Application of Hyperspectral Technology Combined With Bat Algorithm-AdaBoost Model in Field Soil Nutrient Prediction

HUAN WANG¹, LIXIN ZHANG¹,², JIAWEI ZHAO¹, XUE HU¹, AND XIAO MA¹
¹College of Mechanical and Electrical Engineering, Shihezi University, Shihezi 832000, China
²Industrial Technology Research Institute XPCC, Shihezi 832000, China

Corresponding author: Lixin Zhang (lixin_zhang_cn@126.com)

This work was supported by the National Natural Science Foundation of China under Grant 51365048.

ABSTRACT This paper proposes a hyperspectral soil nutrient estimation method based on the bat algorithm (BA)-AdaBoost model. The spectral reflectance, the first derivative of the reflectance, and the reciprocal logarithm of the reflectance are analyzed based on the 800 field soil samples and their hyperspectral data collected. The first derivative of the reciprocal logarithm of the reflectance and the sensitive band was extracted using the correlation coefficient method, and the correlation of the content of soil organic matter, phosphorus, and potassium was solved. The BA is used to optimize the two core parameters of the AdaBoost model (i.e., the maximum number of iterations (n) and the weight reduction coefficient (v) of the weak learner), the classification and regression trees (CART) decision tree is selected as the weak regression learner of the model, and the coefficient of determination is used as parameter optimization. Based on the objective function value, a BA-AdaBoost model was constructed to estimate soil organic matter and phosphorus and potassium contents. The results show that the BA-AdaBoost combined model can better search for globally optimal parameters. The AdaBoost model optimized by BA significantly improved accuracy and reliability. Among the three elements, soil organic matter estimation accuracy is the highest, and the coefficient of determination and the root mean square error are 0.867 and 0.151 g·kg⁻¹, respectively. Compared with the model before optimization, the model accuracy and reliability improved by 29.0% and 24.1%, respectively. The results indicate that hyperspectral technology combined with the BA-AdaBoost model has certain application prospects in field soil nutrient estimation.

INDEX TERMS Hyperspectral, soil nutrient content, machine learning, bat algorithm, AdaBoost model.

I. INTRODUCTION

A. BACKGROUND AND MOTIVATION

Agriculture is the basis of and very important to human survival. At present, the total global population exceeds 7.5 billion and is expected to reach 11 billion by 2100 [1]. By the end of 2020, at least 83 million people existed and possibly as many as 132 million were hungry. In addition, 3 billion or more people cannot afford a healthy diet [2]. Furthermore, the amount of food available is limited, especially in Africa, and the problem of food shortage remains to be solved [3]. With the development of human society and the economy, limited resources cannot meet the needs of the growing population [4]. Under the influence of multiple factors (e.g., nature and social economy), productivity had declined, land degradation, soil erosion, and other problems have continuously emerged, and mankind is facing very severe challenges [5], [6]. Soil nutrients, such as organic matter (OM), phosphorus (P), and potassium (K), play an important role in the process of crop growth [7], [8]. Plants need a large amount of nitrogen to synthesize protein [9], phosphorus can promote the formation and growth of plant roots [10], and potassium can promote photosynthesis [11]. The content of the main nutrients (total nitrogen, available phosphorus, and available potassium) in soil is an important index affecting crop yield and an important basis for
guiding agricultural scientific fertilization [12], [13]. Therefore, accurately grasping the content of soil nutrients is of great significance in guiding agricultural production and scientific fertilization proportioning.

**B. PROBLEM STATEMENT**

Predicting soil element content can guide irrigation and fertilization, ensure a balanced supply of crop nutrients, meet the growth needs of crops, improve fertilizer utilization efficiency, and reduce fertilization to protect the environment, improve yield, and save manpower and material resources [14]. However, the traditional method of measuring the content of soil components is that for field sampling and then laboratory analysis. This method has high requirements for testers and requires expensive testing equipment [15]. Moreover, high testing cost, low efficiency, and inability to conduct large-scale testing simultaneously are its problems. However, hyperspectral analysis technology has the advantages of low cost, being quick, and environmental protection. Thus, it has continuously been widely used in soil nutrient prediction in recent years [16]. Not only can hyperspectral analysis technology be used by inspectors for backpack operations, it can also be used to carry unmanned aerial vehicles and even conduct remote sensing operations to realize omnidirectional soil nutrient detection in space and heaven [17], [18].

Early scholars used correlation analysis to study the relationship between different types of soil organic matter and soil spectral reflectance and different transformation forms, and took the wavelength with a high correlation coefficient as the sensitive band of soil organic matter. For example, Wang and other studies showed that the content of soil organic matter had a strong correlation with the reflectance of a red light band, and the determination coefficient could reach 0.61 [19]; Songo.D. et al. think that the organic matter content of black soil in Northeast China has a very significant negative correlation with the original spectral reflectance in a 545~738 nm band and the first derivative spectrum in a 481~598 nm band, but a very significant positive correlation with the first derivative spectrum in a 816~932 nm band and 1,039~1,415 nm band [20]. In recent years, multivariable regression modeling methods have been gradually applied to the estimation of soil nutrients. For example, Williams et al. used principal component analysis combined with multiple linear regression and an artificial neural network to construct soil organic matter model [21]. Zavyalova et al. used modeling schemes to build estimation models of soil organic matter of different land use types, and the highest determination coefficient $R^2$ of estimation accuracy was 0.89 [22]. Zhang et al. constructed an estimation model for the content of different components of soil organic carbon in Australia [23]. Kong et al. used a partial least squares method in combination with a variety of spectral absorption characteristics to construct an estimation model for soil organic carbon in the European Land Use/Cover Area frame Survey (LUCAS) [24]. Gruszczyński et al. used CNN model to estimate K element in soil [25]. He. B. et al. Combined support vector machine and chaotic whale optimization algorithm to estimate soil moisture of corn [26].

