organic matter content; spectral data processing; forest soil; Guangxi

Academic Editor: Borja Velazquez-Marti

Received: 13 November 2024
Revised: 18 December 2024
Accepted: 29 December 2024
Published: 7 January 2025

Citation: Deng, Y.; Xiao, L.; Shi, Y. Enhanced Hyperspectral Forest Soil Organic Matter Prediction Using a Black-Winged Kite Algorithm-Optimized Convolutional Neural Network and Support Vector Machine. *Appl. Sci.* **2025**, *15*, 503. https://doi.org/10.3390/app15020503

1. Introduction

Soil Organic Matter (SOM) refers to a complex mixture of organic compounds at various stages of decomposition, including plant residues, microbial products, and rhizosphere inputs [1]. It also contains fixed proportions of nitrogen (N), phosphorus (P), and sulfur (S) [2]. SOM plays multiple roles, such as storing water, retaining nutrients, improving soil physical properties, and reducing pollution, making it an important indicator of soil fertility [3]. Rapid and accurate determination of SOM content and monitoring its dynamic changes are crucial for the development of forestry and agriculture.

However, traditional SOM measurement methods are costly, time-consuming, and complex, making it difficult to accurately reflect soil changes [4,5]. Additionally, the use of large amounts of chemical solvents and analytical reagents poses risks to personnel and the environment [6]. In contrast, using hyperspectral reflectance in the visible (350–700 nm)

and near-infrared (700–2500 nm) ranges to predict SOM content offers advantages such as shorter cycles, cost-effectiveness, and no pollution [7,8]. This method has been applied to measure various soil parameters and is widely used in ecology, agriculture, medicine, and food fields [9–11].

Hyperspectral technology has demonstrated its broad effectiveness in monitoring Soil Organic Matter (SOM) content. The vast amount of spectral band information enriches the dimensions and information content of soil spectral data, providing a solid foundation for precise analysis [8]. However, this also brings challenges in data processing, including increased information redundancy and the emergence of multicollinearity issues [12]. These phenomena increase the complexity of data processing and analysis. To effectively utilize this data for predicting soil properties, it is crucial to identify appropriate data preprocessing and modeling methods [13,14].

The main goal of data preprocessing is to improve the accuracy and robustness of models. Common mathematical transformations, Savitzky–Golay (SG) smoothing, Discrete Wavelet Transform (DWT) smoothing, Multiplicative Scatter Correction (MSC), and Standard Normal Variate (SNV), can effectively improve prediction models [15–17]. Mathematical transformations can effectively highlight characteristic bands of the spectrum, smoothing operations can suppress spectral noise and reduce random noise impact, MSC adjusts spectral skew by correcting scattering effects, and SNV reduces scattering effects caused by sample surface irregularities and particle size differences [18–20]. For example, Bao et al. used a wavelet transform combined with nine mathematical transformations in Xinjiang to establish a Random Forest (RF) model, effectively enhancing spectral band features and reducing noise interference [21]. Zhang et al. found that SG denoising combined with first-order differential transformation effectively extracted the spectral characteristics of sandy loam soil in the experimental area [22]. Carvalho et al. used Standard Normal Variate (SNV) preprocessing to establish an SVM model to predict SOM content in southern Brazil [23].

In terms of modeling, more and more researchers are using machine learning (ML) and deep learning (DL) techniques to explore the relationship between hyperspectral data and SOM. PLSR [24], RF [21], and SVM [25] are still robust prediction models in this field. With the improvement of computing power, DL is gradually being applied in various fields. DL can reveal complex nonlinear relationships between spectra and soil properties and has been proven to outperform geostatistics and other existing traditional ML algorithms [26–28]. For example, Haghi et al. used Scottish soil spectral data to predict multiple soil contents and found that CNN performed better than PLSR and SVM [29]. Hao et al. used a dual-branch CNN network for modeling, effectively extracting spectral data features [30].

Differences in soil physical properties lead to changes in spectral data characteristics. These differences also cause complex multicollinearity issues among spectral bands [31–33]. Therefore, targeted complex feature extraction is necessary. This study employs a combination of three widely used denoising methods and nine mathematical transformation methods for preprocessing. We then conduct modeling using SVM, PLSR, BPNN, and CNN to compare their performance and determine the optimal preprocessing method for the spectral data in the study area.

In recent years, research has found that integrating different types of modeling methods provides richer feature representations, significantly enhancing model performance and developing many application scenarios [34,35]. Combining CNN component construction with hyperparameter tuning can significantly advance the application of CNN in soil spectral modeling [15,36]. Based on these points, this study proposes a CNN-SVM prediction model optimized by the Black Kite algorithm. This model leverages the powerful

feature extraction capabilities of CNN and the classification and regression abilities of SVM, combined with the strong optimization ability of BKA, aiming to develop a new method for measuring SOM content in forest soils in Guangxi and similar regions.

2. Materials and Methods

2.1. Study Area

2.1.1. Overview of the Study Area

The study area is located in the state-owned Yachang Forest Farm and Huangmian Forest Farm in Guangxi, both of which belong to the subtropical climate zone. Detailed information is shown in Table 1. The mainland type in this area is forest land, with mountainous soil being the predominant soil type. The planting structure is relatively simple, primarily consisting of eucalyptus, Chinese fir, and pine trees. Due to the lack of effective soil nutrient detection methods, it is not possible to scientifically apply fertilizers. Additionally, unreasonable human planting practices have led to a decline in soil fertility, which in turn affects the overall health of the forest.

Table 1. Tree farm information.

Name	Yachang Forest Farm	Huangmian Forest Farm
Latitude and Longitude	106°08′–106°26′ E, 24°37′–25°00′ N	109°43′–109°58′ E, 24°37′–24°52′ N
Annual Average Rainfall	1058 mm	1750 mm
Annual Average Evaporation	1484.7 mm	1426 mm
Annual Average Temperature	16.8 °C	19 °C

2.1.2. Soil Sample Collection

Soil samples were collected using the S-shaped sampling method in Tianlin County and Luzhai County, Guangxi Zhuang Autonomous Region, with the distribution of sampling points shown in Figure 1. All collected samples were naturally air-dried and finely ground in the laboratory. Some samples were filtered through a 0.2 mm soil sieve, and their Soil Organic Carbon (SOC) content was accurately measured using the potassium dichromate oxidation method. The SOC content was then multiplied by a coefficient of 1.724 to obtain the SOM content. Another portion of the samples was filtered through a 0.149 mm sieve, and detailed spectral data from the visible to near-infrared region (350–2500 nm) were captured using an ASD FieldSpec1 4 Hi-Res (Purchased by Beijing LICA United Technology Limited from Analytical Spectral Devices in Boulder, CO, USA) ground object spectrometer. To enhance the accuracy of the spectral data, each sample's spectral data were collected 10 times, and the arithmetic mean was taken as the final spectral data. Additionally, to eliminate noise caused by external factors during the operation, the noisy edge bands of 350–399 nm and 2401–2500 nm were removed.

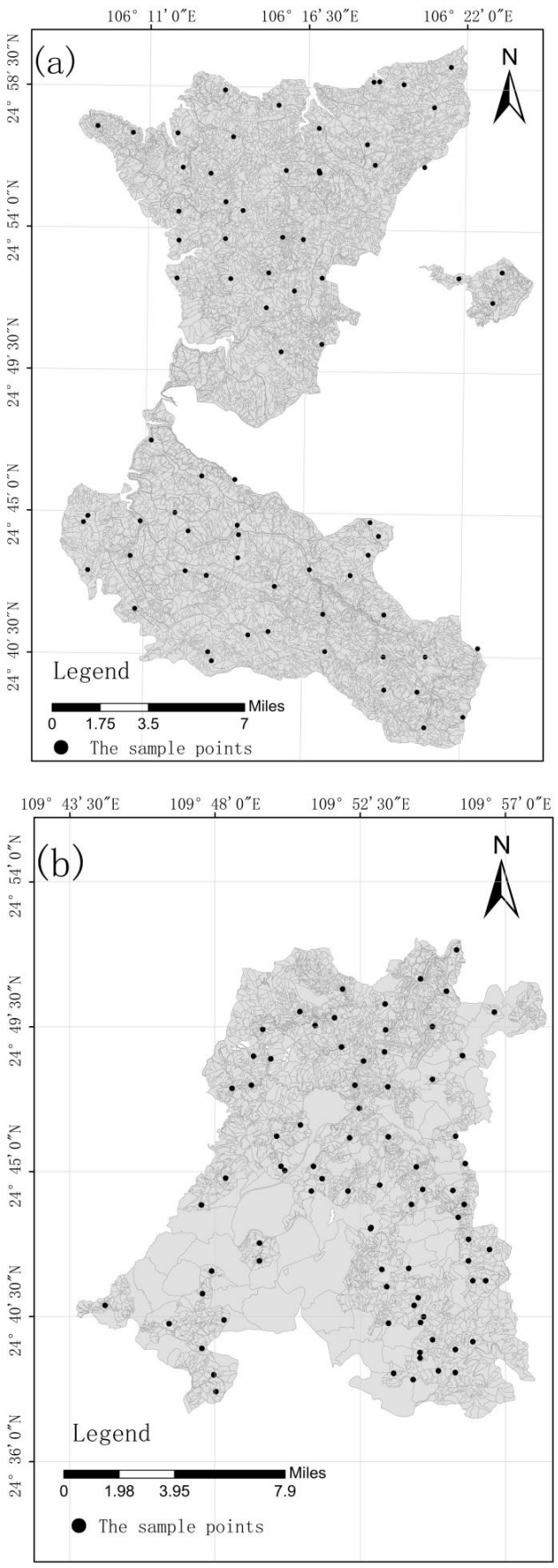

Figure 1. Distribution of sampling points in the study area. (**a**) is Yachang Forest Farm. (**b**) is Huangmian Forest Farm.

2.2. Experimental Design

In this study, spectral measurements were conducted on 206 samples from the study area. Three denoising methods were used: no denoising (N), S-G filter denoising (SG), and discrete wavelet transform denoising (DWT). Nine mathematical transformation methods were applied: original spectral reflectance (R), first-order differential (1DR), second-order differential (2DR), multiplicative scatter correction (MSC), standard normal variate transformation (SNV), logarithmic transformation (logR), first-order differential of logarithmic transformation ((LogR)'), reciprocal transformation (1/R), and first-order differential of reciprocal transformation ((1/R)'). These combinations resulted in 27 preprocessing combinations.

The Pearson correlation coefficient was used to preliminary screen the preprocessing combinations that were more effective in extracting spectral data features. Seven modeling methods were then selected for modeling: Support Vector Machine (SVM), Partial Least Squares Regression (PLSR), Backpropagation Neural Network (BP), Convolutional Neural Network (CNN), CNN-SVM, CNN network improved by the Black-winged Kite Algorithm (BKA-CNN), and CNN-SVM network improved by the Black-winged Kite Algorithm (BKA-CNN-SVM). The optimal preprocessing combination and the best modeling method were selected. The methodological flow of this study is shown in Figure 2.

