A Hyperspectral Feature Selection Method for Soil Organic Matter Estimation Based on an Improved Weighted Marine Predators Algorithm

Kun Tan, *Senior Member, IEEE*, Libin Zhu, Xue Wang

***Abstract*—Soil organic matter (SOM) content is a crucial indicator for assessing soil fertility, and serves as a key factor in sustaining a viable agricultural system. With the continuous advancement and improvement of remote sensing technology, hyperspectral imagery has been employed for the monitoring of SOM. While the numerous bands reveal finer details within the spectral features, this also brings information redundancy and noise interference. Currently, the dimensionality reduction methods designed for hyperspectral imagery encounter difficulties in achieving optimal band combinations. As a result, swiftly and accurately capturing the spectral features of SOM becomes a challenging task. In this paper, aiming to address the inefficiency and instability in hyperspectral feature selection, we propose a metaheuristic-based algorithm—the improved weighted marine predators algorithm (IWMPA)—for hyperspectral feature selection. Specifically, we simulated the process of hyperspectral feature selection using the foraging strategy of marine predators. We employed prior weight coefficients and reverse learning operations in the initialization phase to accelerate the convergence of the population, and introduced mutation operations into the phase of development to prevent the occurrence of local optima traps. We employed the IWMPA feature selection method to establish SOM estimation models within the research area of Yitong Manchu Autonomous County in China. The results demonstrated that the hyperspectral features selected using the IWMPA approach yield favorable outcomes in the SOM estimation models. Specifically, in the best-performing regression model of this study, the R^2 on the test set was 0.7225. These experimental results suggest that, in comparison to the existing methods, the proposed IWMPA method is more adept at capturing the spectral features of SOM.**

***Index Terms*—Hyperspectral imagery, feature selection, marine predators algorithm, soil organic matter (SOM).**

Manuscript received xx 2024. This work is jointly supported by the Shanghai Municipal Science and Technology Major Project (grant no. 22511102800), the Natural Science Foundation of China (grant nos. 42171335), the National Civil Aerospace Project of China (grant no. D040102), the International Research Center of Big Data for Sustainable Development Goals (CBAS2022GSP07) and the Open Foundations of Jiangsu Province Engineering Research Center of Airborne Detecting and Intelligent Perceptive Technology (JSECF2023-10). (*Corresponding author: Xue Wang*).

Kun Tan, Libin Zhu, Xue Wang are with the Key Laboratory of Geographic Information Science(Ministry of Education), the Key Laboratory of Spatial-temporal Big Data Analysis and Application of Natural Resources in Megacities, Ministry of Natural Resources , and also with the School of Geographic Sciences, East China Normal University, Shanghai 200241, China (e-mail: tankuncu@gmail.com ; e-mail: zhulb@stu.ecnu.edu.cn ; e-mail: wx_cumt@yeah.net).

I. INTRODUCTION

SOIL forms the foundation of agricultural production and serves as a crucial medium for human habitat and survival[1]. Soil organic matter (SOM) content serves as a crucial indicator for assessing soil quality[2]. SOM refers to all the carbon-containing organic substances found within the soil, including plant and animal residues, as well as organic products resulting from biological activities[3,4]. SOM is crucial for maintaining soil fertility, water retention, nutrient cycling, and supporting diverse ecosystems[5]. It improves plant growth and enhances the physical properties of soil. It is also crucial for fostering soil structure formation, improving the physical traits of soil, and promoting fertilizer retention, and stands as a key determinant for crop yield[6]. Hence, the rapid and precise assessment of SOM holds immense significance in both monitoring soil fertility and driving advancements in agriculture. The conventional approach for obtaining SOM content typically involves collecting point-based field data and conducting chemical analyses in the laboratory. While this traditional approach offers the advantages of a high precision and accuracy, it comes with challenges, such as the substantial groundwork, high cost, extended timeframe, and the inability to achieve continuous assessment of SOM content across large geographical areas within a short time span[7-9]. As a result, obtaining spatial distribution data for SOM presents significant difficulties. Remote sensing, as an advanced technology, can provide support for the monitoring of SOM.

In recent years, the advancement of hyperspectral sensors has greatly heightened the acquisition capability of hyperspectral data, enabling the possibility of extensive and refined ground monitoring[10]. This advancement provides a substantial amount of hyperspectral data support for soil and environmental quality monitoring[11,12]. As is widely known, hyperspectral data possess the characteristic of a high spectral resolution, which provides rich spectral information[13]. However, the challenges of information redundancy and noise interference brought about by hyperspectral data can also affect the accuracy of SOM estimation. Therefore, it is necessary to perform dimensionality reduction when constructing a hyperspectral SOM regression model. There are two common methods for the dimensionality reduction of hyperspectral data: feature extraction and feature selection. Feature extraction is a method of transforming raw data into a new feature space, aiming to retain as much information as possible in a lower dimension. The common feature extraction methods include mathematical transformations, frequency domain

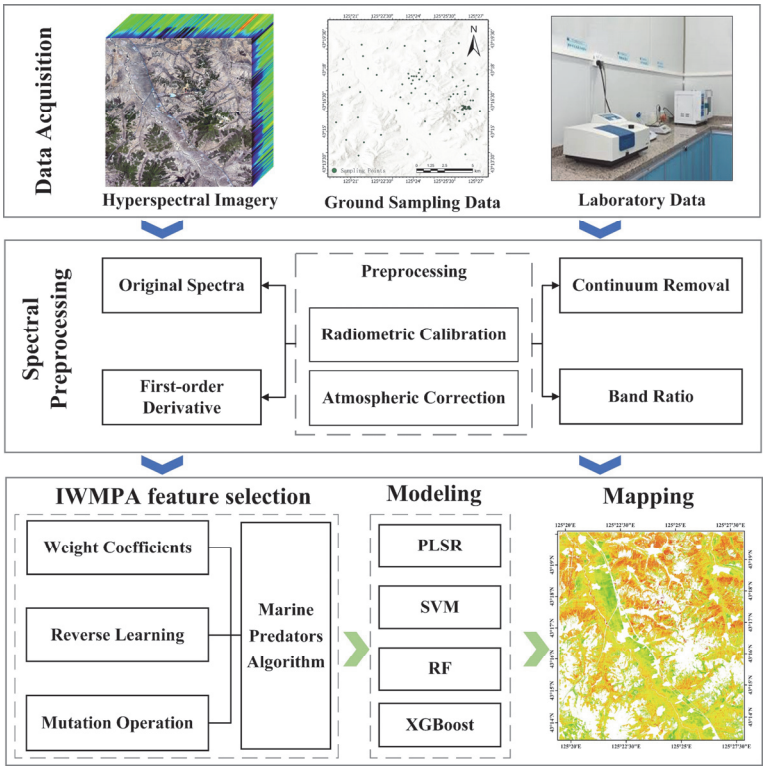

Fig. 1. Technical route.