At present, using CNN, SVR and other algorithms to establish the inversion model between soil hyperspectral characteristics and element content is the main method with which to estimate soil element content [27]. CNN, SVR and other algorithms usually need to set some model parameter values in advance [28]. Because the predefined parameter values may not contain the globally optimal parameter values, the aforementioned model cannot achieve the best effect [29]. In order to overcome the problems of machine learning models in finding the best model parameters, classical optimization algorithms such as the genetic algorithm and the particle swarm optimization algorithm are used to optimize the internal parameters of machine learning models [30]. However, classical optimization algorithms such as particle swarm optimization (PSO) are sensitive to the initial parameter setting, and it is easy to fall into the locally optimal solution in the optimization process, resulting in slow convergence speed in the later stages of the algorithms [31]. The bat algorithm is a new swarm intelligence method. In the process of optimization, the bat algorithm imitates the adaptive adjustment process of bat acoustic pulse loudness and frequency to realize the free switching between global optimization and local optimization so as to balance the global search ability and local search ability of the algorithm and perform well in the optimization calculation of model parameters [32]. AdaBoost makes good use of weak classifiers to cascade and has high prediction accuracy. However, the maximum number of iterations N and the weight reduction coefficient V of weak learners in AdaBoost model are not set well, and training is time-consuming.

This paper studied the applicability of the mixed Ba AdaBoost model in estimating the contents of organic matter, phosphorus and potassium in agricultural landscape soils. Specifically, the objectives are (1) to develop a mixed Ba AdaBoost model using 800 collected field soil samples and their hyperspectral data, and to compare the prediction performance of the mixed Ba AdaBoost model with that of AdaBoost model, MLR, CNN and SVR methods.

**C. CONTRIBUTION**

First, 800 field soil samples and their hyperspectral data were collected, and the content of OM, P, and K was determined in the corresponding samples. Through the hyperspectral characteristic transformation of soil nutrients, the spectral reflectance, the first-order differential of reflectance, the reciprocal logarithm of reflectance, and the first-order differential of the reciprocal logarithm of reflectance were solved. Sensitive bands were extracted using the correlation coefficient method, and the correlation of OM, P, and K content was analyzed. The bat algorithm is used to optimize the two core parameters of the maximum number of iterations (n) and the weight reduction coefficient (v) of the weak learner in the AdaBoost model. The CART decision tree is selected.
as the weak regression learner of the model, and the decision coefficient is used as the objective function value of parameter optimization. The BA-AdaBoost model proposed in this paper can effectively predict the content of OM, P, and K, and its prediction accuracy is 0.867, 0.853, and 0.729, respectively. The model can quickly search for the globally optimal parameters. The accuracy and reliability of the AdaBoost model optimized by the BA are significantly improved and the prediction accuracy of the model is improved by 29.0% and 24.1% in comparison with that before optimization. The results show that hyperspectral technology combined with the BA-AdaBoost model has a certain application prospect in field soil nutrient prediction.

D. PAPER ORGANIZATION
The remainder of this paper is organized as follows: Section 2 discusses related work, Section 3 discusses the proposed method in detail, Section 4 discusses the results and discussion, and Section 5 draws a conclusion.

II. RELATED WORK
The traditional detection method of soil nutrients is that of field sampling followed by laboratory analysis [33], [34]. M Chen et al. used the hybrid composite membrane method to monitor soil nitrate–nitrogen [35]. Moreover, JCV Puno et al. used genetic algorithms for the qualitative level classification of soil nutrients [36]. A large number of scholars have also installed sensor devices in fields to measure soil moisture, electrical conductivity, and pH to invert soil nutrient content [37], [38].

Hyperspectral remote sensing technology has played an increasingly important role in quantitative remote sensing monitoring of soil components due to its high spectral resolution, strong band continuity, and high spatial resolution in recent years. With the support of various mathematical models, the nutrient content was quantitatively calculated by establishing the correlation between the soil sample spectrum and the content. L. L. et al. used a FieldSpec spectrometer to reveal the soil reflectance spectroscopy mechanism to estimate total nitrogen content [39]. In addition, Guigue et al. studied the hyperspectral inversion method of organic matter and total phosphorus and concluded that the influence of different land use methods on the accuracy of organic matter inversion can be ignored but that the total nitrogen needs to be differentiated and modeled [40]. Moreover, Greenberg et al. applied field-walking equipment to solve the problem of feature selection and transformation in the inversion of soil carbon content by soil infrared spectroscopy [41]. Furthermore, Pradhan et al. studied total nitrogen estimation through the spectral carbon content index of wheat leaves. The comprehensive method of the content proves the feasibility of the indirect method [42].

Based on the data acquired, the establishment of the model needs to solve the problems of a large amount of hyperspectral data and the high redundancy between bands. Moreover, machine learning technology is a series of processes of inputting existing knowledge into computers and exploring new knowledge discovery. The introduction of machine learning theory in this field has become a research hotspot. Support vector machines [43], manifold learning [44], firefly algorithms [45], neural networks [46], random forests [47], decision-making trees [48], and other machine learning methods are used in soil classification, content estimation, model optimization, automatic interpretation, and feature recognition. The conventional method matches spectral data or their variation with laboratory data one by one, which are inputted into a computer to obtain the information discovery model with which to calculate the soil composition content represented by the unknown spectrum [49].

Balancing the problem of local and global optimums is difficult due to the large randomness of machine learning model parameter settings [50]. This paper analyzes in detail the spectral characteristics of OM, P, and K elements in the fields, selects the optimal spectral transformation form of the elements and the sensitive band with high correlation, and combines the bat algorithm and the AdaBoost machine learning model to build a soil nutrient content prediction model. The bat algorithm is used to solve the key parameters in the AdaBoost model modeling, and the prediction accuracies before and after the optimization of the model parameters are compared, providing an efficient new method for the hyperspectral prediction of soil nutrient content in fields.

III. PROPOSED METHODS
A. SOIL SAMPLE COLLECTION
The soil sample collection was conducted in the Xinjiang Autonomous Region, China, and the location of the soil sample collection is shown in Figure 1. A 5 km grid was set as the sampling unit. To enable the sampling points to be arranged so as to represent the soil properties of the sampling unit, 800 soil samples were collected in the experimental area. During the sample collection process, the sampling distance was at least 150 m from the road. The position of each sampling point was the center, and the samples were collected within a range of 5 × 5 m in circumference. Five black soil samples were collected at each sampling point, and the sampling depth was within 15 cm of the topsoil. The samples were fully mixed and then placed inside of sampling bags.