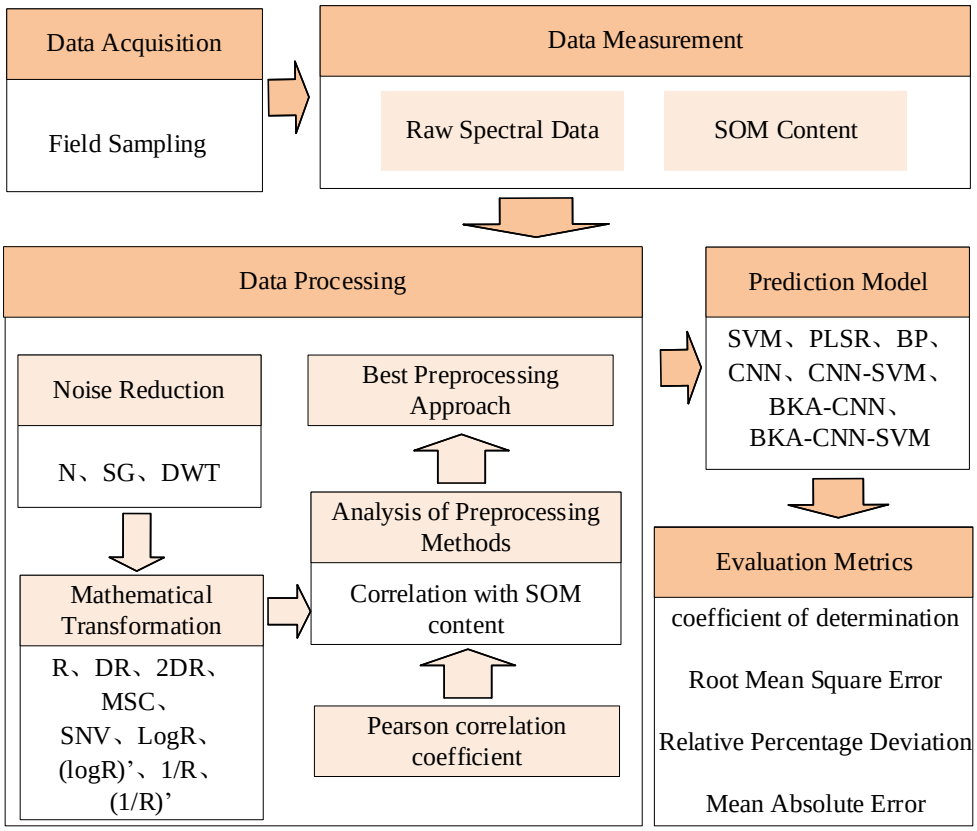

Figure 2. Method flowchart.

2.3. Spectral Preprocessing

2.3.1. Savitzky–Golay Filter

The SG (Savitzky–Golay) filter is widely used for smoothing and denoising spectral data, effectively handling data in most experimental areas [37,38]. It eliminates noise while preserving the width and shape of the signal. This filter is based on local polynomials

and uses a weighted moving average through a convolution window to achieve the least squares method to fit the curve. The formula is shown in Equation (1).

$$\widehat{y_j} = \frac{\sum_{i=-m}^{m} C_i X_{j+i}}{n} \quad (1)$$

where $\widehat{y_j}$ is the smoothed data, X_{j+i} is the data to be processed, C_i is the smoothing coefficient, m is the window size, and ($n = 2m + 1$). In this study, the smoothing coefficient is set to 3, and the window size is set to 21.

2.3.2. DWT Denoising

Discrete Wavelet Transform (DWT) achieves wavelet denoising by decomposing and filtering high-frequency noise layer by layer [21]. Data processed through DWT not only reduces noise interference but also enhances feature expression to a certain extent [39]. The steps of wavelet denoising include discrete wavelet decomposition, threshold processing, and wavelet reconstruction. Threshold processing involves filtering the detail coefficients decomposed at each layer, while wavelet reconstruction involves adding the approximate coefficients of the last layer and the detail coefficients processed by the threshold at each layer to reconstruct a one-dimensional signal. This study uses the Donoho denoising method [40], with the threshold formula shown in Equation (2).

$$threshold = \sigma\sqrt{2log_e M} \quad (2)$$

where σ is the predicted standard deviation of the wavelet coefficients, and M is the total number of sample wavelet coefficients. In this study, the threshold is set to half of the threshold value, and the wavelet function used is db4.

2.4. SPXY Algorithm

The SPXY algorithm, developed from the Kennard-Stone (KS) algorithm, is a statistically-based sample partitioning method widely used in spectral data partitioning [41,42]. Compared to the KS algorithm, it can simultaneously consider two variables—in this case, spectral reflectance and SOM content. By using these two variables to calculate the Euclidean distance, the SPXY algorithm ensures the maximum retention of sample distribution, effectively covering the multidimensional vector space to enhance model stability. The distance calculation formula is given by Equations (3)–(5).

$$d_x(p,q) = \sqrt{\sum_{j=1}^{N}\left[x_p(j) - x_q(j)\right]^2}, (p,q \in [1,N]) \quad (3)$$

$$d_y(p,q) = \left|y_p - y_q\right|, (p,q \in [1,N]) \quad (4)$$

$$d_{xy}(p,q) = \frac{d_x(p,q)}{maxd_x(p,q)} - \frac{d_y(p,q)}{maxd_y(p,q)}, (p,q \in [1,N]) \quad (5)$$

where j represents the bands in the spectrum, and $x_p(j)$ and $x_q(j)$ denote the spectral reflectance of samples p and q at band j, respectively. N is the total number of samples. $d_x(p,q)$ represents the Euclidean distance between two samples in the spectral feature space. $d_{xy}(p,q)$ is an upgraded version of $d_x(p,q)$ in the KS algorithm, ensuring that the sample data have equal weights in both the x and y spaces.

2.5. Black-Winged Kite Algorithm

2.5.1. Algorithm Overview

The Black-winged Kite Algorithm (BKA) is an optimization algorithm proposed by Wang et al. in March 2024 [43]. It is inspired by the high adaptability and intelligent behavior exhibited by black-winged kites in nature, particularly during their attacks, migrations, and hunting processes. These unique biological characteristics inspired researchers to develop a new swarm intelligence optimization algorithm aimed at better handling complex problems. The algorithm not only captures the flight and hunting behaviors of black-winged kites but also deeply simulates their high adaptability to environmental changes and target locations. This endows the algorithm with strong evolutionary capabilities, fast search speed, and excellent ability to find optimal solutions. The unique bio-inspired features of this algorithm provide it with robust dynamic search capabilities, enabling it to effectively cope with constantly changing optimization environments. The pseudocode of the algorithm is shown in Algorithm 1 (BKA Algorithm Pseudocode).

Algorithm 1 Black-winged kite algorithm

Input: The population size p, maximum number of iterations T, and variable dimension d.
Output: The optimal solution to the problem to be solved is obtained through BKA.

Initialize the position of each bk and select the population leader Y_L.
While (t < T)**do**
  if p < r
    $b_{t+1}^{i,j} = b_t^{i,j} + n(1 + \sin(r)) \times b_t^{i,j}$
  else if do
    $b_{t+1}^{i,j} = b_t^{i,j} + n \times (2r - 1) \times b_t^{i,j}$
  end if
  if $F_i < F_{ri}$**do**
    $b_{t+1}^{i,j} = b_t^{i,j} + C(0,1) \times \left(b_t^{i,j} - L_t^j\right)$
  else if do
    $b_{t+1}^{i,j} = b_t^{i,j} + C(0,1) \times \left(L_t^j - m \times b_t^{i,j}\right)$
  end if
  if $b_{t+1}^{i,j} < L_t^j$
    $Y_L = b_{t+1}^{i,j}$, $F_{best} = f\left(b_{t+1}^{i,j}\right)$
  else if do
    $Y_L = L_t^j$, $F_{best} = f\left(L_t^j\right)$
  end if
end while
ReturnY_L and F_{best}

2.5.2. Population Initialization

In the BKA, an initial set of random solutions is generated, with each Black-winged Kite (BK) positioned as described in Equation (6).

$$BK = \begin{bmatrix} BK_{1,1} & BK_{1,2} & \cdots & BK_{1,d} \\ BK_{2,1} & BK_{2,2} & & BK_{2,d} \\ \vdots & & \ddots & \vdots \\ BK_{p,1} & BK_{p,2} & \cdots & BK_{p,d} \end{bmatrix} \quad (6)$$

where p represents the number of potential solutions and d denotes the dimensionality of the problem. Subsequently, each Black-winged Kite (BK) is uniformly distributed as described in Equation (7).

$$Y_i = BK_{lb} + rand(BK_{ub} - BK_{lb}) \tag{7}$$

where i is an integer between 1 and p, lb and ub represent the algorithm's lower and upper bounds, respectively, and rand is a random value between 0 and 1. During initialization, the BKA selects the best fitness value as the population leader Y_L. The expression for Y_L, using the minimum value as an example, is provided in Equations (8) and (9).

$$f_{best} = \min(f(Y_i)) \tag{8}$$

$$Y_L = Y(find(f_{best} == f(Y_i))) \tag{9}$$

2.5.3. Aggressive Behavior

As predators of small mammals and insects on the grasslands, BK adjust their wing and tail angles according to wind speed during flight. They hover to observe their prey before diving to attack. This process includes various attack patterns for prey search, as described in Equations (10) and (11).

$$b_{t+1}^{i,j} = \begin{cases} b_t^{i,j} + n(1 + \sin(r)) \times b_t^{i,j}, p < r \\ b_t^{i,j} + n \times (2r - 1) \times b_t^{i,j}, \ else \end{cases} \tag{10}$$

$$n = 0.05e^{-2(\frac{t}{T})^2} \tag{11}$$

where $b_t^{i,j}$ represents the position of the i-th BK in the j-th dimension at the t-th iteration, r is a random number between 0 and 1, p is a constant set to 0.9, and (T) is the number of iterations completed so far.

2.5.4. Migration Behavior

Bird migration is a complex behavior influenced by seasonal changes, climate, and food availability, typically led by a population leader. When the current population's fitness is lower than that of a random population, the leader relinquishes leadership and joins the migration queue. Conversely, if the current population's fitness is higher, the leader continues to guide the population to the destination. This dynamic leader selection strategy ensures successful migration, as mathematically modeled in Equations (12) and (13).

$$b_{t+1}^{i,j} = \begin{cases} b_t^{i,j} + C(0,1) \times \left(b_t^{i,j} - L_t^j\right), F_i < F_{ri} \\ b_t^{i,j} + C(0,1) \times \left(L_t^j - m \times b_t^{i,j}\right), \ else \end{cases} \tag{12}$$

$$m = 2 \times \sin\left(r + \frac{\pi}{2}\right) \tag{13}$$

where L_t^j represents the leader of the j-th dimension for the t-th iteration, and F_i denotes the fitness of any BK in the j-th dimension at the t-th iteration. $C(0,1)$ refers to the Cauchy mutation. The one-dimensional Cauchy distribution is a continuous probability distribution with two parameters. When δ = 1 and μ = 0, it converts to the standard form of the probability density function, as shown in Equation (14).

$$f(x, \delta, \mu) = \frac{1}{\pi} \frac{\delta}{\delta^2 + (x - \mu)^2} = \frac{1}{\pi} \frac{1}{x^2 + 1}, -\infty < x < \infty \tag{14}$$

2.6. CNN-SVM Network Enhanced by the Black-Winged Kite Algorithm

Convolutional Neural Networks (CNNs) are a crucial architecture in the field of deep learning. Their basic structure typically includes an input layer, convolutional layers, activation functions, pooling layers, fully connected layers, and an output layer, along with regularization techniques to prevent overfitting.