transformations, and modified Gaussian transformations[14]. For example, Zhang et al.[15] applied first-order derivative, second-order derivative, continuum removal (CR), and wavelet transform preprocessing to the original spectra, effectively eliminating noise and baseline drift, and highlighting the positions of the spectral feature bands. Wu et al.[16] improved the quality of input data for the organic matter estimation model by applying characteristic transformations such as Savitzky-Golay filter, spectral continuum removal to the raw spectral curves of ZY1-02D ,and the model coefficient of determination reaches 0.829. Song et al.[17] utilized fractional order differentiation to preprocess the spectra, combined with PLSR model to enhance the spectral information of leaf, and successfully demonstrated that leaf spectra can be used to predict changes in photosynthetic capacity.

Feature selection is the process of selecting the optimal subset from all the given features using specific evaluation criteria. This procedure aims to identify the most valuable features for addressing problems or establishing models, thereby reducing the dimensionality, enhancing the model performance, minimizing the computational cost, and aiding in the elimination of irrelevant or redundant information[18-21]. Generally speaking, feature selection algorithms must take into account the following four factors: 1) the initial search point to define the starting point for exploration and the search direction for sequencing candidate subsets; 2) the search strategy for identifying the optimal subset within the search space; 3) an assessment function for evaluating the subset being considered; and 4) a termination criterion for determining when to stop the process. Commonly used feature selection methods include variable importance projection (VIP)[22], the successive projections algorithm (SPA)[23], the Pearson product-moment correlation coefficient (PPMCC)[24], competitive adaptive reweighted sampling (CARS)[25], the genetic algorithm (GA)[26], and others. Shi et al.[27] applied fractional order derivatives and baseline correction as feature extraction methods, combined with feature selection methods like CARS and VIP to estimate SOM in Xinjiang Uygur Autonomous Region of China. The results indicate that this combined approach notably improves the accuracy of organic matter estimation. Zhang et al.[28] utilized the CARS method to extract spectral features related to available copper (ACu) and SOM from Unmanned aerial vehicle hyperspectral data. Subsequently, they established estimation models for ACu and SOM in the tailings area of the Jianghan Plain, China. The results demonstrated the effectiveness of the established models in prediction. While the feasibility of utilizing the sensitive bands selected by the existing feature selection methods for estimating SOM content has been demonstrated, the complex environmental factors can disrupt the accuracy of soil parameter prediction in real-world scenarios. In addition, the strategies commonly adopted by the current feature selection methods often involve directly discarding features with lower importance. Furthermore, the presence of inherent random processes in the algorithms frequently leads to significant variations in the selected features, resulting in unstable algorithm performance that directly impacts the accuracy of SOM monitoring.

The marine predators algorithm (MPA)[29] is a metaheuristic optimization algorithm inspired by the survival of the fittest theory, where marine predators employ optimal foraging strategies during their Lévy flight or Brownian motion. The MPA adheres to a framework similar to that of other metaheuristic algorithms. It first employs a set of candidate

solutions for a given optimization problem. Then, by utilizing several mechanisms to enhance this set of candidate solutions, they transition from the initial random state to gradually approximating the global optimum for the given problem. There are three main phases in the MPA as follows: 1) high velocity, in which the speed of a marine predator is higher than that of the prey; 2) same velocity, in which the speed of a marine predator is identical to that of the prey; and 3) low velocity, in which the speed of a marine predator is less than that of the prey. The MPA possesses distinct advantages when compared to other optimization algorithms, and it is characterized by being derivative-free, parameter less, adaptable, robust, and comprehensive[30]. The MPA has been widely applied in the fields of optimization problems and path planning [31]. Moreover, many studies have explored its usage in high-dimensional hyperspectral feature selection. The MPA can improve the performance of classification and prediction models by selecting the most relevant features from the high-dimensional data[32]. The problem of feature selection can be viewed as an optimization problem within a binary search space. Therefore, researchers have modified the MPA to deal with optimization problems within a binary search space. For example, Elminaam et al.[33] proposed a binary version of the MPA for feature selection, referred to as the MPA with *k*-nearest neighbors (MPA-KNN) algorithm. The application of the sigmoid function transforms the optimization problem from the continuous MPA into a binary one. The simulation results obtained in this study confirmed the efficacy of the MPA-KNN algorithm, compared to other competing methods across all the evaluation metrics. Yousri et al.[34] proposed an improved MPA for global optimization and feature selection called the fractional-order comprehensive learning MPA (FOCLMPA). In the FOCLMPA method, the integration of the comprehensive learning strategy and memory perspective principles from fractional calculus with the MPA method was aimed at avoiding local solutions and mitigating premature convergence. In the field of hyperspectral classification, Shang et al.[35] conducted comparisons between various swarm intelligence algorithms and evolutionary algorithms for the task of hyperspectral feature selection. The results indicated that the MPA method exhibits a notable performance, securing a third-place ranking. Consequently, the MPA method is considered well-suited for feature selection tasks. However, in the field of hyperspectral inversion, research related to the MPA is still relatively limited.