Soil samples were collected from three test fields in Turushuke, Shihezi and Tacheng (Xinjiang, China), which are operated and managed by Shihezi University. The longitude and latitude ranges are 40°04′ ~ 39°36′ N and 78°38′ ~ 79°50′ E in the Turushuke area. The annual average temperature is 11.6 °C, the average temperature in the hottest month (July) is 25.0 °C ~ 26.7 °C, the average temperature in the coldest month (January) is ~6.6 °C ~ 7.3 °C, the annual average frost-free period is about 225 days, the annual precipitation is 38.3 mm, and the altitude is 425 ~ 437m. The longitude and latitude ranges of the Shihezi area are 44°25′ ~ 44°27′ N and 85°40′ ~ 86°10′ E,
the annual average temperature is 9.2 °C, the average temperature in the hottest month (July) is 25.1 ~ 26.1 °C, the average temperature in the coldest month (January) is −20.6 °C ~ −7.3 °C, the annual average frost-free period is about 170 days, the annual precipitation is 43.5mm, and the altitude is 510 ~ 520m. The longitude and latitude ranges are 46°31' ~ 46°37' N and 83°37' ~ 83°41' E in the Tacheng area. The annual average temperature is 7.6 °C; the average temperature in the hottest month (July) is 24.5 ~ 25.8 °C, the average temperature in the coldest month (January) is −26.6 °C ~ −9.6 °C, the annual average frost-free period is about 155 days, the annual precipitation is 27.1mm, and the altitude is 1,100 ~ 1,150m. The cash crops are mainly cotton and wheat, and the soil properties are mainly loamy and SIL clay. Ten 5km grids were set up in Tumushuke, Shihezi and Tacheng test fields as sampling units. In order to make each sampling point representative of the soil properties of the sampling unit, 80 soil nutrients were sampled for each sampling, and a total of 800 soil samples were collected. Five meters at each sampling point according to the five-point sampling method × samples shall be collected within 5m. A total of five soil samples were collected at each sampling point. The samples shall be fully mixed and loaded into sampling bags. At the same time, in order to avoid some lead pollution with different content caused by human activities, ensure that the sampling distance from the highway of at least 150m was ensured during the sampling process. Since soil macro-nutrients are mainly concentrated on the earth’s surface, soil samples are collected in the soil layer at a depth of about 15cm according to traditional soil sampling technology.

The distance between the sample collection area and the laboratory is long, so it is inevitable that the temperature and moisture of the soil will be affected during the transportation of soil samples. We also attach great importance to samples in the process of transportation. After collecting soil samples, we should put them in a dark and low-temperature (4 °C) closed environment as soon as possible to maintain the stability of soil water content. The dark environment is to avoid the growth of algae in the soil under light. The low temperature is to reduce bacterial reproduction and maintain the stability of microbial flora. Polyethylene bags shall be used as packing materials for loose binding. In addition, in order to avoid overlapping of sample bags during storage, the method of paving should be adopted to avoid damaging the original aggregate structure of the soil and causing the underlying samples to be in an anaerobic environment.

B. SOIL NUTRIENT CONTENT MEASUREMENT AND SAMPLE DIVISION

The process comprised removing the plant tissue, gravel and other sundries in the sample, and then grinding and screening after air drying so that the particle size of the soil was <0.25 mm; The samples were divided into two parts, namely one for the determination of soil element content and the other for indoor hyperspectral measurement. The potassium dichromate volumetric method was used to determine the content of soil organic matter. The delta professional 4050 portable X-ray fluorescence spectrometer (Olympus of the United States) was used to measure the content of soil phosphorus and potassium. The equipment was equipped with an Au target micro X-ray exciter and adopted an SDD detector cooled by Peltier semiconductors. The detector area was 25 mm² and the energy resolution is 128ev, It could simultaneously measure the content of 38 elements (including soil phosphorus and potassium). Thereafter, the measurement results of element content were statistically analyzed according to [51], and the results are presented in Table 1. 800 soil samples were divided into 200 groups according to the element content from low to high. One sample was randomly selected from each group and put into
TABLE 1. Statistics of nutrient content in soil samples.

| Soil nutrients | Number of samples | Minimum (g·kg⁻¹) | Maximum (g·kg⁻¹) | Mean (g·kg⁻¹) | Standard deviation (g·kg⁻¹) |
|---------------|------------------|------------------|------------------|--------------|-----------------------------|
| OM            | 800              | 2.016            | 5.107            | 3.209        | 0.487                       |
| P             | 800              | 0.415            | 0.963            | 0.571        | 0.081                       |
| K             | 800              | 1.669            | 2.172            | 1.910        | 0.659                       |

FIGURE 2. Laboratory data acquisition device.

C. HYPERSPECTRAL ACQUISITION AND PREPROCESSING

The FieldSpec 4 surface feature spectrometer was used to collect soil hyperspectral data. In order to eliminate the interference of scattered light, a light source and black box were used, as shown in Figure 2. The light source was provided by two 50W (650 lx) quartz halogen tungsten lamps. The lamp source was 130cm high from the measured soil samples, and irradiated the measurement target at a depression angle of about 50°. The measurement target was located in the center of the light spot of the light source to ensure that the light source in the field of view was uniform, and the spectrum of the light source was as close as possible to the solar spectrum. The soil samples were placed in a Petri dish (diameter 10 cm, depth 1.5 cm) completely blackened inside, with their surface being leveled with a ruler. The ASD optical fiber probe is vertically fixed at a height of about 3 cm from the measured soil samples. The sensor probe adopted a 25° field angle probe. The area in which the probe received the soil spectrum was about a circle with a diameter of 1.33cm. The field of view of the spectrometer was much smaller than the area of the Petri dish, a circle with a diameter of 0.67cm, so as to ensure that the reflection spectrum of the soil samples was received by the probe. White board correction was carried out before the test. Ten spectral curves of each soil samples, taking their arithmetic mean as the actual reflection spectral data of the soil samples. The data before 400nm and after 2,450nm were excluded because spectral data in the ranges of 350–399nm and 2,451–2,500nm has high noise and a low signal-to-noise ratio, which interferes with analysis of the relationship between soil elements and reflectance. The sampling interval of the spectrometer was 1 nm (i.e., 2,051 bands were obtained in the range of 400–2,450 nm). Due to the high spectral resolution and a large number of bands, information overlap may happen between adjacent bands, which is more susceptible to noise. Therefore, the spectral data were resampled and the sampling interval was set to 10 nm. Based on denoising and resampling processing, the original spectral reflectance was transformed with features such as the first-order differential, reciprocal logarithm, and so on. Different transformation forms can help to find peaks and valleys accurately and quickly and determine the corresponding wavelengths through peaks and valleys to determine a sensitive band.