Support Vector Machines (SVMs) are a significant algorithm in machine learning, widely used for classification and regression tasks. They effectively handle both linearly separable and non-linearly separable data, constructing optimal decision boundaries in high-dimensional spaces.

The CNN-SVM approach combines the strengths of deep learning and machine learning [44]. CNN acts as feature extractors, learning high-level feature representations from spectral data, which are then classified by SVMs. The overall workflow is illustrated in Figure 3. SVMs offer excellent generalization and robustness, effectively handling high-dimensional feature spaces with numerous spectral bands. This approach avoids the excessive parameters and overfitting issues associated with traditional CNN fully connected layers. For spectral data, CNN can extract features from both global and local regions through convolution operations, preserving the spatial structure of spectral bands. SVM kernel functions then map these features to high-dimensional spaces, capturing nonlinear relationships more effectively and simplifying computations in the original feature space.

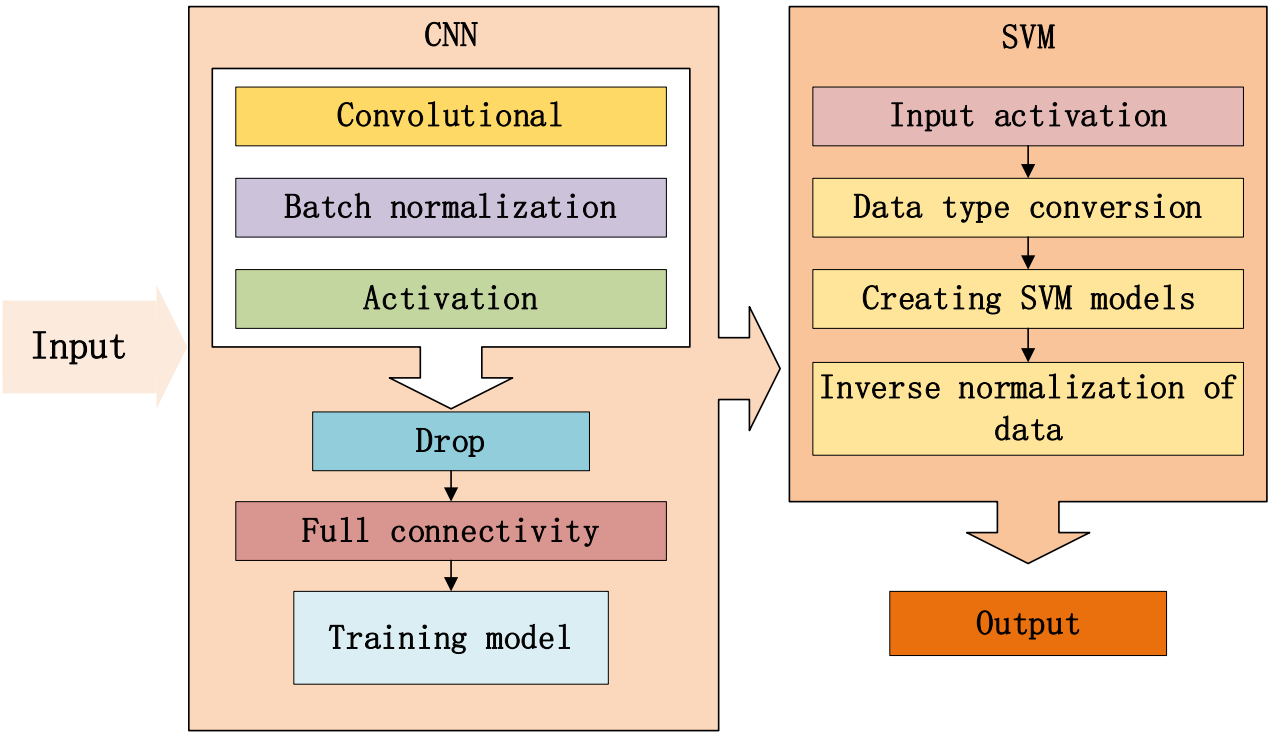

Figure 3. CNN-SVM flowchart.

Optimization is performed using the BKA, leveraging its robust search capabilities and balance between local and global optima to find the optimal parameters. A model is then constructed based on these optimal results, and the data are trained to obtain feature parameters. These extracted feature parameters are input into the SVM network to complete the prediction, as illustrated in Figure 4.

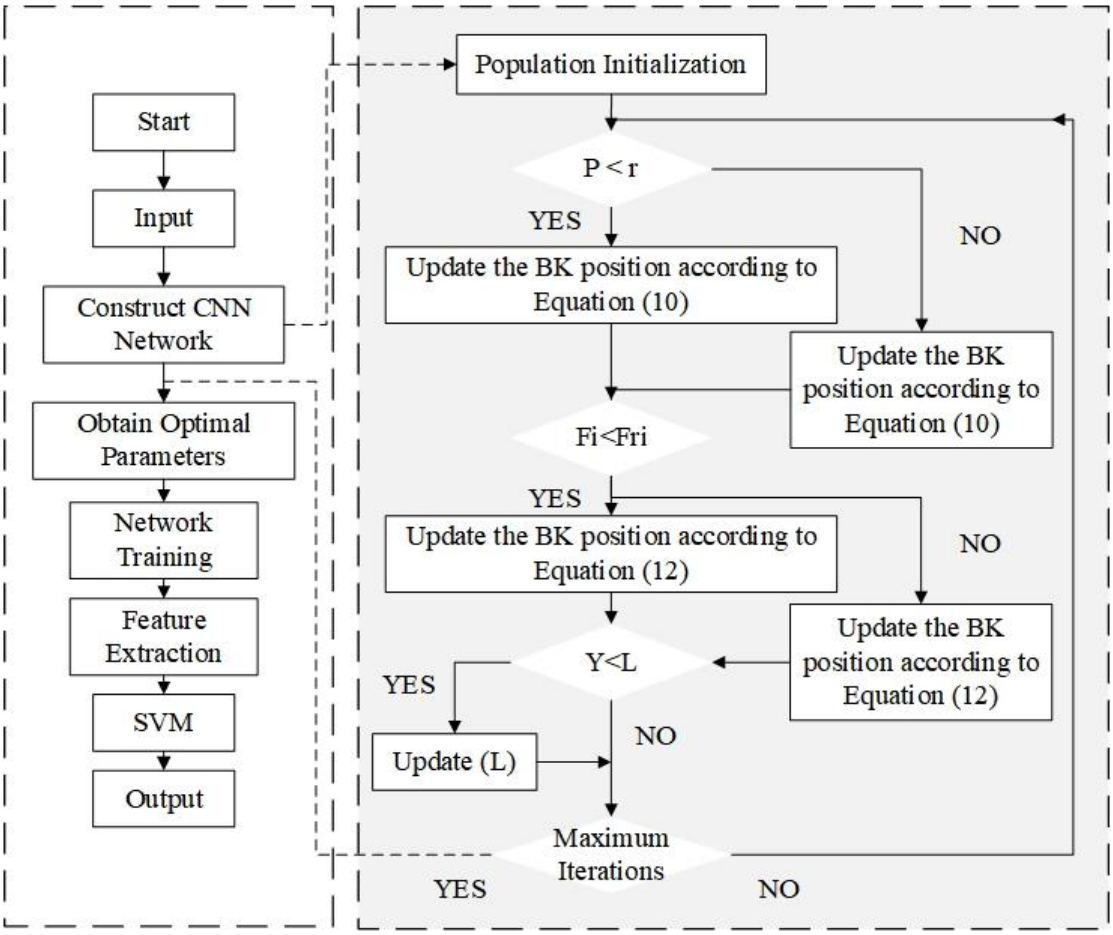

Figure 4. BAK-CNN-SVM model flowchart.

2.7. Evaluation Metrics

The study employs six evaluation metrics to assess model performance: the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), mean absolute relative error (MARE), mean bias error (MBE), and the ratio of performance to interquartile range (RPIQ).

R^2, RMSE, MAE, MARE, and MBE are widely used evaluation metrics in the fields of soil spectral data prediction and artificial intelligence [45]. R^2 reflects the correlation between the actual soil SOM content and the model's predicted values. The closer R^2 is to 1, the higher the model's accuracy, as shown in Equation (16). RMSE indicates the deviation between the actual soil SOM content and the model's predicted values. It is negatively correlated with model stability, as shown in Equation (17). MAE measures the accuracy of the predictive model. The smaller the MAE, the higher the model's accuracy, as shown in Equation (18). MBE measures the tendency to overestimate or underestimate parameter values. A positive bias means the error is overestimated, while a negative bias means the error is underestimated. This metric quantifies overall bias and captures the average deviation in predictions, as shown in Equation (19). MARE measures the degree of deviation between the predicted and actual values by calculating the relative error between them and taking the average of their absolute values, as shown in Equation (20).

The Relative Percent Difference (RPD) reflects both the reliability and accuracy of a model, and it is widely used by many soil researchers [46,47]. However, there is no consensus on the threshold values based on statistical studies. For example, Lei et al. defined the RPD threshold for evaluating SOM prediction models as follows: RPD < 2 indicates that the model does not meet prediction requirements, while RPD > 3 indicates that the

model performs excellently [48]. In contrast, Shi et al. defined the two thresholds as 1.4 and 2 [49]. Additionally, the standard deviation (SD) used in RPD cannot accurately describe the distribution range of a population in a skewed distribution. Spectral data are influenced by various physicochemical properties of the sampling points, and ideal normal distribution conditions are rare, further affecting the evaluation of models using the RPD metric. RPD is shown in Equation (20).

Bellon Maurel et al. [50] proposed a new evaluation metric based on quartiles: the Ratio of Performance to Interquartile Distance (RPIQ), which better describes the distribution of a population regardless of its shape. Quartiles play a key role in population distribution: Q1 indicates that 25% of the samples are below this value, while Q3 indicates that 75% of the samples are below this value. The interquartile range (IQ) is calculated as Q3 minus Q1, representing the span of the middle 50% of the population around the median. Compared to RPD, RPIQ is not affected by the SD and addresses the existing limitations of using near-infrared spectroscopy for soil characterization. RPIQ is shown in Equation (21).

$$R^2 = 1 - \sum_{i=1}^{n} \frac{(x_i - \widehat{x_i})^2}{(x_i - \overline{x_i})^2} \tag{15}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \widehat{x_i})^2}{n}} \tag{16}$$

$$MAE = \frac{\sum_{i=1}^{n}|x_i - \widehat{x_i}|}{n} \tag{17}$$

$$MARE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - \widehat{x_i}}{x_i}\right| \tag{18}$$

$$MBE = \frac{\sum_{i=1}^{n}(x_i - \widehat{x_i})}{n} \tag{19}$$

$$RPD = \frac{SD}{RMSE} \tag{20}$$

$$RPIQ = \frac{Q3 - Q1}{RMSE} \tag{21}$$

where n represents the total number of samples. x_i, $\widehat{x_i}$, and $\overline{x_i}$ denote the actual value, predicted value, and mean value of the soil SOM content for the i-th sample, respectively.