In the field of hyperspectral inversion, our main focus is on the spectral features strongly correlated with the inversion parameters. The MPA can flexibly adjust its optimization strategy based on the characteristics of the inversion indicators and the requirements of the inversion modeling, enabling the selection of improved spectral features for SOM prediction. In this paper, to address the issue of indistinctive spectral features in the process of estimating SOM content using hyperspectral data, we present an improved weighted marine predators algorithm (IWMPA) for the purpose of feature selection. By incorporating reverse learning and feature weighting coefficients, we enhance the diversity of the initial population, thereby expediting the search efficiency. Simultaneously, we introduce the concept of mutation into the third phase of the algorithm, preventing the model from becoming trapped in a local optimum solution. The objectives of the present study were: 1) to develop a metaheuristic algorithm for feature selection to be employed in the hyperspectral inversion of SOM content; 2) to convert from the continuous IWMPA method to binary by using a threshold (0.5), enabling the algorithm to be suitable for the hyperspectral feature selection task; 3) to accelerate the convergence of the population by employing prior weight coefficients and reverse learning operations in the initialization phase; 4) to prevent the occurrence of local optima traps by introducing mutation operations into the phase of development; and 5) to apply the selected features to hyperspectral imagery and investigate the distribution patterns of SOM within the study area.

II. Study Area And Materials

A. Study Area and Experimental Design

The study area is situated within Yitong Manchu Autonomous County (125.33°E–125.47°E, 43.22°N–43.33°N), which is located in Jilin province, China. The study area spans approximately 139 km^2, with an average elevation of around 305m.

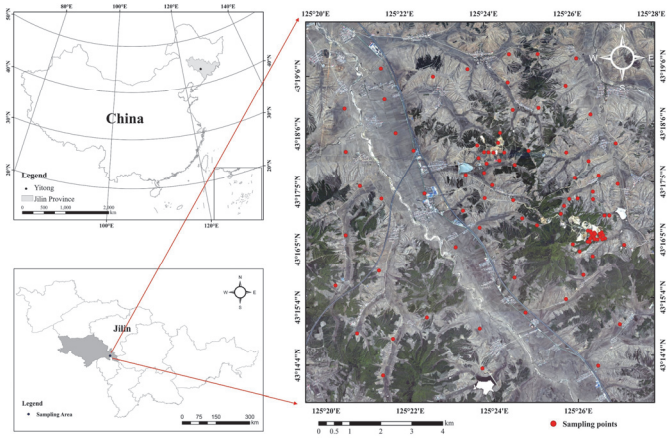

Fig. 2. Locations of the soil sampling points.

Surface soil (0–20cm) was collected from 93 sampling sites within the study area. The collected soil samples were then transported to the laboratory for the processes of air-drying, grinding, and passing through a 100-mesh sieve. The organic matter content of the samples was assessed using the potassium dichromate gravimetric method. The range of SOM values spanned from 14.76 g/kg to 49.84 g/kg, with an average of 30.48 g/kg. The standard deviation (SD) was calculated to be 6.38 g/kg, resulting in a coefficient of variation (CV) of 20.93%.

TABLE I
STATISTICAL CHARACTERIZATION OF THE SOM OF THE STUDY AREA

Max (g/kg)	Min (g/kg)	Mean (g/kg)	SD (g/kg)	CV (%)
49.84	14.76	30.48	6.38	20.93

B. Datasets and Preprocessing

In this study, the hyperspectral imagery was acquired between 18 April and 22 April 2017. The HyMap-C airborne

hyperspectral spectrometer consists of four sensors and covers a wavelength range from 400 nm to 2500 nm, encompassing a total of 144 spectral bands. Nine bands, primarily ranging from 400 to 450 nm, were excluded prior to the image preprocessing. After undergoing radiometric calibration, atmospheric correction, and geometric correction, along with the removal of water vapor bands, a total of 101 bands were retained for further analysis. More detailed HyMap-C parameters are shown in Table II. We utilized the Euclidean distance and spectral angle as constraints to augment the training dataset. This involved adding unlabeled samples that were spatially adjacent and spectrally similar to the new training set, in order to enhance the effectiveness of the model training and its stability.

TABLE II
HYMAP-C PARAMETER

Technical indicator	Parameter description
Number of bands	144
Spectral range	400-2500nm
Spatial resolution	4.5m
Spectral resolution	15nm(400-1440nm) 18nm(1440-2500nm)
Removed bands' wavelengths(nm)	1355, 1369, 1383, 1398, 1412, 1418, 1431, 1449, 1465, 1481, 1498, 1514, 1530, 1778, 1804, 1820, 1835, 1851, 1866, 1882, 1897, 1912, 1928, 1943, 1958, 1965, 1981, 1996, 2013, 2030, 2044, 2059, 2455, 2470

The First-order Derivative method, Continuum Removal, and Band Ratio method were used to extract features from the original spectral data. The following provides a detailed introduction to three preprocessing methods.

（1） First-order Derivative (FOD)

The principle of the first-order derivative is to compute the derivative of the original spectrum, which can mitigate the influence of atmospheric interference on reflectance and remove some linear background noise[36]. The calculation formula for the first-order derivative is as follows:

$$FOD_{\lambda_i} = \frac{R_{i+1} - R_{i-1}}{\lambda_{i+1} - \lambda_{i-1}} \quad (1)$$

In the formula, R_{λ_i} represents the reflectance value corresponding to the wavelength λ_i, λ_{i+1} and λ_{i-1} represent the wavelength values of two adjacent bands, FOD_{λ_i} represents the first-order derivative corresponding to the wavelength λ_i.The interval refers to the difference between the wavelength values of two adjacent bands.

（2） Continuum Removal (CR)

The Continuum Removal method first normalizes the spectral reflectance to help highlight the absorption and reflection features of the spectrum, and then identifies the maximum and minimum points on the spectral curve to obtain the envelope of each spectral curve[37]. The spectral curve after CR is calculated using Equation (2).

$$S_{cr} = S/C \quad (2)$$

In (2), S_{cr} represents the reflectance after CR, S denotes the original reflectance, and C represents the envelope reflectance value obtained for the corresponding band.