D. MODEL OVERVIEW

1) BAT ALGORITHM

The bat algorithm is a heuristic search algorithm. It simulates bats using sonar to detect prey and avoid obstacles. Moreover, the optimization search process simulates the process of bats flying to find prey. In the calculation process, the fitness value of the problem is used to select the position of the bats using the evolutionary process of survival of the fittest to simulate the iterative search process in which the better feasible solution replaces the poorer feasible solution [52]. Because Ba has the characteristics of fast speed and few parameters, it has been widely concerned in many optimization problems and has shown good performance. Based on the basic BA principle, the heuristic search starts at a random position (Z) in the d-dimensional search space after the parameters of the algorithm are initialized. The prey was searched at a fixed frequency of different wavelengths and sound intensities. Bat algorithm flowchart. As shown in Figure 3. During the search process, a bat automatically adjusts the wavelength following the distance to the prey. After the global search, the flight speed and spatial position of each bat are updated and the fitness value of the objective function is calculated. The updated formula for the speed and spatial position is Formula (1):

\[ f_t = f_{\text{min}} + (f_{\text{max}} - f_{\text{min}}) \beta \]
\[ v_{t+1} = v_t + (z_t - z^*) \beta \]
\[ z_{t+1} = z_t + v_{t+1} \]

(1)
All variables are explained in Appendix A nomenclature and Appendix B nomenclature.

The sound intensity and frequency in the bat algorithm will be updated and calculated following the impulse loudness attenuation coefficient and the impulse frequency increase coefficient after each iteration.

2) ADABOOST ALGORITHM
The basic idea of the AdaBoost algorithm is to build a strong regressor with strong training ability through the superposition method for the weak regressor with general training ability in the same training set. The algorithm itself is realized by changing the weight of the sample distribution. According to the prediction accuracy of each sample in each training set and the overall prediction accuracy of the last time, the weight of each weak regressor is calculated and the distribution weight of each sample is updated simultaneously. Finally, the weighted summation of the regressor results obtained from each training is used as the final output result of the strong regressor. In the modeling process, the maximum number of iterations (n) of the weak learner and the weight reduction coefficient (v) of the weak learner are two important parameters. If the number of iterations (n) is too small, the model will not fit. However, if n is too large, it will lead to the model being overfitted. A smaller weight reduction coefficient (v) means that more iterations of the weak learner are required. Thus, these two parameters should be usually optimized together.

![Bat algorithm flowchart.](image)

3) MULTIVARIATE LINEAR REGRESSION MODEL
Ba AdaBoost was compared with multiple linear regression (MLR) equation for prediction performance. Based on the classical least square method, a multi-objective linear programming is constructed.

4) CNN MODEL
The CNN model uses the hidden layer for training, in which the filter size of the convolution layer (conv1d) is 128 and the core size is 2. The same filling type, ELU activation type and unified core initializer are used. At the same time, a pooling layer Max pooling is applied, the pooling size is 1, and the maximum value of each neuron cluster is in the previous layer. After the pooling layer of CNN, the dropout layer with a rate of 0.001 is used to adjust the overfitting. The pooling layer also helps to overcome overfitting. A fully connected layer, called 'dense', is applied to the output layer, where a linear activation function is used. Finally, the model was compiled using the MSE loss function and adagrad optimizer with a learning rate of 0.01.

The structure of CNN model is:
Convolutional layer (128 filters + 2 filter sizes + ELU activation + padding 'same' + 1 stride) + Maxpooling (1 pooling size) + Flatten + Dense layer (30 neurons, SELU activation) + 1 Dropout layer (0.001) + 1 Dense layer (1 neuron) + Compile (MSE loss, Adagrad optimizer with a learning rate of 0.002)

5) SVR MODEL
Epsilon SVR karyotype equation was used to construct SVR model. When developing SVR models, different types of kernel functions are considered, such as linear, radial basis function (RBF), sigmoid and polynomial. RBF is the best kernel and the prediction error is the smallest. It has fewer hyperparameters, which reduces the complexity of the model, so its performance is better than other kernel functions. When using RBF kernel function to train SVR model, set the regulator (C), epsilon (ε) and epsilon (ε). Hyperparameters. In the training phase, the grid search with ten fold cross validation proposed by Duan et al. is used for super parameter optimization. For OM prediction, C ε and γ are used. The optimal values are 1, 0.01, and 1 respectively, and the optimal values for P and K prediction are 1, 0.1, and 0.1 respectively. Python’s scikit learn Library Pedregosa et al. [53] was used to implement the SVR model.

E. PRINCIPLE OF SPA CHARACTERISTIC BAND SELECTION
The selection of a characteristic wavelength is very important in establishing a stable spectral model. When using full band spectral data to establish the model, not only is the calculation workload heavy, but the prediction accuracy of the calibration model will make it difficult to reach the optimal value. Therefore, it is necessary to select the characteristic wavelength before establishing the calibration model. At present, commonly used characteristic wavelength selection methods include the correlation coefficient method, analysis of variance, stepwise multiple linear regression, particle swarm optimization, the inverse interval partial least squares method, and the continuous projection algorithm. Among them, the continuous projection algorithm is a new wavelength selection method which has become more and more widely used. This study also uses this algorithm to extract the characteristic bands. SPA is a variable group that can fully find the minimum redundant information from the spectral information so as to minimize the collinearity between the variables. It is a deterministic search method, and its variable
selection results are reproducible, which will be more robust to the variable selection of the verification set. It can extract effective information from the severely overlapping spectral information, minimize the collinearity between the spectral variables, simplify the correction model and improve the speed and efficiency of modeling.