3. Results

3.1. Statistical Characteristics of Soil Organic Matter Content

In this study, we collected 206 samples according to the method described in Section 2.1. The SPXY algorithm was used to divide the dataset into training and validation sets, resulting in 164 training samples and 42 validation samples in a 4:1 ratio. The statistical characteristics of soil organic matter content are shown in Table 2. The coefficient of variation, calculated from the standard deviation and mean, reflects the degree of data dispersion. The coefficients of variation for the three types of samples indicate moderate dispersion, suggesting the need for appropriate preprocessing methods to enhance model performance.

Table 2. Statistics on the characteristics of soil organic matter content.

Sample Classification	Sample Size	Mean ($g \cdot kg^{-1}$)	Minimum Value ($g \cdot kg^{-1}$)	Maximum Value ($g \cdot kg^{-1}$)	Standard Deviation ($g \cdot kg^{-1}$)	Coefficient of Variation (%)
Total Samples	206	25.72	4.26	80.04	13.57	52.75
Training Set	164	26.03	4.26	80.04	14.04	53.95
Validation Set	42	24.50	8.23	71.15	11.59	47.30

The spectral reflectance corresponding to different organic matter contents measured in this study is shown in Figure 5. The overall reflectance trends for different SOM contents are generally similar, with reflectance inversely related to SOM content. In the 400–800 nm visible light range, SOM exhibits strong absorption, while in the 800–2400 nm near-infrared range, SOM shows strong reflectance. An absorption valley influenced by iron oxide appears at a wavelength of 950 nm, and absorption valleys influenced by moisture appear at wavelengths of 1500 nm, 1900 nm, and 2350 nm.

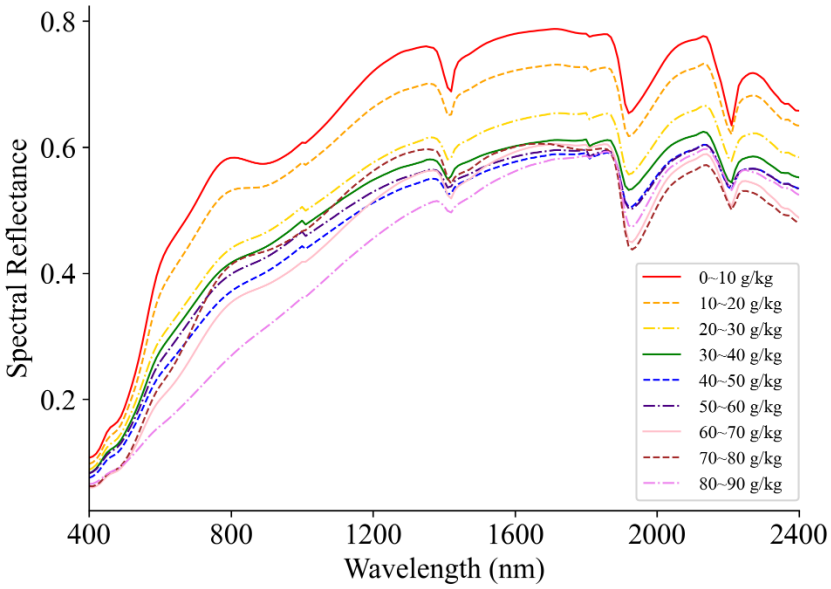

Figure 5. Different organic matter contents correspond to reflectivity.

3.2. Impact of Different Preprocessing Methods on Data and Model Accuracy

3.2.1. Impact of Different Preprocessing Methods on Data

Extracting spectral feature bands is fundamental to model construction, as it accurately reflects spectral characteristics closely related to soil organic matter (SOM). The preprocessing process highlights spectral information characteristic of SOM, effectively removing redundant information and enhancing the correlation between data and SOM. This strategy simplifies data processing complexity, strengthens the model's ability to invert SOM content, and consequently improves prediction accuracy and reliability [51]. This study employed three denoising methods: no denoising (N), Savitzky–Golay filter denoising (SG), and discrete wavelet transform denoising (DWT). Additionally, nine mathematical transformation methods were used: raw spectral reflectance (R), first derivative (1DR), second derivative (2DR), multiplicative scatter correction (MSC), standard normal variate transformation (SNV), logarithmic transformation (logR), first derivative of the logarithm ((LogR)'), reciprocal transformation (1/R), and first derivative of the reciprocal ((1/R)'). The combinations are shown in Table 3.

Table 3. Different pretreatment combinations.

Denoising Methods		N	SG	DWT
Mathematical Transformations	R	N-R	SG-R	DWT-R
	1DR	N-1DR	SG-1DR	DWT-1DR
	2DR	N-2DR	SG-2DR	DWT-2DR
	MSC	N-MSC	SG-MSC	DWT-MSC
	SNV	N-SNV	SG-SNV	DWT-SNV
	LogR	N-LogR	SG-LogR	DWT-LogR
	(LogR)′	N-(LogR)′	SG-(LogR)′	DWT-(LogR)′
	1/R	N-1/R	SG-1/R	DWT-1/R
	(1/R)′	N-(1/R)′	SG-(1/R)′	DWT-(1/R)′

Correlation analysis is a statistical method used to explore the degree of association between two or more variables [52]. The Pearson correlation coefficient, as an indicator of correlation analysis, reflects the relationship between adjacent spectral bands and soil organic matter (SOM) content. Based on this, a correlation heatmap is constructed to reveal the relationship and degree of influence between band reflectance and SOM content under different preprocessing methods. Using the data preprocessed according to the combinations in Table 3, a Pearson correlation coefficient heatmap of spectral band reflectance and SOM content is shown in Figure 6. In the heatmap, Pearson correlation coefficients are represented by different colors. The heatmaps for the mathematical transformations R, logR, and 1/R show relatively uniform correlation changes. In contrast, MSC and SNV exhibit diverse correlation changes, but their color distribution is concentrated, indicating less noticeable changes in individual spectral data. Observations reveal that the heatmaps for the four mathematical transformations 1DR, 2DR, (LogR)′, and (1/R)′ show more diverse correlation changes and more uniform color distribution. Data based on these four transformations effectively present the correlation characteristics between spectral band reflectance and SOM content, providing valuable insights for subsequent modeling.

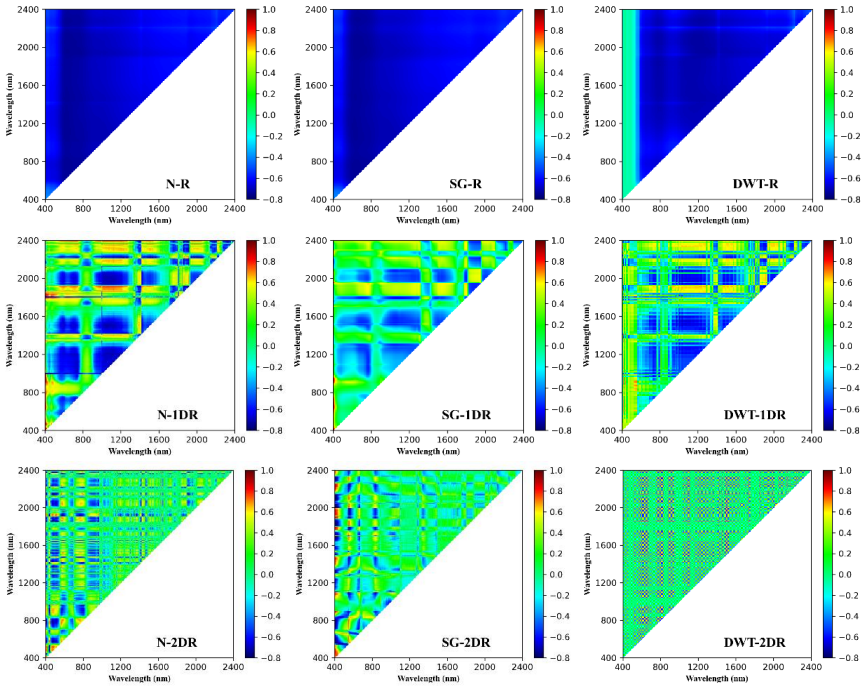

Figure 6. *Cont*.

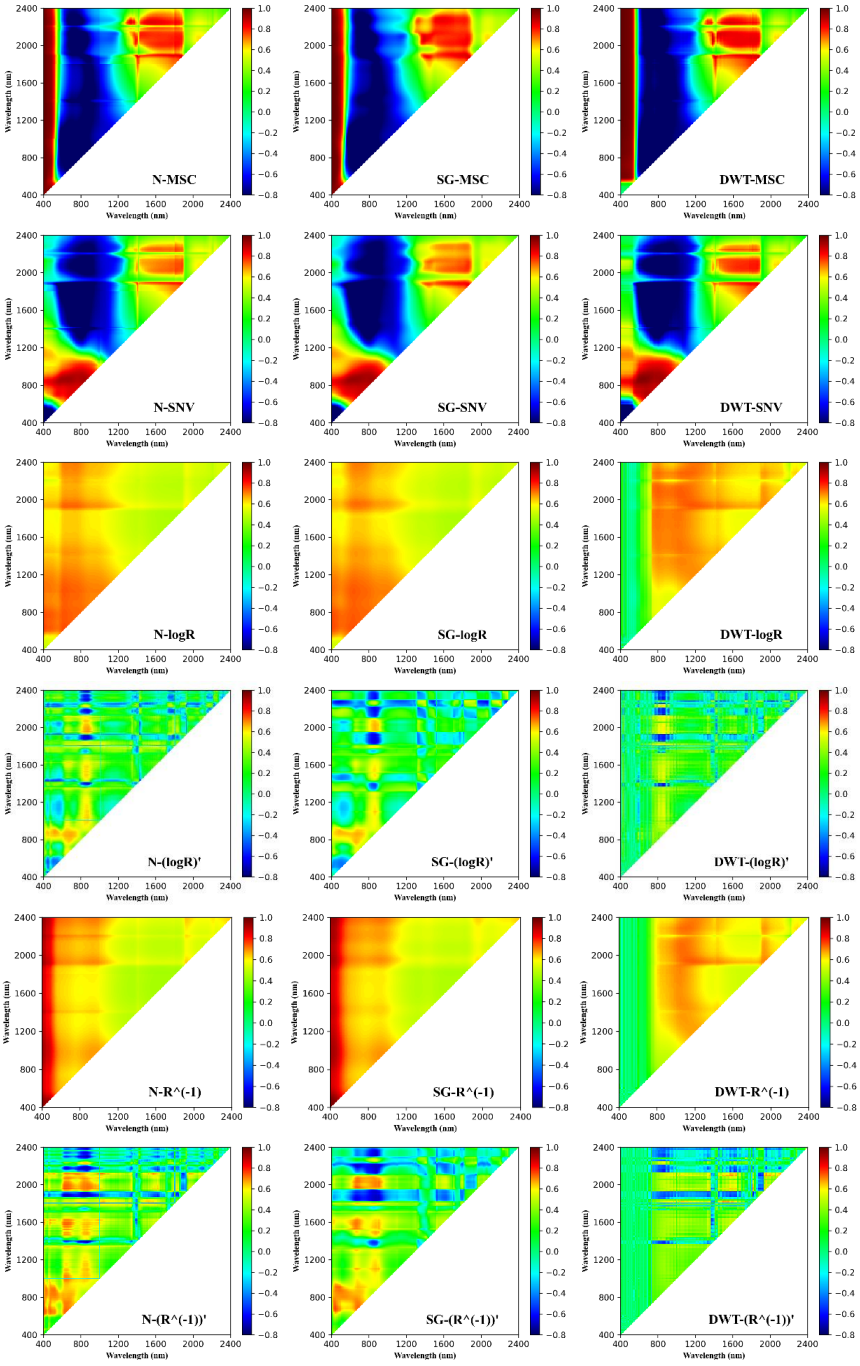

Figure 6. Heat map of Pearson correlation coefficient under different pretreatment methods.