（3） Band Ratio(BR)

For each band, calculate the ratio with other bands one by one, as shown in Equation (3), and compute the correlation coefficient between each band's ratio spectrum and soil organic matter content[38]. Select the combination of band ratios for each band that has the highest correlation coefficient with organic matter content, and add it to the subset of features to be selected.

$$F_{\lambda_i} = \frac{R_{\lambda_i}}{R_{\lambda_j}} \quad (3)$$

In (3), R_{λ_i} represents the reflectance value at wavelength λ_i, and F_{λ_i} represents the spectrum after band ratio.

Ultimately, each sample was characterized by 385 spectral features, which include 101 original spectra, 84 FOD spectra (with 17 bands removed due to all-zero values), 99 spectra after CR (excluding the first and last spectra), and 101 BR spectra. The results are as shown in Fig. 3.

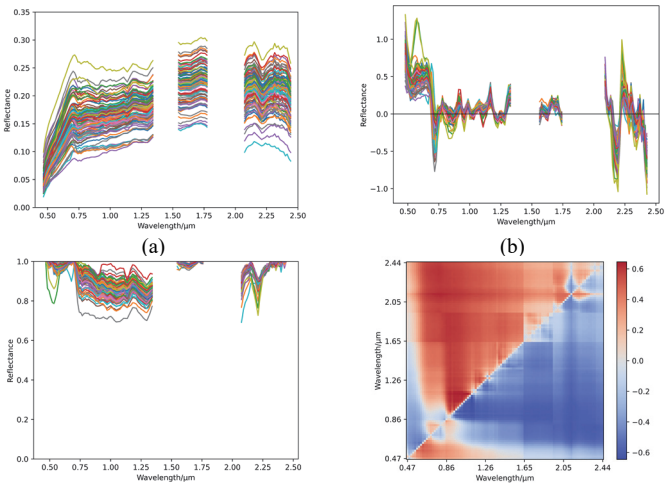

Fig. 3. Spectral preprocessing: (a) original spectra; (b) FOD; (c) CR; (d) BR.

III. METHODS

A. Marine Predators Algorithm (MPA)

The marine predators algorithm (MPA)[29] models the interactions between predator and prey in the ocean enviroent. This algorithm utilizes predator behaviors such as pursuit, predation, and competition to explore and discover the optimal solution for a given problem. The MPA includes the following main steps.

1) **Initialization**

In the MPA, the elite predator matrix and the predator matrix are represented as:

$$\vec{E} = \begin{bmatrix} X^I_{1,1} & X^I_{1,2} & \cdots & X^I_{1,d} \\ X^I_{2,1} & X^I_{2,2} & \cdots & X^I_{2,d} \\ \vdots & \vdots & \vdots & \vdots \\ X^I_{N,1} & X^I_{N,2} & \cdots & X^I_{N,d} \end{bmatrix}_{N^* D} \quad (4)$$

$$Pr = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,d} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,d} \\ \vdots & \vdots & \vdots & \vdots \\ X_{N,1} & X_{N,2} & \cdots & X_{N,d} \end{bmatrix}_{N^* D} \quad (5)$$

$$X_{i,j} = X_{min} + rand\left(X_{max} - X_{min}\right) \quad (6)$$

In (4), the vector $\overrightarrow{X^I}$ represents the top-level predator, which refers to the predator with the highest fitness. The vector $\overrightarrow{X^I}$ is duplicated N times to construct the elite matrix, which has the same dimension as $\vec{Pr}$. N is the number of search agents, and D represents the dimensionality of the search space, corresponding to the dimension of the features. In (5), the vector $X_{i,j}$ is randomly initialized according to (6), representing the position information of the ith prey in the jth dimension. In accordance with the requirements of the feature selection task, X_{min} and X_{max} are respectively set to 0 and 1. $rand$ represents generating a random number within the range of [0, 1].

2) **Exploration**

This phase occurs during the initial one-third of the total iterations, assuming that the predators' movement speed significantly surpasses that of the preys' speed. The update of positions and their iterations during this phase are described by (7). The mathematical model for this phase is as follows:

while $Iter < \frac{1}{3} Max\_Iter$

$$\vec{s}_i = \vec{R}_B \otimes \left(\vec{E}_i - \vec{R}_B \otimes \overrightarrow{Pr_i} \right) \; i = 1, ..., N \tag{7}$$

$$\overrightarrow{Pr_i} = \overrightarrow{Pr_i} + P.\vec{R} \otimes \vec{s}_i \tag{8}$$

where $\vec{s}_i$ represents the moving step length; $\vec{R}_B$ stands for a random vector based on the normal distribution of Brownian motion; $\otimes$ denotes the element-wise multiplication operation; P is a constant, which is equal to 0.5; $\vec{R}$ is uniformly distributed randomly in the range of [0, 1]; and *Max_iter* is the maximum number of iterations.

3) **Exploration and Development**

This phase marks the transition of the population from the exploration phase to the development phase. In this phase, both the predators and preys move at the same speed. The preys utilize *Levy* motion for development, while the predators continue to employ Brownian motion for exploration. This phase can be described as follows:

while $\frac{1}{3} Max\_Iter < Iter < \frac{2}{3} Max\_Iter$

if $i = 1, \ldots, N/2$

$$\vec{s}_i = \vec{R}_L \otimes \left(\vec{E}_i - \vec{R}_L \otimes \overrightarrow{Pr_i} \right) \tag{9}$$

$$\overrightarrow{Pr_i} = \overrightarrow{Pr_i} + P.\vec{R} \otimes \vec{s}_i \tag{10}$$

else

$$\vec{s}_i = \vec{R}_B \otimes \left(\vec{R}_B \otimes \vec{E}_i - \overrightarrow{Pr_i} \right) \tag{11}$$

$$\overrightarrow{Pr_i} = \vec{E}_i + P.CF \otimes \vec{s}_i \tag{12}$$

$$CF = \left(1 - \frac{Iter}{Max\_Iter}\right)^{\left(2 \times \frac{Iter}{Max\_Iter}\right)} \tag{13}$$