First establish the spectral matrix $X^{l\times p}$ and the sample property vector $y$, where $l$ is the sample capacity and $p$ is the number of full spectrum wavelengths.

SPA characteristic band selection steps are as follows:

Step 1: let $i = 1$, assign the $k$-th column of the spectral matrix $(X)$ to $x_k^{(i)}$, that is, $k(1) = k, x_k^{(i)} = x_k$, and let $z_i = x_k^{(i)}$, $x_k^{(i)} = x_j, j = 1, 2, ..., p$;

Step 2: the wavelength vector $(x_j)$ that has not been selected; the position set of is marked as $S_i$, with $S_i = \{j, 1 \leq j \leq p, j \notin \{k(1), k(1), ..., k(i)\}\}$;

Step 3: construct orthogonal projection operator $P_i = I - z_i(z_i^T)z_i$ based on $z_i$, where $I$ is the identity matrix of $l \times l$;

Step 4: calculate the orthogonal projection vector $x_j^{(i+1)} = P_jx_j^{(i)}$ of each $x_j$, and select the wavelength position, $k(x+1) = \arg(\max(||x_j^{(i+1)}||)), J \in S_i$;

Step 5: make $i = i+1, z_i = x_j^{(i)}(i)$; if $i < m$, return to Step 2 to select the next wavelength vector.

All variables are explained in Appendix A nomenclature and Appendix B nomenclature.

F. MODEL EVALUATION CRITERIA

To evaluate the predictive ability and stability of the model, the coefficient of determination ($R^2$) and the root mean square error (RMSE) are selected to evaluate the modeling effect of the model. The $R^2$ and RMSE evaluation formulas are those of Formulas (2) and (3):$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (2)$$  $$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (3)$$

The lower the RMSE calculated by the prediction model and the closer $R^2$ is to 1, the higher the accuracy and stability of the prediction model.

IV. RESULTS AND DISCUSSION

A. SPECTRAL FEATURE TRANSFORMATION

The spectral curve of the treated indoor soil sample is shown in Figure 3(a). The original reflectance of the collected soil samples is between 0 and 0.9, and the spectral curve of each sample shows a similar fluctuation. With the wavelength length, the reflectance gradually increases until the reflectance tends to be stable at 1,200 nm. The spectral reflectance of the soil in the near-infrared region is generally higher than that in the visible-light region. Two wave valleys are distributed around 1,400 nm and 1,900 nm, which are mainly caused by the absorption of residual moisture in the soil and water vapor in the air, and a slight depression at 2,200 nm is influenced by the presence of clay minerals in the soil. Figures 3(b)–(d) are the spectra of the original reflectance transformed by the first-order differential ($R'$), the reciprocal logarithm ($lg1/R'$), and the inverse logarithmic first-order differential ($[lg1/R']$), respectively. Moreover, the first-order differential transformation can amplify the original spectral changes, and the reflectivity fluctuates more at 1,400 nm, 1,900 nm, and 2,200 nm after transformation. To optimize the display effect, only 30 random groups of soil sample spectra and their transformation curves are presented in Figure 3.

B. CORRELATION ANALYSIS AND FEATURE BAND SELECTION

SPA was used to calculate the correlation coefficient of soil organic matter, phosphorus, and potassium content and soil reflectance, and the correlation coefficient curve was drawn (Figure 4). From the correlation coefficients of the original spectra in Figure 4(a), soil organic matter and phosphorus element contents are observed to be negatively correlated with spectral reflectance, while potassium element content is inversely correlated. Compared with the original spectral reflectance, the correlation between the transformed spectral data and OM, P, and K content is higher, and the correlation coefficient between the first-order differential transformation and OM, P, and K content is positive and negative. More peaks and troughs were noted, and the highest correlation coefficient of each element was significantly improved after the first-order differential transformation (Figure 4(b, c and d)).

In order to enable original reflectance, $R'$, $lg1/R'$ and $[lg1/R']$ transformation curves to select a certain number of characteristic bands, the median of 0.4 of the medium correlation coefficient (0.3-0.5) was taken as the selected threshold. The sensitive band with the absolute value of the correlation coefficient >0.4 was selected as the sample input in Table 2 of the prediction model. The statistical results show that the correlation coefficient between soil element content and spectral reflectance improved after different spectral feature transformations (Figure 5(b)). The absolute value of the coefficient increased to 0.788, with the wavelength range being 1,376 nm. The best transformation form of the soil phosphorus element is the first-order differential of the reciprocal logarithm, with the absolute value of the correlation coefficient being up to 0.590, and the wavelength is around 541 nm. For the soil potassium element, the corresponding optimal transformation form is the reciprocal logarithmic first-order differential, with the absolute value of the correlation coefficient being up to 0.634, and the wavelength around 580 nm (Figure 5(d)). According to different prediction objects (OM, P and K), the model input curve will also change, and the best transformation form should be used as the model input.

C. PREDICTION BASED ON BA-ADABOOST MODEL

The sensitive bands and the corresponding element content values of the best spectral transformation form of 600 training
samples are selected as the modeling data of the AdaBoost model. The number of modeling bands for the three elements is 86, 75, and 63, respectively, and the CART decision tree is selected, which is the weak regression learner of the AdaBoost model. The bat search algorithm is used to optimize the maximum number of iterations (\( n \)) of the AdaBoost model and the weight reduction coefficient (\( v \)) of the weak learner.