3.2.2. Impact of Different Preprocessing Methods on Model Accuracy

By comparing the heatmaps, it was found that the preprocessing methods for the four mathematical transformations 1DR, 2DR, (LogR)', and (1/R)' were superior. Next, we will compare the modeling results of these four mathematical transformations under different denoising methods. This study employs two machine learning methods, SVM and PLSR, and two deep learning methods, BP and CNN, for modeling. The modeling results are shown in Table 4.

Table 4. Different preprocessing modeling results.

Preprocessing Methods		Modeling Methods	Validation Set					
			RMSE	R^2	MAE	MARE	MBE	RPIQ
N-	1DR	SVM	5.247	0.790	4.167	0.182	1.035	3.904
		PLSR	5.768	0.746	4.938	0.215	0.247	4.035
		BP	5.991	0.726	4.779	0.211	0.786	4.514
		CNN	5.736	0.749	4.633	0.221	1.103	5.338
	2DR	SVM	6.427	0.685	4.948	0.216	1.918	5.203
		PLSR	6.670	0.661	5.541	0.263	1.412	4.069
		BP	7.052	0.621	5.628	0.279	1.936	5.161
		CNN	6.286	0.699	4.983	0.236	1.181	4.647
	(LogR)′	SVM	4.426	0.851	3.535	0.172	1.558	3.866
		PLSR	4.490	0.846	3.644	0.174	1.095	3.034
		BP	5.575	0.747	4.637	0.205	1.678	4.318
		CNN	5.652	0.756	4.745	0.235	2.215	3.782
	(1/R)′	SVM	4.281	0.860	3.448	0.159	0.613	3.290
		PLSR	4.181	0.867	3.467	0.151	0.580	3.719
		BP	5.860	0.738	4.535	0.233	1.809	3.831
		CNN	5.791	0.744	4.609	0.214	0.890	5.210
SG-	1DR	SVM	5.165	0.797	4.161	0.183	0.469	4.255
		PLSR	6.221	0.705	5.080	0.268	0.354	3.478
		BP	4.471	0.848	3.574	0.159	0.130	2.401
		CNN	5.191	0.794	4.096	0.195	0.447	3.697
	2DR	SVM	4.100	0.809	4.135	0.190	0.832	4.583
		PLSR	5.472	0.772	4.408	0.226	0.488	3.617
		BP	5.887	0.736	4.946	0.226	0.637	4.804
		CNN	5.779	0.745	4.848	0.245	1.049	3.384
	(LogR)′	SVM	4.800	0.825	4.000	0.192	1.149	4.284
		PLSR	6.485	0.679	5.316	0.271	1.259	5.709
		BP	5.693	0.753	4.302	0.190	0.451	3.708
		CNN	5.482	0.771	4.449	0.202	1.540	4.598
	(1/R)′	SVM	4.745	0.828	3.729	0.172	0.480	4.620
		PLSR	5.447	0.774	4.439	0.210	0.535	3.199
		BP	4.887	0.818	3.826	0.183	1.154	3.790
		CNN	5.578	0.763	4.096	0.195	1.236	4.615
DWT-	1DR	SVM	5.034	0.806	3.945	0.170	0.444	4.310
		PLSR	5.834	0.740	4.445	0.242	0.565	5.071
		BP	5.290	0.787	4.187	0.215	1.407	4.475
		CNN	5.396	0.778	4.468	0.222	1.595	4.246
	2DR	SVM	5.365	0.780	4.169	0.184	1.407	5.302
		PLSR	6.618	0.666	5.371	0.293	1.379	4.121
		BP	6.327	0.695	4.894	0.240	2.949	4.638
		CNN	6.197	0.707	5.158	0.241	2.960	4.690
	(LogR)′	SVM	4.886	0.818	4.012	0.195	1.323	3.416
		PLSR	5.227	0.792	4.174	0.203	1.267	4.083
		BP	6.164	0.710	4.757	0.206	0.311	5.208
		CNN	5.573	0.763	4.428	0.205	1.259	3.611
	(1/R)′	SVM	4.792	0.825	3.984	0.182	0.131	3.858
		PLSR	4.995	0.810	3.840	0.189	1.267	4.334
		BP	5.601	0.761	4.520	0.205	2.239	3.953
		CNN	5.713	0.751	4.288	0.199	1.268	4.946

Among the 16 combinations of four mathematical transformations and four modeling methods, the combinations with SG denoising outperformed the other two denoising methods in several metrics: RMSE was lower in 9 out of 16 cases, R^2 was higher in 10 out of 16 cases, MAE was lower in 8 out of 16 cases, MARE was lower in 8 out of 16 cases, MBE was closer to 0 in 9 out of 16 cases, and RPIQ was lower in 10 out of 16 cases. Therefore, for the soil spectral data in the experimental area, SG denoising is superior to N and DWT denoising methods. The comparison of the four mathematical transformations and four modeling methods after SG denoising is shown in Figure 7. Among these, 1DR-BP achieved the best results in R^2, MAE, MARE, MBE, and RPIQ. Other 1DR preprocessing results were also excellent, such as 1DR-CNN, which performed best in RMSE, MAE, MARE, MBE, and RPIQ among the four mathematical transformations using CNN modeling, and 1DR-SVM, which performed best in MARE, MBE, and RPIQ among the four mathematical transformations using SVM modeling. This indicates that 1DR preprocessing effectively reflects spectral data characteristics. The experimental results on the impact of different preprocessing methods on data and model accuracy show that SG-1DR preprocessing is the best preprocessing method. This conclusion is consistent with the findings of Nan Feng et al. [53] and Zhang et al. [25], as well as with Chen et al. [54], who found the optimal preprocessing method for soils in the northwestern part of Yunnan Province, which is close to the experimental area. This effectively highlights the correlation between spectral data and SOM content and improves model generalization ability.

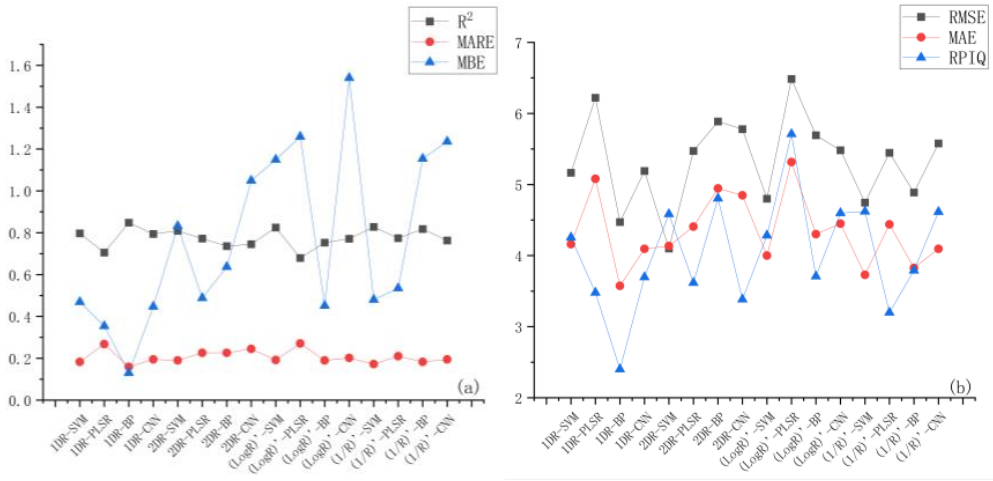

Figure 7. Different modeling evaluation indexes under SG noise reduction. (**a**) shows the details of R^2, MARE and MBE. (**b**) shows the details of RMSE, MAE and RPIQ.

3.3. CNN-SVM Prediction Model Improved by the Black-Winged Kite Algorithm

3.3.1. Training Process

This study employs the BKA for optimization, leveraging its robust search capabilities and balance between local and global optima. The BKA iterates within the CNN-SVM network to identify three optimal parameters: the best mini-batch size (best1), the best initial learning rate (best2), and the best learning rate decay factor (best3) [55]. The BKA iterates according to the parameters in Table 5, with the loss function recording the match between predicted and actual values. As shown in Figure 8, the loss fluctuates significantly in the early stages but generally decreases, stabilizing after the 60th generation, indicating that the model converges to the true values.

Table 5. BKA detailed parameters.

Definition		Parameters
Population Size		60
Maximum Iterations		250
Dimension		3
Upper Bound	best1	512
	best2	5×10^{-2}
	best3	1×10^{-1}
Lower Bound	best1	128
	best2	1×10^{-3}
	best3	1×10^{-4}

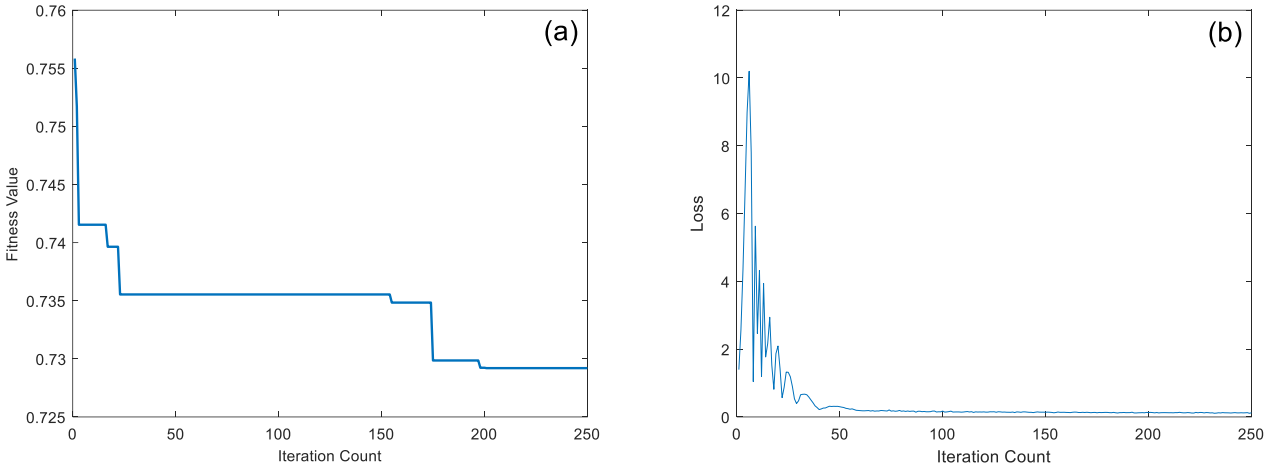

Figure 8. BKA iteration. (**a**) shows the change of the fitness value. (**b**) shows the change of the Loss.