4) **Development**

During the final one-third of the iterations, the optimal strategy for predators is to utilize *Levy* motion, for which the mathematical model is as follows:

while $Iter > \frac{2}{3} Max\_Iter$

$$\vec{s}_i = \vec{R}_L \otimes \left(\vec{R}_L \otimes \vec{E}_i - \overrightarrow{Pr_i} \right) \; i = 1, ..., N \tag{14}$$

$$\overrightarrow{Pr_i} = \vec{E}_i + P.CF \otimes \vec{s}_i \tag{15}$$

5) **Eddy Formation and the Effect of Fish Aggregating Devices (FADs)**

The MPA takes the effect of eddy formation or fish aggregating devices (FADs) into account to avoid falling into a locally optimal solution, which is mathematically expressed as follows:

$$\overrightarrow{Pr_i} = \begin{cases} \overrightarrow{Pr_i} + CF\left[\vec{X}_{\min} + \vec{R} \otimes \left(\vec{X}_{\max} - \vec{X}_{\min} \right) \right] \otimes \vec{U} & \text{if } r \le \text{ FADs} \\ \overrightarrow{Pr_i} + [\text{ FADs}(1\text{-r}) + r] \left(\overrightarrow{Pr_{r1}} - \overrightarrow{Pr_{r2}} \right) & \text{if } r > \text{ FADs} \end{cases} \tag{16}$$

where $FADs = 0.2$ is the probability of the FADs' effect on the optimization process. $\vec{U}$ is a binary vector constructed by generating a random vector within the range of [0,1]. If an element in the vector is less than the FADs, the element is set to 0; if it is greater than the FADs, the element is set to 1. r is a uniform random number in the range of [0,1]. $\vec{X}_{min}$ and $\vec{X}_{max}$ are composed of elements consisting of 0s and 1s. *r1* and *r2* represent random indices of the prey matrix.

B. Improved Weighted Marine Predators Algorithm (IWMPA)

The IWMPA is an enhancement of the MPA. The following explanation outlines the improvements made in this study. The flowchart of the IWMPA is shown in Fig. 4.

1) **Initialization**

In the initialization phase of the population, incorporating weight coefficients can enhance the probability of selecting sensitive features. Given that the initial population is randomly generated, there is a possibility that certain features might consistently remain unselected. To ensure the completeness of the search space, a reverse learning operation is introduced. This operation aims to enhance the diversity of the initial population. Therefore, incorporating prior weight coefficients and reverse learning operations for population initialization contributes to the rapid convergence of the population toward the optimal solution. The search agent initialization in the IWMPA can be achieved through the following computation:

$$w_j = \frac{|b_j|}{\sum_{i=1}^{D} |b_j|} \tag{17}$$

$$X_{i,j} = 0.5 \times \left[X_{min} + rand\left(X_{max} - X_{min} \right) + w_j \right] \tag{18}$$

$$\overrightarrow{Pr'} = 1 - \overrightarrow{Pr} \tag{19}$$

$$\overrightarrow{Pr_{new}} = \begin{bmatrix} \overrightarrow{Pr} \\ \overrightarrow{Pr'} \end{bmatrix}_{2N \times D} \tag{20}$$

where b_j represents the weight coefficients calculated by partial least squares regression (PLSR) for the jth feature dimension. In (19), $\vec{Pr}$ is calculated using the $X_{i,j}$ from (18). In the IWMPA, the shape of $\vec{E}$ is $2N \times D$.

2) **Development**

In the third phase of the IWMPA, we introduce a mutation operation. This operation targets the search agents with the highest fitness within the population. It randomly selects certain features for modification, transitioning them from their initial unselected state to a selected state, and vice versa. The mathematical expression for this is as follows:

while $Iter > \frac{2}{3} Max\_Iter$

while *num*<*round*(*D*×*m*)

if $X_{i,fitness} > x$

$$X_{i,j}^{new} = 1 - X_{i,j}^{old} \tag{21}$$

where *m* is the mutation probability, which was set to 0.05 in this study; *round* is an integer function; *x* is the fitness of the search agents ranked in the last one-third; and $X_{i,j}^{new}$ is the value of the *j*th dimension of the *i*th search agent after mutation. $X_{i,j}^{old}$ is the value of the *j*th dimension of the *i*th search agent before mutation.

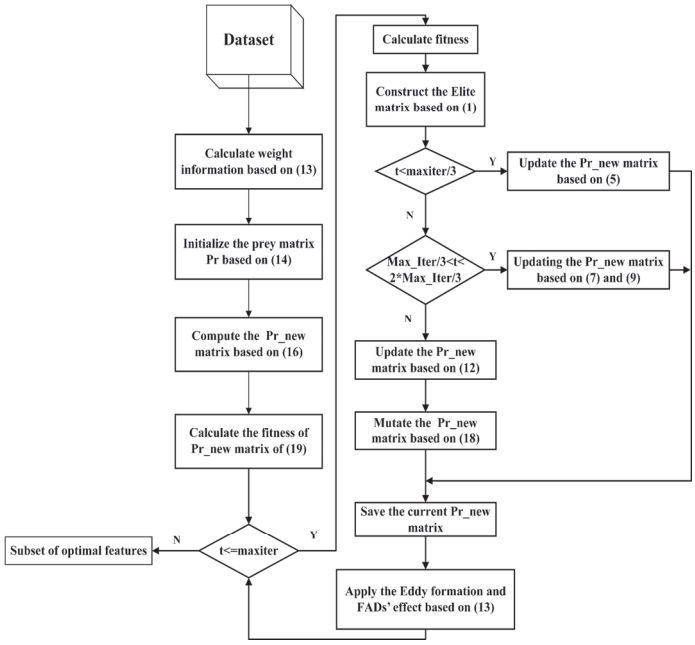

Fig. 4. IWMPA flowchart.

3) **Fitness Calculation**

The fitness of each search agent is determined through 10-fold cross-validation using PLSR, where both the *RMSE* and the number of selected features collectively determine the search agent's fitness. Each dimension of the search agent needs to go through the calculation of (22) to determine whether it is selected. The search agent with the highest fitness is then chosen as the current optimal search agent. The feature selection method and the fitness calculation formula are as follows:

$$B_{i,j} = \begin{cases} 1 & if\ X_{i,j} > threshold \\ 0 & otherwise \end{cases} \tag{22}$$

$$X_{i,fitness} = a \times RMSE + (1-a) \times \frac{\sum_{j=1}^{j=D} B_{i,j}}{D} \tag{23}$$

where the threshold is set to 0.5, the weight coefficient *a* is set to 0.8, *RMSE* is the root-mean-square error, and $X_{i,fitness}$ is the fitness of the *i*th search agent. It is noteworthy that the fitness calculation method used in MPA in this paper is consistent with that of IWMPA.