Based on the basic BA principle, the parameters of the model need to be initialized first. BA has many initialization parameters, but other parameters are not sensitive except for the number of iterations. The default parameters can be selected for initialization. The search space of the BA algorithm is a two-dimensional space composed of the maximum number of iterations (\( n \)) and the weak learner weight reduction coefficient (\( v \)) as the coordinate axis. The iterative search

### TABLE 2. Maximum correlation coefficients and sensitive bands.

| Soil nutrients | Absolute value maximum | Corresponding band | Number of sensitive bands | Absolute value maximum | Corresponding band | Number of sensitive bands |
|----------------|------------------------|--------------------|---------------------------|------------------------|--------------------|---------------------------|
| OM             | 0.634                  | 0.788              | 635                       | 0.634                  | 0.707              |
| P              | 0.459                  | 0.588              | 635                       | 0.590                  | 0.486              |
| K              | 0.634                  | 0.478              | 635                       | 0.436                  | 0.478              |
|                | 24                     | 17                 | 63                        | 47                     | 63                 |

**FIGURE 4.** Soil spectral reflectance curves and its transformations. (a) Original spectral reflectance curve. (b) \( R' \) transformation curves. (c) \( \text{lg1/R} \) transformation curves. (d) \( \text{lg1/R}' \) transformation curves.
process starts at $L$ random positions in the search space. The two-dimensional coordinates ($n$ and $v$) of the position of each bat in space are used as the initialization parameters of the AdaBoost model. Consequently, the model is trained on the sample data based on the AdaBoost model, and the prediction results are calculated. According to the calculation results, the position corresponding to the maximum $R$ value is selected as the current optimal position of a bat, and the spatial position of each bat is updated using Formula (1). The population size ($L = 20$), pulse frequency range ($f_{\text{min}} = 0$, $f_{\text{max}} = 1$), pulse sound intensity range ($A_{\text{min}} = 0$, $A_{\text{max}} = 1$), pulse loudness attenuation coefficient ($\alpha = 0.9$), and pulse frequency increase coefficient ($\gamma = 0.9$) are set. The default parameters are used, and $R^2$ corresponding to different iterations is calculated (Figure 6). Moreover, with the increase in the number of iterations ($T$), the value of $R^2$ gradually increases. Among the three elements, potassium has the fastest convergence rate. The value of $R^2$ reaches the maximum when the number of iterations reaches 10, and organic matter and phosphorus are maximized when the number of iterations reaches 12 and 16, respectively. When modeling and predicting, the values of $n$ and $v$ corresponding to the maximum $R^2$ are selected as the modeling parameters of the AdaBoost model.

D. MODEL ACCURACY ANALYSIS

To analyze the prediction accuracy of the BA-AdaBoost model, the $R^2$ and RMSE of the training and verification set model MLR, CNN, SVR and Ba before and after optimization
are calculated (Table 3). The table shows that the hyperspectral prediction accuracy of organic matter in both training and verification sets is higher than that of phosphorus and potassium. After BA optimization, the accuracy of the AdaBoost model greatly improved. From the results of the verification set, the $R^2$ of phosphorus changed most significantly, from 0.714 to 0.853. The root mean square error of validation of organic matter decreased from 0.199 to 0.151 g·kg$^{-1}$. Among the three elements, the prediction accuracy of soil organic matter is the highest, and the determination coefficient and the root mean square error are 0.867 and 0.151 g·kg$^{-1}$, respectively. Compared with the model before optimization, the accuracy and reliability of the model improved by 29.0% and 24.1%, respectively. Furthermore, the prediction accuracy of phosphorus and potassium are 0.853 and 0.729, respectively.

The efficiency of the model was examined using the Taylor diagram (Fig. 7). For the prediction of OM, the correlation coefficient of Ba AdaBoost is 0.982, and the normalized standard deviation is 0.523, which is the best result. The correlation coefficient of SVR model is 0.782, and the normalized standard deviation is 1.0. The result is the worst. Similar to the OM prediction, the correlation coefficient of Ba AdaBoost is 0.960 and the normalized standard deviation is 0.407 for the prediction of P, and the result is the best. The correlation coefficient of SVR model is 0.715, and the normalized standard deviation is 1.0. The result is the worst. Similar to OM prediction, for K prediction, Ba AdaBoost’s correlation coefficient is 0.923 and normalized standard deviation is 0.619, with the best result. The correlation coefficient of SVR model is 0.736, and the normalized standard deviation is 1.0. The result is the worst. To sum up,
Ba AdaBoost model has the best prediction effect on OM, P and K. It can be seen from the box plot comparison model (Fig. 8) that for the predictions of OM, P and K, the median value generated by Ba AdaBoost model is the most similar to the observed value. The model is different in the lower quartile, 25th percentile (Q25) and data range (maximum and minimum), but it is better than the prediction effect.
of other models. For the variables of OM, P and K, the statistical variable difference of SVR model is the largest, indicating that it performs worst in all models. In general, the Ba AdaBoost model is very close to the observed value in terms of the predicted value, but the SVR model cannot. The SVR model needs well structured data to make more accurate predictions.

It can be seen from the violin graph comparison model (Fig. 8) that for the prediction of OM, Ba AdaBoost, CNN and SVR have higher performance compared with MLD and AdaBoost models (BA AdaBoost fits perfectly, and RF and RF fit well). For the prediction of P, Ba AdaBoost has high performance. For the prediction of K, AdaBoost, Ba AdaBoost and SVR have better performance than MLD and CNN models.

It is observed that the developed Ba AdaBoost hybrid machine learning model has good effects in the prediction of OM, P and K, although the implementation of Ba AdaBoost model is more complex than the traditional ML model, they have higher accuracy in nonlinear learning.

To sum up, the performance of the hybrid Ba AdaBoost model is better than that of the independent AdaBoost model, CNN and SVR model, because AdaBoost model has good estimation accuracy (but the distribution state and probability density fitting ability are weak), while Ba has good global search ability. The results show that the mixed Ba AdaBoost model established in this study can be used as an alternative to the prediction of soil OM, P and K.

Although this study confirms that Ba AdaBoost model has a good estimation ability for OM, P and K, it does not explain the relationship among OM, P and K. This unexplained difference may be related to some factors ignored in this study. For example, farming is a common practice in this area, and its impact on soil erosion is serious [54]. In addition, fertilization has a significant impact on the distribution of OM, P and K in the surface soil of the study area [55]. Other factors such as irrigation also affect the distribution of soil organic matter and soil moisture content.