After the BKA iterations, the optimal parameters selected were best1 = 164, best2 = 0.0720, and best3 = 0.0026. These three parameters were incorporated into the CNN network training. The CNN was set with a maximum of 800 training iterations, a mini-batch size of best1, an initial learning rate of best2, a learning rate decay factor of best3, and an L2 regularization coefficient of 0.1. The training set was shuffled for each training session. The specific network structure is detailed in Table 6.

Table 6. CNN structure settings.

Definition	Parameters
Input	199, 1×1
Conv1	16, 3×1
BN1	-
Relu1	-
Maxpool1	2×1
Conv2	32, 3×1
BN2	-
Relu2	-
Maxpool2	2×1
Dropout	0.5
Fc	1

The optimal parameters were obtained by traversing the values of the penalty factor (c) and the Gamma function (g) using the grid search cross-validation method. The initial

value of the penalty factor (c) for the SVM model was set to 0.5 with a step size of 0.1; the initial value of the Gamma function (g) was set to 0.05 with a step size of 0.01, and the SG-1DR data were trained. As shown in Figure 9, the model results were optimal when (c) was 1.1 and (g) was 0.11. The final SVM parameter settings were based on the optimal (c) and (g) values and the commonly used parameters for SVM modeling of soil spectral data [22,23]. The SVM type selected was e-SVR, with a penalty factor (c) of 1.1 and a loss function (p) of 0.01; the kernel function was the RBF function, and the Gamma function in the kernel function was set to 0.11.

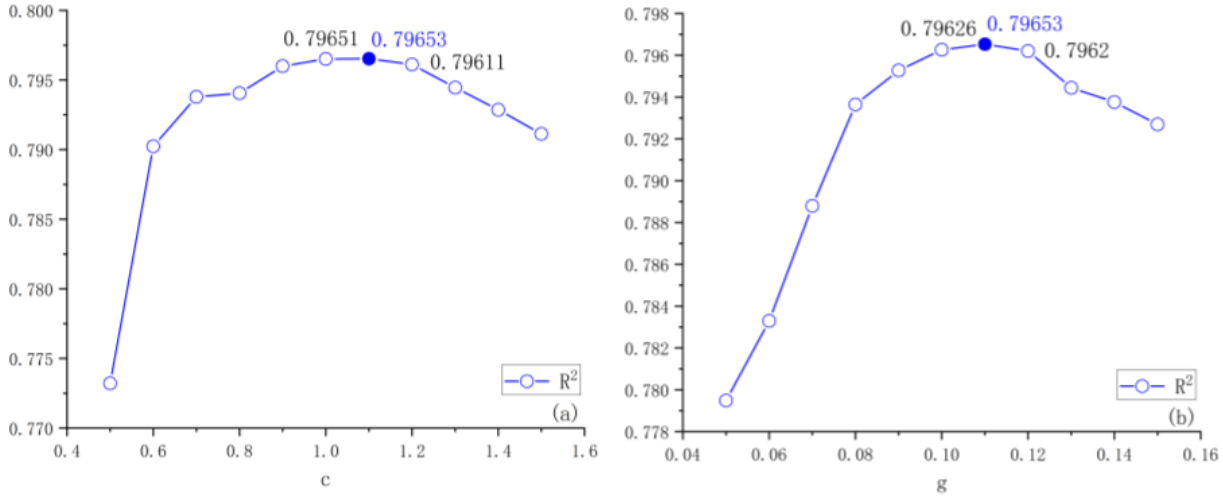

Figure 9. SVM parameter optimization process. (**a**) shows the optimization situation of parameter c. (**b**) shows the optimization situation of parameter g.

3.3.2. Model Performance Comparison

To validate the effectiveness of the model optimization results, we uniformly used the training and test samples obtained through SG-1DR preprocessing as divided in Table 2. This study selected five modeling methods—CNN, SVM, CNN-SVM, BKA-CNN, and BKA-CNN-SVM—to compare the prediction accuracy of SOM content. The modeling results are shown in Table 7. To visually understand the differences between the predicted and actual values for each model, regression scatter plots of the predicted and actual values for each model's test set were drawn, as shown in Figure 10. The closer the fitted regression line is to the 1:1 line, the more accurate the model's predictions. The BKA-CNN-SVM model was the best in all four evaluation metrics, with its scatter points more evenly distributed around the 1:1 line. The RMSE of the validation set was 3.042, which decreased by 2.022, 2.123, 1.599, and 1.702 compared to other models. The R^2 was 0.930, which increased by 12.3%, 13.3%, 9.2%, and 9.9% compared to other models. The MAE was 2.254, which decreased by 1.716, 2.137, 1.487, and 1.221 compared to other models. The MARE was 0.1, which decreased by 0.077, 0.083, 0.078, and 0.063 compared to other models. The RPIQ was 1.436, which decreased by 3.560, 2.819, 2.316, and 0.876 compared to other models. It can be seen that the BKA-CNN-SVM model has a high degree of matching between predicted and actual values, indicating that the addition of SVM improves the generalization ability of regression predictions. The Black-winged Kite Algorithm significantly enhances the accuracy of the CNN-SVM network, showing a more significant and positive impact on the effective management of forest soil in the study area and enhancing the scientific and practical effectiveness of soil improvement strategies.

Table 7. Modeling results based on SG-1DR preprocessed data.

Modeling Methods	Training Set						Validation Set					
	RMSE	R^2	MAE	MARE	MBE	RPIQ	RMSE	R^2	MAE	MARE	MBE	RPIQ
CNN	2.148	0.976	1.605	0.073	0.073	1.776	5.064	0.807	3.970	0.177	0.229	4.996
SVM	3.505	0.937	0.713	0.079	0.005	0.655	5.165	0.797	4.391	0.183	0.469	4.255
CNN-SVM	2.307	0.973	1.327	0.062	0.305	0.350	4.641	0.838	3.741	0.178	1.372	3.752
BKA-CNN	1.756	0.984	1.279	0.069	0.347	0.366	4.744	0.831	3.475	0.163	1.036	2.312
BKA-CNN-SVM	1.609	0.987	1.287	0.066	0.348	0.836	3.042	0.930	2.254	0.100	0.890	1.436

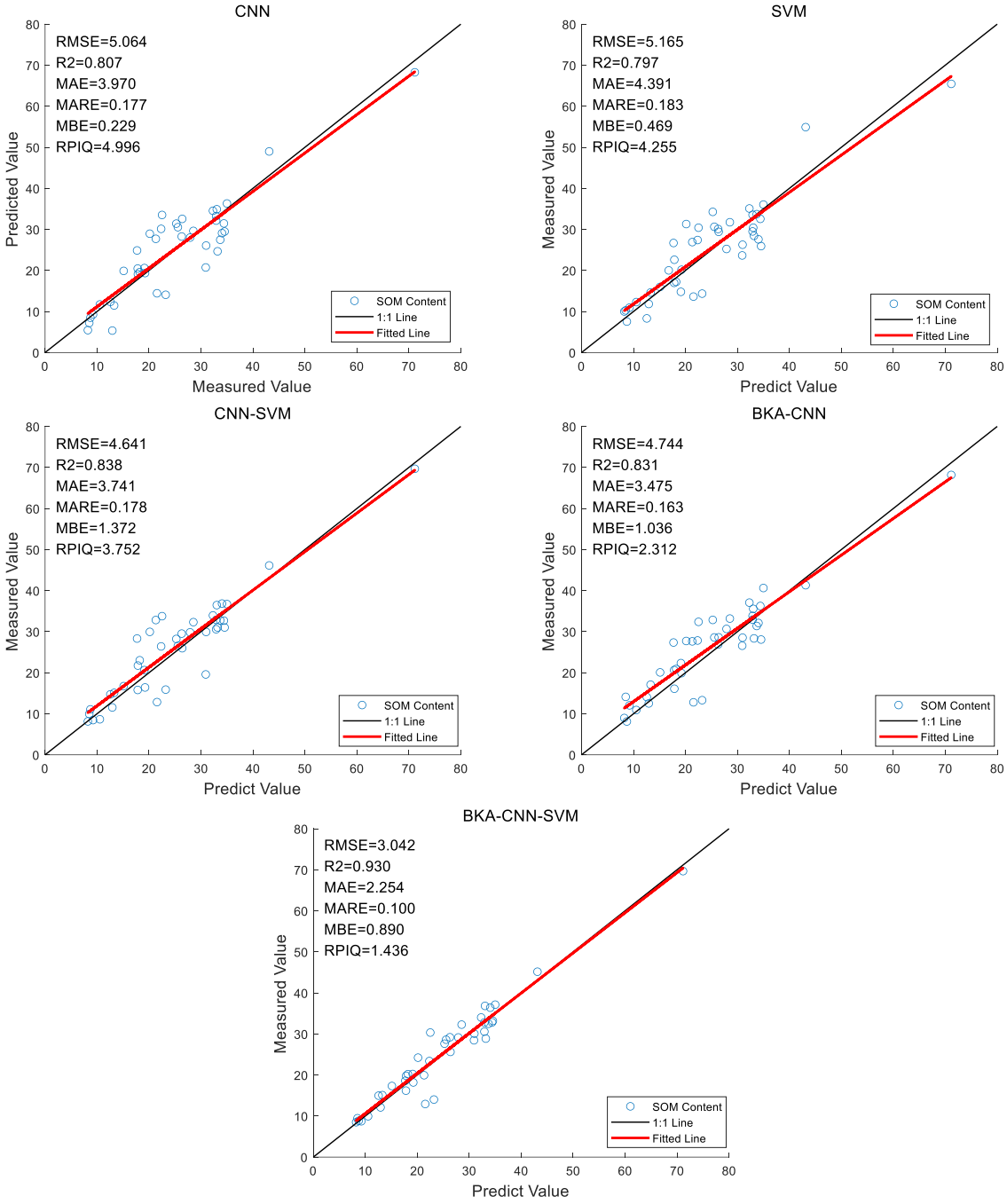

Figure 10. Scatter plots of predicted and measured SOM content.

4. Discussion

This study analyzed soil samples from the state-owned Yachang and Huangmian Forest Farms in Guangxi to determine the effectiveness of the SG-1DR preprocessing

method in the experimental area. It also constructed a CNN-SVM model optimized by the Black-winged Kite Algorithm for SOM prediction. This achievement provides an important reference for SOM determination in this area and offers a new method for forest soil monitoring.