C. Comparison Algorithms

To validate the effectiveness of the approach proposed in this paper, the spectral feature data processed as described in Section II were subjected to both the IWMPA and four commonly used feature reduction methods. Subsequently, the reduced features obtained from each method were used for the subsequent modeling. The effectiveness of the proposed method was demonstrated by comparing both the number of reduced features and the modeling accuracy, thereby supporting the efficacy of the approach introduced in this paper.

1) **Principal Component Analysis (PCA)**

In principal component analysis (PCA)[39], high-dimensional data are transformed into a lower-dimensional format through linear transformation. The process includes projecting the data into the principal component space, effectively reducing the dimension while retaining the crucial information. We utilized PCA to extract the principal components contributing up to 99% of the variance, which were subsequently employed as the refined features for analysis.

2) **Variable Importance Projection (VIP)**

VIP[22] takes into account the correlation between independent and dependent variables. In the VIP method, the contribution of each feature is evaluated by training a partial least squares (PLS) model. The importance of the variables is assessed through their contributions to the projections within the model. Subsequently, the variables are ranked based on their contribution levels, and those with higher contributions are selected for further modeling analysis.

3) **Genetic Algorithm (GA)**

The GA[26] is a biomimetic optimization method. It simulates the process of biological evolution, starting from a population of candidate solutions and continuously evolving toward better solutions through operations such as crossover and mutation. In each generation, the quality of the solutions is evaluated using a fitness function. The GA is suitable for addressing problems with complex solution spaces, such as large-scale optimization and parameter tuning. In feature selection, the GA explores the feature space to obtain the optimal search agents or subsets, which represent the selected features. The hyperparameter settings in this study were as follows: the maximum number of iterations was 300, the number of search agents was 30, and the experiment count was 5.

4) **Competitive Adaptive Reweighted Sampling (CARS)**

CARS is a feature selection technique used to reduce the dimensionality of high-dimensional data. By iteratively selecting information-rich features and automatically adjusting the weights, CARS is able to effectively preserve the crucial features of the data.

D. Modeling Method

The prerequisite for employing statistical regression methods in SOM inversion is to ensure the efficacy of the feature reduction method. In this study, to mitigate the variability inherent in a single method, we utilized four common machine learning approaches to model the SOM assessment within the study area. By comparing the accuracies of the various modeling methods, the aim was to clarify the superior modeling capacity of features extracted using the IWMPA method.

1) **Partial Least Squares Regression (PLSR)**

PLSR[40] is a statistical learning method that establishes a predictive model and explores the relationships between variables. It combines the principles of PCA and linear regression to model the relationship between the input and output variables, making it robust for high-dimensional data and multicollinearity.

2) **Support Vector Machine (SVM)**

Support vector machine[41] (SVM) is a supervised machine learning algorithm. In regression tasks, it establishes a regression model by finding a hyperplane that maximizes the margin between the predicted and actual values, while ensuring that the prediction errors are controlled within a certain range.

3) **Random Forest (RF)**

Random forest[42] (RF) is an integrated learning method that performs a regression task by constructing multiple decision trees and averaging their predictions. Each decision tree is constructed based on a randomly selected subset of features and a randomly sampled training sample. In the prediction phase, the RF model calculates the average or weighted average of the predictions of each decision tree to obtain the final regression result.

4) **Extreme Gradient Boosting (XGBoost)**

Extreme gradient boosting[43] (XGBoost) is an improved version of the gradient boosting tree algorithm. It progressively improves the performance of the model by serially training multiple decision tree models and optimizing the loss function in each iteration.

E. Modeling Evaluation

In this paper, the evaluation metrics used to assess the accuracy of the models include the coefficient of determination (R^2), the root-mean-square error (*RMSE*), and the mean absolute error (*MAE*). The formulas for each metric are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \tag{24}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{25}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{26}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, $\bar{y}_i$ is the average value of all the actual values, and N represents the sample size.

IV. Results and Discussion

A. Modeling Results

In this study, we evaluated the proposed IWMPA method by comparing its results with the results of the commonly used dimensionality reduction methods described in Section III. For each dimensionality reduction method, we evaluated the inversion performance using four modeling approaches on an aerial hyperspectral dataset. The dataset was made up of organic matter assay data from 93 samples, and 385 spectral features were extracted from the airborne imagery, as described in Section II. In the experiments, these samples were divided into training and test sets with a ratio of 2:1. The performance of each model is listed in TABLE III .

TABLE III presents the modeling accuracy of each dimensionality reduction method. From TABLE III, we can infer that PCA shows the lowest performance. Despite its capability of preserving a significant amount of information while removing noise, the features extracted by PCA are unable to explain the spectral characteristics of SOM. PCA is guided by the aim of retaining as much information as possible, without being optimized for a specific task. Therefore, the features extracted may not effectively represent the characteristics of SOM. Both VIP and the GA incorporate the content of SOM as input to define the optimization objective of the algorithm. However, the features extracted by VIP show limited effectiveness in enhancing the model accuracy. On the other hand, although the GA contributes to improving the model accuracy, its failure to restrict the selected feature quantity prevents it from achieving effective spectral dimensionality reduction. In comparison to the preceding three classical algorithms, the MPA not only efficiently reduces the spectral dimensionality but also enhances the model accuracy. CARS and the IWMPA demonstrate the most superior predictive accuracy, with an 0.05 increase in R^2 on the test set. Notably, the IWMPA not only selects the fewest number of bands but also achieves the highest precision.