Therefore, it is necessary to further study the relationship among OM, P and K, explain the impact of soil erosion, farming and irrigation on soil nutrients and water, and more accurately grasp the content of soil nutrients for guiding agricultural production, scientifically matching irrigation and fertilization. And study the robustness of Ba AdaBoost model under different soil types to explore the ability of the model to estimate different soil moisture content.

V. CONCLUSION

Taking 800 soil samples as the research object, the spectral characteristics of OM, P, and K were analyzed and the BA-AdaBoost model was constructed to predict the element content. The main conclusions are as follows:

1) The characteristic transformation processing of spectral data can enhance spectral characteristics. The first-order differential change effectively highlights the peaks and troughs of the spectral curve and improves the correlation coefficient between spectral reflectance and element content. The correlation coefficients of OM, P, and K reach the maximum at 1,376nm, 541nm and 580 nm, respectively.

2) The BA-AdaBoost soil content prediction model is constructed by combining the BA and AdaBoost models. The combined model only needs to set the search space and then automatically search for the optimal parameter value of the model. Compared with the prediction accuracy before and after optimization of the BA algorithm, the $R^2$ of the BA-AdaBoost model increased, the RMSE decreased, and the prediction accuracy significantly improved, which shows that the BA-AdaBoost model has certain applicability in the hyperspectral prediction of soil element content and expands the application of machine learning models in the prediction of soil composition.

3) Through the comparison of Taylor diagram, box diagram and violin diagram, the mixed Ba AdaBoost model is very close to the observation value, and has a good distribution state and probability density fitting ability. However, it is still necessary to further study the relationship among OM, P and K, verify the robustness of Ba AdaBoost model under different soil types, and verify its ability to explore different soil moisture contents.

APPENDIX A

NOMENCLATURE

\[ y = \text{The average of the measured values.} \]

APPENDIX B

NOMENCLATURE

\[ A_t = \text{The loudness, which helps update the bat’s position.} \]

\[ \beta = \text{A random number between } [0, 1]. \]

\[ f_t = \text{The pulse frequency of individual bat } t \text{ at time } t. \]

\[ r_t = \text{The pulse frequency, global search of control algorithm.} \]

\[ \text{It decreases with the iteration from 1 to 0, because the closer the bat is to the prey (the best point), the lower the loudness. The initial value is 1.} \]

\[ f_{\text{max}} = \text{The maximum frequency range.} \]

\[ f_{\text{min}} = \text{The minimum frequency range.} \]

\[ \text{The position of individual bat } t \text{ at time } t. \]

\[ z^* = \text{The globally optimal position.} \]

REFERENCES

[1] J. McBride, “Climate change, global population growth, and humanoid robots,” J. Future Robot Life, vol. 2, nos. 1–2, pp. 23–41, Dec. 2021.

[2] In Brief to The State of Food Security and Nutrition in the World, FAO, Rome, Italy, 2020.

[3] P. A. Sanchez, “Soil fertility and hunger in Africa,” Science, vol. 295, no. 5562, pp. 2019–2020, 2002.

[4] J. K. Steinberger and J. T. Roberts, “From constraint to sufficiency: The decoupling of energy and carbon from human needs, 1975–2005,” Ecol. Econ., vol. 70, no. 2, pp. 425–433, 2011.
[5] D. Mandal, M. Chandrakala, N. M. Alam, T. Roy, and U. Mandal, “Assessment of soil quality and productivity in different phases of soil erosion with the focus on land degradation neutrality in tropical humid region of India,” CATENA, vol. 204, Sep. 2021, Art. no. 105440.

[6] E. Assefa and B. Hans-Rudolf, “Farmers’ perception of land degradation and traditional knowledge in southern ethiopia-resilience and stability,” Land Degradation Develop., vol. 27, no. 6, pp. 1552–1561, Aug. 2016.

[7] M. Yuan, F. G. Fernández, C. M. Pettitkow, K. D. Greer, and D. Schaefer, “Soil and crop response to phosphorus and potassium management under conservation tillage,” Agronomy J., vol. 112, no. 3, pp. 2302–2316, May 2020.

[8] F. Schneider and W. A. Amelung Don, “Origin of carbon in agricultural soil profiles deduced from depth gradients of C:N ratios, carbon fractions, 413C and 815N values,” Plant Soil, vol. 460, no. 1, pp. 123–148, 2020.

[9] C. Hu, V. O. Sadrans, G. Lu, P. Zhang, Y. Han, L. Liu, J. Xie, X. Yang, and S. Zhang, “A global meta-analysis of split nitrogen application for improved wheat yield and grain protein content,” Soil Tillage Res., vol. 213, Sep. 2021, Art. no. 105111.

[10] A. Petrova, G. Sibgatullina, T. Gorshkova, and L. Kozlova, “Dynamics of cell wall polysaccharides during the elongation growth of RYE primary roots,” Plant, vol. 255, no. 5, pp. 1–17, May 2022.

[11] J. O. Omondi, N. Lazarovich, S. Rachmilevitch, T. Kukew, U. Yermiyahu, and H. Yasour, “Potassium and storage root development: Focusing on photosynthesis, metabolites and soluble carbohydrates in cassava,” Physiologia Plantarum, vol. 169, no. 2, pp. 169–178, Jun. 2020.

[12] R. Haynes and K. Goh, “Seasonal levels of available nutrients under grazed down, cultivated and zero tilled orchard soil management practices,” Soil Sci. Res., vol. 3, p. 363, 1978.

[13] C. Wang, B. Liu, L. Liu, Y. Zhu, J. Hou, P. Liu, and X. Li, “A review of synthetic color learning machine (SCLM) method and hyperspectral data,” Geoderma, vol. 380, p. 114664, Dec. 2020.

[14] I. Greenberg, M. Seidel, M. Vohland, and B. Ludwig, “Performance of soil quality and productivity in different phases of soil erosion with the focus on land degradation neutrality in tropical humid region of India,” CATENA, vol. 9, pp. 32517–32548, 2021.