In terms of preprocessing, due to different natural conditions, the research results only represent the optimal preprocessing effect in the experimental area and provide a reference method for similar regions. Secondly, although a wide variety of mathematical transformations and models were used in the experiment, it is still possible that not all potentially effective methods were covered, leaving room for the discovery of better combinations. Due to the high difficulty of sample collection, most researchers used a limited number of samples from the experimental area [56,57]. Although it is not yet concluded whether sample size is a key factor affecting experimental results [58], with sufficient sample size, it is possible to try using multivariate feature selection [59,60] and the method of concatenating original and preprocessed bands to form a new dataset [61]. More experiments can explore additional possibilities for the optimal preprocessing methods for the experimental area data.

Regarding predictive models, the results show that CNN can effectively extract rich feature information from spectral data, and SVM can handle the emergence of multicollinearity issues contained in the data processed by CNN. Currently, many researchers use traditional machine learning methods alone, which may struggle to extract effective spectral features when faced with redundant soil spectral data in the experimental area [30,62]. The complexity and high dimensionality of such data often require more advanced and sophisticated methods, such as the combination of CNN and SVM used in this study or the establishment of more diverse model structures like stacked machine-learning methods [45]. Additionally, some researchers have deployed predictive models in remote sensing to achieve real-time SOM content prediction [63,64]. This approach has great potential as it allows for large-scale and continuous monitoring of soil conditions, providing valuable information for timely decision-making in agriculture and forestry. There are also studies that combine remote sensing images with field measurement data for model training, yielding good results [65]. Combining spectral data with image data can improve the model's generalization ability, but due to the differences in the forms of these two types of data, feature extraction may not be optimal, requiring further experiments tailored to the specific conditions of the experimental area [66].

5. Conclusions

The conclusions of this study are as follows:

1. Among the 27 preprocessing combinations used in this study, the SG-1DR method effectively extracts data features and reduces noise interference. The heatmaps for the four mathematical transformations—1DR, 2DR, (LogR)', and (1/R)'—exhibit rich correlation variations and relatively uniform color distributions. Data based on these four transformations effectively present the correlation characteristics between spectral band reflectance and SOM content. Through modeling comparisons, it was found that the first derivative transformation after SG denoising yielded the best results, demonstrating the highest feature expression capability for spectral data.
2. In the preprocessing and modeling comparison experiments, it was found that deep learning methods have a superior ability to extract spectral data features compared to machine learning methods.
3. After combining CNN with SVM, the evaluation metrics improved compared to using CNN alone. Specifically, RMSE decreased by 0.423, R^2 increased by 3.1%, MAE decreased by 0.229, MARE differed by only 0.001, and RPIQ decreased by 1.244. This

indicates that for predicting SOM, it is feasible to use CNN for feature extraction followed by SVM for classification and regression.

4. The CNN-SVM network improved by the Black-winged Kite Algorithm showed significant enhancements compared to the original CNN-SVM network: RMSE decreased by 1.599, R^2 increased by 9.2%, MAE decreased by 1.487, MARE decreased by 0.078, MBE decreased by 0.482, and RPIQ decreased by 2.316. This makes it the best among the five compared networks. Therefore, the BKA-CNN-SVM model is highly effective in predicting the SOM content of soil in Guangxi forest farms.

Author Contributions: Conceptualization, L.X. and Y.D.; methodology, L.X.; software, Y.S.; validation, Y.D., L.X. and Y.S.; formal analysis, Y.D.; investigation, Y.S.; resources, L.X.; data curation, Y.D.; writing—original draft preparation, L.X.; writing—review and editing, L.X.; visualization, Y.D.; supervision, Y.D.; project administration, Y.D.; funding acquisition, Y.D. and Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Guangxi Natural Internal Medicine Foundation project (GuikeAB24010338), the central government guides local science and technology development fund projects (GuikeZY22096012), and the National Natural Science Foundation of China (32360374).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this article are not readily available because they come from state-owned forests and are restricted by privacy agreements. If you have relevant needs, you can contact the author to obtain access. Requests to access the datasets should be directed to 2002078@glut.edu.cn.

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

References

1. Kögel-Knabner, I.; Rumpel, C. Chapter One—Advances in Molecular Approaches for Understanding Soil Organic Matter Composition, Origin, and Turnover: A Historical Overview. *Adv. Agron.* **2018**, *149*, 1–48.
2. Pribyl, D.W. A critical review of the conventional SOC to SOM conversion factor. *Geoderma* **2010**, *156*, 75–83. [CrossRef]
3. Zhang, X.; Yao, Y.; Yan, X. Research progress on prediction of soil organic matter content by mid-infrared spectroscopy. *Soil Fertil. Sci. China* **2021**, *4*, 327–336.
4. Zhao, L.; Liu, X.; Wang, Y.; Ren, T. Thermal analysis determining soil organic matter content and thermal stability. *Trans. Chin. Soc. Agric. Eng.* **2016**, *32*, 105–114.
5. ACambule, H.; Rossiter, D.G.; Stoorvogel, J.J.; Smaling, E.M.A. Soil organic carbon stocks in the Limpopo National Park, Mozambique: Amount, spatial distribution and uncertainty. *Geoderma* **2014**, *213*, 46–56. [CrossRef]
6. Sun, J.; Wang, G.; Zhang, H.; Xia, L.; Zhao, W.; Guo, Y.; Sun, X. Detection of fat content in peanut kernels based on chemometrics and hyperspectral imaging technology. *Infrared Phys. Technol.* **2020**, *105*, 103226. [CrossRef]
7. Lu, P.; Wang, L.; Niu, Z.; Li, L.; Zhang, W. Prediction of soil properties using laboratory VIS–NIR spectroscopy and Hyperion imagery. *J. Geochem. Explor.* **2013**, *132*, 26–33. [CrossRef]
8. Conforti, M.; Buttafuoco, G.; Leone, A.P.; Aucelli, P.P.; Robustelli, G.; Scarciglia, F. Studying the relationship between water-induced soil erosion and soil organic matter using Vis–NIR spectroscopy and geomorphological analysis: A case study in southern Italy. *Catena* **2013**, *110*, 44–58. [CrossRef]
9. Caporaso, N.; Whitworth, M.B.; Fisk, I.D. Near-infrared spectroscopy and hyperspectral imaging for non-destructive quality assessment of cereal grains. *Appl. Spectrosc. Rev.* **2018**, *53*, 667–687. [CrossRef]
10. Chen, H.; Chen, A.; Xu, L.; Xie, H.; Qiao, H.; Lin, Q.; Cai, K. A deep learning cnn architecture applied in smart near-infrared analysis of water pollution for agricultural irrigation resources. *Agric. Water Manag.* **2020**, *240*, 106303–106310. [CrossRef]
11. Zeng, J.; Guo, Y.; Han, Y.; Li, Z.; Yang, Z.; Chai, Q.; Wang, W.; Zhang, Y.; Fu, C. A review of the discriminant analysis methods for food quality based on near-infrared spectroscopy and pattern recognition. *Molecules* **2021**, *26*, 749. [CrossRef] [PubMed]
12. Zhao, M.; Gao, Y.; Lu, Y.; Wang, S. Hyperspectral Modeling of Soil Organic Matter Based on Characteristic Wavelength in East China. *Sustainability* **2022**, *14*, 8455. [CrossRef]

13. Gholizadeh, A.; Soom, M.A.; Saberioon, M.; Farming, S. Visible and near infrared reflectance spectroscopy to determine chemical properties of paddy soils. *J. Food Agric. Environ.* **2013**, *11*, 859–866.
14. Hau, N.-X.; Tuan, N.-T.; Trung, L.-Q.; Chi, T.-T. Estimation of soil organic carbon content using visible and near-infrared spectroscopy in the Red River Delta, Vietnam. *Chemom. Intell. Lab. Syst.* **2024**, *255*, 105253. [CrossRef]
15. Cao, L.; Sun, M.; Yang, Z.; Jiang, D.; Yin, D.; Duan, Y. A Novel Transformer-CNN Approach for Predicting Soil Properties from LUCAS Vis-NIR Spectral Data. *Agronomy* **2024**, *14*, 1998. [CrossRef]
16. Choi, J.-Y.; Kim, H.-C.; Moon, K.-D. Geographical origin discriminant analysis of Chia seeds (*Salvia hispanica* L.) using hyperspectral imaging. *J. Food Compos. Anal.* **2021**, *101*, 103916. [CrossRef]
17. Terra, F.S.; Demattê, J.A.M.; Rossel, R.A.V. Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis–NIR and mid-IR reflectance data. *Geodermas* **2015**, *255–256*, 81–93. [CrossRef]
18. Shen, L.; Gao, M.; Yan, J.; Li, Z.-L.; Leng, P.; Yang, Q.; Duan, S.-B. Hyperspectral Estimation of Soil Organic Matter Content Using Different Spectral Preprocessing Techniques and Plsr Method. *Remote Sens.* **2020**, *12*, 1206. [CrossRef]
19. Mishra, P.; Lohumi, S. Improved prediction of protein content in wheat kernels with a fusion of scatter correction methods in NIR data modelling. *Biosyst. Eng.* **2021**, *203*, 93–97. [CrossRef]
20. Zhang, Y.; Biswas, A.; Ji, W.; Adamchuk, V.I. Depth-Specific Prediction of Soil Properties In Situ using vis-NIR Spectroscopy. *Soil Sci. Soc. Am. J.* **2017**, *81*, 993–1004. [CrossRef]
21. Bao, Q.-l.; Ding, J.-l.; Wang, J.-z.; Cai, L.-h. Hyperspectral detection of soil organic matter content based on random forest algorithm. *Arid. Land Geogr.* **2019**, *42*, 1404–1414.
22. Zhang, J.; Xi, L.; Yang, X.; Xu, X.; Guo, W.; Cheng, T.; Ma, X. Construction of hyperspectral estimation model for organic matter content in Shajiang black soil. *Trans. Chin. Soc. Agric. Eng.* **2020**, *36*, 135–141.
23. Carvalho, J.K.; Moura-Bueno, J.M.; Ramon, R.; Almeida, T.F.; Naibo, G.; Martins, A.P.; Santos, L.S.; Gianello, C.; Tiecher, T. Combining different pre-processing and multivariate methods for prediction of soil organic matter by near infrared spectroscopy (NIRS) in Southern Brazil. *Geoderma Reg.* **2022**, *29*, e00530. [CrossRef]
24. Fidêncio, P.H.; Poppi, R.J.; de Andrade, J.C.; Cantarella, H. Determination of Organic Matter in Soil Using near-Infrared Spectroscopy and Partial Least Squares Regression. *Commun. Soil Sci. Plant Anal.* **2002**, *33*, 1607–1615. [CrossRef]
25. Zhang, S.; Lu, X.; Nie, G.-g.; Li, Y.-r.; Shao, Y.-t.; Tian, Y.-q.; Fan, L.-q.; Zhang, Y.-j. Estimation of Soil Organic Matter in Coastal Wetlands by SVM and BP Based on Hyperspectral Remote Sensing. *Spectrosc. Spectr. Anal.* **2020**, *40*, 556–561.
26. Zhang, L.; Shao, Z.; Liu, J.; Cheng, Q. Deep learning based retrieval of forest aboveground biomass from combined LiDAR and landsat 8 data. *Remote Sens.* **2019**, *11*, 1459. [CrossRef]
27. Shahrayini, E.; Noroozi, A.A.; Eghbal, M.K. Prediction of Soil Properties by Visible and Near-Infrared Reflectance Spectroscopy. *Eurasian Soil Sci.* **2020**, *53*, 1760–1772. [CrossRef]
28. Rossel, R.A.V.; Behrens, T. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* **2010**, *158*, 46–54. [CrossRef]
29. Haghi, R.K.; Pérez-Fernández, E.; Robertson, A.H.J. Prediction of various soil properties for a national spatial dataset of Scottish soils based on four different chemometric approaches: A comparison of near infrared and mid-infrared spectroscopy. *Geoderma* **2021**, *396*, 115071. [CrossRef]
30. Li, H.; Ju, W.; Song, Y.; Cao, Y.; Yang, W.; Li, M. Soil organic matter content prediction based on two-branch convolutional neural network combining image and spectral features. *Comput. Electron. Agric.* **2024**, *217*, 108561. [CrossRef]
31. Hu, W.; Huang, Y.; Wei, L.; Zhang, F.; Li, H. Deep Convolutional Neural Networks for Hyperspectral Image Classification. *J. Sens.* **2015**, *2015*, 258619. [CrossRef]
32. Ladoni, M.; Bahrami, H.A.; Alavipanah, S.K.; Norouzi, A.A. Estimating soil organic carbon from soil reflectance: A review. *Precis. Agric.* **2010**, *11*, 82–99. [CrossRef]
33. Chen, H.; Zhao, G.; Zhang, X.; Wang, R.; Sun, L.; Chen, J. Improving estimation precision of soil organic matter content by removing effect of soil moisture from hyperspectra. *Trans. Chin. Soc. Agric. Eng.* **2014**, *30*, 91–100.
34. Li, T.; Dong, X.; Lin, J.; Peng, Y.; Li, T.; Dong, X.; Lin, J.; Peng, Y.; Li, T.; Dong, X.; et al. A Transformer-Cnn Parallel Network for Image Guided Depth Completion. *Pattern Recognit.* **2024**, *150*, 110305. [CrossRef]
35. Sun, W.; Chang, L.C.; Chang, F.J. Deep Dive into Predictive Excellence: Transformer's Impact on Groundwater Level Prediction. *J. Hydrol.* **2024**, *636*, 131250. [CrossRef]
36. Omondiagbe, O.P.; Lilburne, L.; Licorish, S.A.; MacDonell, S.G. Soil Texture Prediction with Automated Deep Convolutional Neural Networks and Population-Based Learning. *Geoderma* **2023**, *436*, 116521. [CrossRef]
37. Cao, Y.; Yang, W.; Li, H.; Zhang, H.; Li, M. Development of a vehicle-mounted soil organic matter detection system based on near-infrared spectroscopy and image information fusion. *Meas. Sci. Technol.* **2024**, *35*, 045501. [CrossRef]
38. Romsonthi, C.; Tawornpruek, S.; Watana, S. In situ near-infrared spectroscopy for soil organic matter prediction in paddy soil, Pasak watershed, Thailand. *Plant Soil Environ.* **2018**, *64*, 70–75. [CrossRef]