According to TABLE III, the modeling accuracy of the PLSR model is inferior to that of SVM, RF, and XGBoost. The PLSR model operates by extracting latent variables for regression modeling. Although the IWMPA and MPA utilize the cross-validation accuracy of the PLSR model for feature selection, their performance with the PLSR model is comparatively weaker than that for the other models. We believe there are two primary reasons for this: 1) SVM, RF, and XGBoost all possess the capability to handle non-linear relationships, enabling them to capture complex non-linear features within the data. The spectral features and organic matter content can exhibit non-linear relationships, while PLSR, as a linear method, might not fully capture these intricate relationships. 2) SVM, RF, and XGBoost exhibit greater robustness, making them more effective at handling noise and anomalous data. In addition, RF and XGBoost are ensemble learning algorithms that combine predictions from multiple models to reduce the overfitting risk, obtaining outstanding performances in handling complex datasets.

TABLE III
ESTIMATION RESULTS FOR SOM

Method	Model	Train			Test		
		R_c^2	$RMSE_c$	MAE_c	R_p^2	$RMSE_p$	MAE_p
PCA	PLSR	0.5968	4.1216	6.3778	0.3386	5.0640	6.8257
	SVM	0.7612	3.1714	1.8510	0.6347	3.7636	3.0929
	XGBoost	0.9999	0.0040	0.0027	0.6187	3.8445	3.2677
	RF	0.7337	3.3495	2.5884	0.5254	4.2896	3.4345
VIP	PLSR	0.5090	4.5481	6.1228	0.5025	4.3919	6.1872
	SVM	0.7168	3.4538	2.2059	0.5676	4.0941	3.2989
	XGBoost	0.9999	0.0009	0.0005	0.6265	3.8055	3.1563
	RF	0.7450	3.2776	2.6231	0.6459	3.7052	3.1000
GA	PLSR	0.5211	4.4917	6.1920	0.4660	4.5502	6.1877
	SVM	0.7189	3.4412	2.0539	0.6432	3.7193	2.7431
	XGBoost	0.9999	0.0239	0.0171	0.6542	3.6614	3.1214
	RF	0.7836	3.0190	2.3970	0.6540	3.6627	3.0476
CARS	PLSR	0.5527	4.3425	6.2667	0.4555	4.5946	6.1451
	SVM	0.8257	2.7090	1.5220	0.6919	3.4564	2.8891
	XGBoost	0.9585	1.3208	1.0397	0.6881	3.4772	2.6704
	RF	0.7782	3.0563	2.4388	0.6835	3.5031	2.8446
MPA	PLSR	0.5612	4.2995	6.2832	0.4940	4.4294	6.3739
	SVM	0.8103	2.8268	1.5334	0.6622	3.6191	2.8081
	XGBoost	0.9332	1.6773	1.2270	0.6544	3.6603	2.9371
	RF	0.7687	3.1216	2.4492	0.6705	3.5743	2.8374
IWMPA	PLSR	0.5324	4.4383	6.2285	0.5043	4.3840	6.3138
	SVM	0.7177	3.4484	2.2796	0.7041	3.3872	2.6053
	XGBoost	**0.9537**	**1.3966**	**1.0163**	**0.7225**	**3.2801**	**2.4928**
	RF	0.7485	3.2553	2.5456	0.7015	3.4019	2.6745

* R_c^2, $RMSE_c$, and MAE_c represent the evaluation metrics for the training set, while R_p^2, $RMSE_p$, and MAE_p represent the evaluation metrics for the test set. The highest precision values are highlighted in bold.

The scatter plots for the SOM estimation are shown in Fig. 5. The measured-predicted points are closely distributed around the 1:1 line for SOM, indicating stable model performances. Among the six methods, the IWMPA-XGBoost model obtains the highest accuracy on the test set.

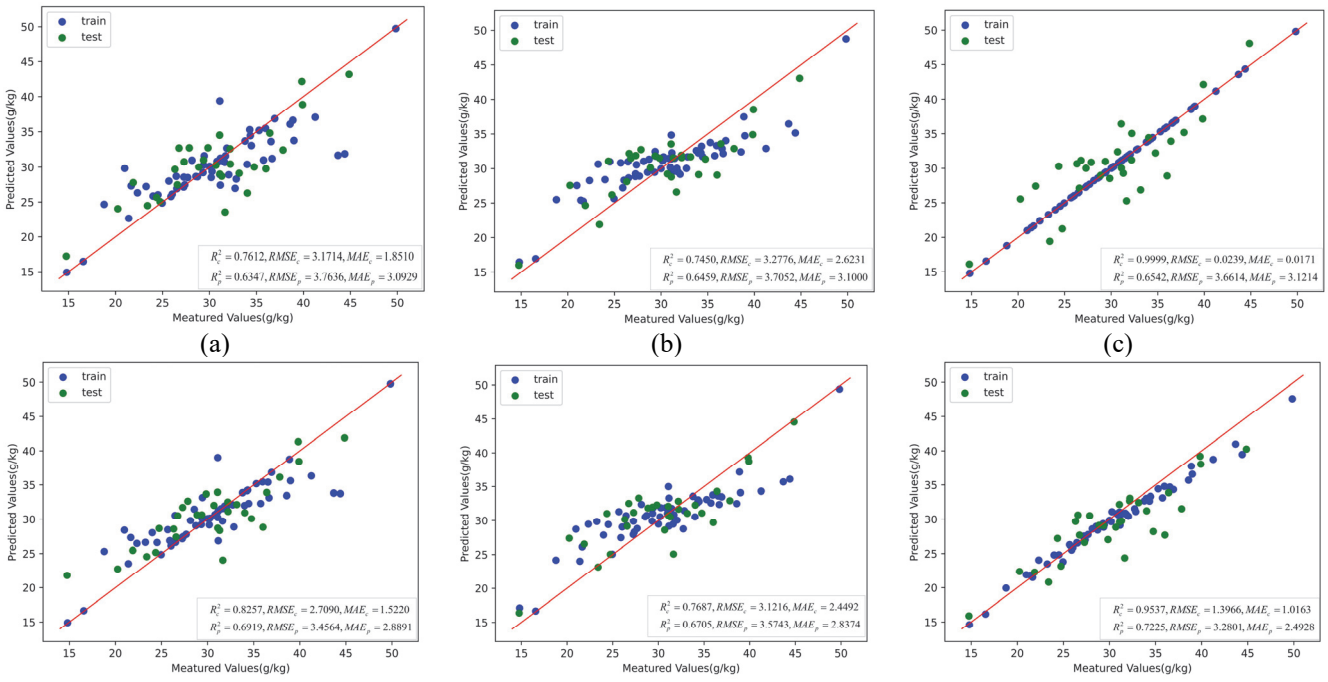

Fig. 5. The scatter plots for the SOM estimation. (a) PCA-SVM. (b) VIP-RF. (c) GA-XGBoost. (d) CARS-SVM. (e) MPA-RF. (f) IWMPA-XGBoost