[15] S. I. Hassan, M. M. Alam, U. Illahi, and M. A. Al Ghamdi, “A systematic review on monitoring and advanced control strategies in smart agriculture,” IEEE Access, vol. 9, pp. 32517–32548, 2021.

[16] Y.-P. Wang and Y. Shen, “Identifying and characterizing yield limiting soil factors with the aid of remote sensing and data mining techniques,” Precis. Agricult., vol. 16, no. 1, pp. 99–118, Feb. 2015.

[17] C. Chen, E. Arthur, M. Muller, H. Zhou, X. Wang, J. Shang, K. Huang, Y. Zha, and J. Yu, “Real-time detection of rice phenology through convolutional neural network using handheld camera images,” Precis. Agricult., vol. 22, no. 1, pp. 154–178, Feb. 2021.

[18] L. Lin and Z. X. G. Liu, “Estimation of soil total nitrogen using the synthetic color learning machine (SCLM) method and hyperspectral data,” Geoderma, vol. 380, p. 114664, Dec. 2020.

[19] J. Guigue, C. Just, S. Luo, M. Fagt, M. Schlötzer, I. Kögel-Knabner, and E. Hobley, “Spatial molecular heterogeneity of POM during decomposition at different soil depths resolved by VNIR hyperspectral imaging,” Eur. J. Soil Sci., vol. 73, Dec. 2021, Art. no. e13207.

[20] S. Pradhan, “Estimation of leaf chlorophyll content in wheat using hyperspectral vegetation indices,” Current Sci., vol. 119, no. 2, pp. 174–175, 2020.

[21] A. Siedliska, P. Banarowski, J. Pustuszk-Woźniak, M. Zubiak, and J. Krzyszczak, “Identification of plant leaf phosphorus content at different growth stages based on hyperspectral reflectance,” BMC Plant Biol., vol. 21, no. 1, p. 28, Dec. 2021.

[22] S. Rabab, P. Badenhorst, Y.-P.-P. Chen, and H. D. Daetwyler, “A template-free machine vision-based crop row detection algorithm,” Precis. Agricult., vol. 22, no. 1, pp. 124–153, Feb. 2021.

[23] J. Lu, W. Li, M. Yu, X. Zhang, Y. Ma, X. Su, X. Yao, T. Cheng, Y. Zhu, W. Cao, and Y. Tian, “Identification of rice plant phosphorus accumulation based on negative matrix factorization using hyperspectral reflectance,” Precis. Agricult., vol. 22, no. 1, pp. 51–74, Feb. 2021.

[24] J. B. Sankey, T. T. Sankey, J. Li, S. Ravi, G. Wang, J. Caster, and A. Kasprak, “Quantifying plant-soil-nutrient dynamics in rangelands: Fusion of UAV hyperspectral-LiDAR, UAV multispectral-photogrammetry, and ground-based LiDAR-digital photography in a shrub-encroached desert grassland,” Remote Sens. Environ., vol. 253, Feb. 2021, Art. no. 112223.

[25] F. Li, L. Xu, A. Wong, and D. A. Clausi, “Feature extraction for hyperspectral imagery via ensemble localized manifold learning,” IEEE Geosci. Remote Sens. Lett., vol. 12, no. 12, pp. 2486–2490, Dec. 2015.

[26] H. R. Shahdoozi and Z. Tabatabaei, “Object-based feature extraction for hyperspectral data using firefly algorithm,” Int. J. Mach. Learn. Cybern., vol. 11, no. 6, pp. 1277–1291, Jun. 2020.

[27] A. K. Patel, J. K. Ghosh, S. Pande, and S. U. Sayyad, “Deep-learning-based approach for estimation of fractional abundance of nitrogen in soil from hyperspectral data,” IEEE J. Sel. Topics Appl. Earth Observ., Remote Sens. 2020, vol. 13, pp. 6495–6511, 2020.

[28] W. Zhu, H. Yang, L. Xie, H. Li, L. Huang, Y. Zhao, and T. Yue, “Hyperspectral inversion of soil heavy metals in three-river source region based on random forest model,” CATENA, vol. 202, no. 12, 2021, Art. no. 10522.

[29] S. Delaileux, B. Somers, B. Haest, T. Spanhove, J. V. Borre, and C. A. Mühker, “Heathland conservation status mapping through integration of hyperspectral mixture analysis and decision tree classifiers,” Remote Sens. Environ., vol. 126, pp. 222–231, Nov. 2012.

[30] G. Wen, “Machine learning-based canola yield prediction for site-specific nitrogen recommendations,” Nutrient Cycling Agroecosystems, vol. 121, no. 2, pp. 241–256, 2021.
HUAN WANG was born in Shaanxi, China, in 1993. He received the master’s degree in mechanical engineering from Shihezi University, China, in 2019, where he is currently pursuing the Ph.D. degree in mechanical engineering. He joined the National Natural Science Foundation of China. His research interests include machine learning and smart agriculture.

LIXIN ZHANG was born in Shandong, in 1967. He is currently pursuing the Ph.D. degree in engineering. He is also a Professor. He is the Vice President of the School of Mechanical and Electrical Engineering, Shihezi University. He is the Head of the First-Level Discipline of Mechanical Engineering with the School of Mechanical and Electrical Engineering, Shihezi University. He is the Head of the Specialty of Mechanical Design, Manufacturing and Automation. He is an Editorial Board Member of Journal of Shihezi University (Natural Science Edition). He is a Correspondence Reviewer of the Department of Engineering and Materials Science, National Natural Science Foundation of China. He is mainly engaged in the direction with mechatronics technology, new energy, and smart agriculture.

JIAWEI ZHAO was born in Anhui, in 1995. He received the master’s degree in mechanical and electronic engineering from Shihezi University, China, in 1997. He is currently teaching with the Department of Electronic Information Engineering, School of Mechanical and Electrical Engineering, Shihezi University. He has four years of work experience in a large state-owned enterprise research institute, and has very rich experience in engineering construction. He is mainly engaged in the research of sensor detection, automatic control, and intelligent agriculture.

XUE HU, photograph and biography not available at the time of publication.

XIAO MA, photograph and biography not available at the time of publication.

***