39. Zhu, H.; Hu, W.; Jing, Y.; Cao, Y.; Bi, R.; Yang, W. Soil organic carbon prediction based on scale-specific relationships with environmental factors by discrete wavelet transform. *Geoderma* **2018**, *330*, 9–18. [CrossRef]
40. Donoho, D.L.; Johnstone, I.M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **1994**, *81*, 425–455. [CrossRef]
41. Li, C.; Chen, H.; Zhang, Y.; Hong, S.; Ai, W.; Mo, L. Improvement of NIR prediction ability by dual model optimization in fusion of NSIA and SA methods. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2022**, *276*, 121247. [CrossRef] [PubMed]
42. Tian, H.; Zhang, L.; Li, M.; Wang, Y.; Sheng, D.; Liu, J.; Wang, C. Weighted SPXY method for calibration set selection for composition analysis based on near-infrared spectroscopy. *Infrared Phys. Technol.* **2018**, *95*, 88–92. [CrossRef]
43. Wang, J.; Wang, W.C.; Hu, X.X.; Qiu, L.; Zang, H.F. Black-winged kite algorithm: A nature-inspired meta-heuristic for solving benchmark functions and engineering problems. *Artif. Intell. Rev.* **2024**, *57*, 98. [CrossRef]
44. Wan, J.X.; Ma, Y. Multi-scale Spectral-Spatial Remote Sensing Classification of Coral Reef Habitats Using CNN-SVM. *J. Coast. Res.* **2020**, *102*, 11–20. [CrossRef]
45. Kazemi, F.; Asgarkhani, N.; Jankowski, R. Optimization-based stacked machine-learning method for seismic probability and risk assessment of reinforced concrete shear walls. *Expert Syst. Appl.* **2024**, *255 Pt D*, 124897. [CrossRef]
46. Lelago, A.; Bibiso, M. Performance of mid infrared spectroscopy to predict nutrients for agricultural soils in selected areas of Ethiopia. *Heliyon* **2022**, *8*, e09050. [CrossRef]
47. Liu, Y.; Wang, C.; Xiao, C.; Shang, K.; Zhang, Y.; Pan, X. Prediction of multiple soil fertility parameters using VisNIR spectroscopy and PXRF spectrometry. *Soil Sci. Soc. Am. J.* **2021**, *85*, 591–605. [CrossRef]
48. Zhang, L.; Yang, X.; Drury, C.; Chantigny, M.; Gregorich, E.; Miller, J.; Bittman, S.; Reynolds, D.; Yang, J. Infrared spectroscopy prediction of organic carbon and total nitrogen in soil and particulate organic matter from diverse Canadian agricultural regions. *Can. J. Soil Sci.* **2018**, *98*, 77–90. [CrossRef]
49. Shi, Z.; Yin, J.; Li, B.; Sun, F.; Miao, T.; Cao, Y.; Shi, Z.; Chen, S.; Hu, B.; Ji, W. Comparison of Depth-Specific Prediction of Soil Properties: MIR vs. Vis-NIR Spectroscopy. *Sensors* **2023**, *23*, 5967. [CrossRef]
50. Bellon-Maurel, V.; Fernandez-Ahumada, E.; Palagos, B.; Roger, J.M.; McBratney, A. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *Trends Anal. Chem.* **2010**, *29*, 1073–1081. [CrossRef]
51. He, S.-F.; Zhou, Q.; Wang, F. Local wavelet packet decomposition of soil hyperspectral for SOM estimation. *Infrared Phys. Technol.* **2022**, *125*, 104285. [CrossRef]
52. Guo, L.; Zhang, H.; Shi, T.; Chen, Y.; Jiang, Q.; Linderman, M. Prediction of soil organic carbon stock by laboratory spectral data and airborne hyperspectral images. *Geoderma* **2019**, *337*, 32–41. [CrossRef]
53. Nan, F.; Zhu, H.; Bi, R. Hyperspectral Prediction of Soil Organic Matter Content in the Reclamation Cropland of Coal Mining Areas in the Loess Plateau. *Sci. Agric. Sin.* **2016**, *49*, 2126–2135.
54. Chen, Y.; Wang, J.; Liu, G.; Yang, Y.; Liu, Z.; Deng, H. Hyperspectral Estimation Model of Forest Soil Organic Matter in Northwest Yunnan Province, China. *Forests* **2019**, *10*, 217. [CrossRef]
55. Zhou, P.; Zhou, G.; Zhu, Z.; Tang, C.; He, Z.; Li, W.; Jiang, F. Health Monitoring for Balancing Tail Ropes of a Hoisting System Using a Convolutional Neural Network. *Appl. Sci.* **2018**, *8*, 1346. [CrossRef]
56. Pham, V.; Weindorf, D.C.; Dang, T. Soil profile analysis using interactive visualizations, machine learning, and deep learning, Comput. *Electron. Agric.* **2021**, *191*, 106539. [CrossRef]
57. Pullanagari, R.R.; Dehghan-Shoar, M.; Yule, I.J.; Bhatia, N. Field spectroscopy of canopy nitrogen concentration in temperate grasslands using a convolutional neural network. *Remote Sens. Environ.* **2021**, *257*, 112353. [CrossRef]
58. Wang, Y.; Chen, S.; Hong, Y.; Hu, B.; Peng, J.; Shi, Z. A comparison of multiple deep learning methods for predicting soil organic carbon in Southern Xinjiang, China. *Comput. Electron. Agric.* **2023**, *212*, 108067. [CrossRef]
59. Wu, J.; Guo, D.; Li, G.; Guo, X.; Zhong, L.; Zhu, Q.; Guo, J.; Ye, Y. Multivariate methods with feature wavebands selection and stratified calibration for soil organic carbon content prediction by Vis-NIR spectroscopy. *Soil Sci. Soc. Am. J.* **2022**, *86*, 1153–1166. [CrossRef]
60. Shekofteh, H.; Ramazani, F.; Shirani, H. Optimal feature selection for predicting soil CEC: Comparing the hybrid of ant colony organization algorithm and adaptive network-based fuzzy system with multiple linear regression. *Geoderma* **2017**, *298*, 27–34. [CrossRef]
61. Wang, X.; Zhang, M.-W.; Guo, Q.; Yang, H.-L.; Wang, H.-L.; Sun, X.-L. Estimation of soil organic matter by in situ Vis-NIR spectroscopy using an automatically optimized hybrid model of convolutional neural network and long short-term memory network. *Comput. Electron. Agric.* **2023**, *214*, 108350. [CrossRef]
62. Meng, X.; Bao, Y.; Wang, Y.; Zhang, X.; Liu, H. An advanced soil organic carbon content prediction model via fused temporal-spatial-spectral (TSS) information based on machine learning and deep learning algorithms. *Remote Sens. Environ.* **2022**, *280*, 113166. [CrossRef]
63. Xiang, H.; Liu, W.; Peng, J. Predicting Organic Matter Content in Paddy Soil Using Meth-od of Continuum Removal in Southern Xinjiang, China. *Soils* **2016**, *48*, 389–394.

64. Odebiri, O.; Mutanga, O.; Odindi, J.; Naicker, R.; Masemola, C.; Sibanda, M. Deep learning approaches in remote sensing of soil organic carbon: A review of utility, challenges, and prospects. *Environ. Monit. Assess.* **2021**, *193*, 802. [CrossRef]
65. Zhai, M. Inversion of organic matter content in wetland soil based on Landsat 8 remote sensing image. *J. Vis. Commun. Image Represent.* **2019**, *64*, 102645. [CrossRef]
66. Wang, S.; Guan, K.; Zhang, C.; Lee, D.; Margenot, A.J.; Ge, Y.; Peng, J.; Zhou, W.; Zhou, Q.; Huang, Y. Using soil library hyperspectral reflectance and machine learning to predict soil organic carbon: Assessing potential of airborne and spaceborne optical soil sensing. *Remote Sens. Environ.* **2022**, *271*, 112914. [CrossRef]