TABLE IV
SPECTRAL CHARACTERISTIC BANDS OF SOM

Method	Quantity	Characteristic band wavelength (μm)
IWMPA	31	0.49, 0.51, 0.55, 0.59, 0.64, 0.65, 0.69, 0.7, 0.72, 0.77, 0.78, 0.82, 0.84, 0.89, 0.91, 0.97, 1.21, 1.65, 1.68, 1.69, 1.74, 1.76, 2.09, 2.11, 2.12, 2.18, 2.21, 2.3, 2.36, 2.41, 2.43
MPA	58	0.51, 0.56, 0.57, 0.59, 0.65, 0.67, 0.7, 0.72, 0.73, 0.75, 0.78, 0.79, 0.81, 0.82, 0.83, 0.84, 0.88, 0.89, 0.91, 0.93, 0.94, 0.97, 1, 1.02, 1.03, 1.17, 1.18, 1.2, 1.21, 1.23, 1.24, 1.25, 1.27, 1.28, 1.3, 1.31, 1.33, 1.34, 1.55, 1.56, 1.6, 1.65, 1.68, 1.71, 1.74, 1.76, 1.77, 2.15, 2.17, 2.18, 2.19, 2.23, 2.25, 2.28, 2.3, 2.34, 2.36, 2.41
GA	92	0.47, 0.48, 0.49, 0.52, 0.53, 0.55, 0.56, 0.59, 0.60, 0.61, 0.63, 0.64, 0.65, 0.67, 0.69, 0.70, 0.72, 0.73, 0.74, 0.75, 0.78, 0.79, 0.82, 0.83, 0.84, 0.86, 0.87, 0.88, 0.91, 0.93…
VIP	39	0.51, 0.52, 0.53, 0.57, 0.59, 0.60, 0.61, 0.64, 0.65, 0.67, 0.70, 0.72, 0.73, 0.75, 0.78, 0.81, 0.82,0.84, 0.86, 0.87, 0.88, 0.91, 0.96, 0.97, 1.00,1.14,1.17, 1.60, 1.65, 1.69, 1.76, 2.07, 2.12, 2.15, 2.24, 2.30, 2.34, 2.36, 2.41
CARS	64	0.48, 0.49, 0.51, 0.52, 0.53, 0.56, 0.57, 0.59, 0.6, 0.61, 0.67, 0.68, 0.7, 0.72, 0.74, 0.75, 0.78, 0.81, 0.82, 0.84, 0.86, 0.88, 0.89, 0.93, 0.96, 0.97, 1.03, 1.05, 1.08, 1.14, 1.17, 1.2, 1.23, 1.24, 1.25, 1.27, 1.31, 1.6, 1.63, 1.68, 1.69, 1.74, 1.76, 1.77, 2.07, 2.09, 2.12, 2.14, 2.15, 2.19, 2.21, 2.24, 2.25, 2.27, 2.28, 2.3, 2.31, 2.33, 2.34, 2.36, 2.37, 2.39, 2.4, 2.41

* The characteristic band consists of selected characteristic wavelengths, including both the selected wavelengths from the original spectrum and those from the feature-extracted spectrum.

B. Analysis of the Spectral Features

TABLE IV lists the bands associated with the features selected by each dimensionality reduction method. A smaller quantity signifies a higher concentration of feature bands, indicating an improved dimensionality reduction effect. From TABLE IV, it can be seen that the feature bands selected by the IWMPA range from 0.55 μm to 0.81 μm, which is consistent with the range of SOM spectral features reported by certain scholars[38,44,45]. In comparison to the five other comparison algorithms, the IWMPA method selects the fewest number of spectral features, indicating its effectiveness in the dimensionality reduction of hyperspectral data, while also being able to extract features related to organic matter. The GA method selects the maximum number of band features, which shows that it cannot effectively complete the dimensionality reduction task for hyperspectral data. Both the VIP and MPA methods can effectively reduce the dimension of the hyperspectral data, but the band combination selected by VIP does not greatly improve the modeling accuracy. The CARS method performs quite well in this experiment, with the modeling results only slightly behind those of the IWMPA method. The feature subset selected by CARS shows a high level of consistency with the results of the IWMPA method, which confirms the reliability of the chosen features.

C. Aerial Image Inversion Results

It can be seen from TABLE III that IWMPA-XGBoost model has the highest accuracy and the best inversion result. Fig. 6 displays the SOM estimation maps obtained using the various feature dimensionality reduction methods. These maps exhibit consistency in their spatial distribution, showcasing regions with high SOM values extending from the northwest to the southeast. The organic matter content in the study area is concentrated within the range of 0 to 50 g/kg. It is evident that regions with higher organic matter content are primarily situated along the riverbanks, whereas the soil surrounding the mountainous areas exhibits a lower organic matter content. From the distribution plot of the results, it is clear that the features selected by the IWMPA method more clearly reflect the full range of organic matter, whereas the other methods do not perform as well over the full range, which can be observed through the scatter plots of the modeling results. In comparison, the mapping results based on CARS and the IWMPA are the most favorable, accurately depicting the spatial distribution of SOM in the study area. Notably, the IWMPA exhibits greater clarity in capturing the spatial variations of SOM, indicating the superior performance of the proposed method in the spectral feature selection task.

D. Shortcomings and Prospects

The algorithm proposed in this paper has shown improvements in terms of reliability and stability, compared to the original MPA. However, the IWMPA, being an optimal solution optimization algorithm based on the foraging strategy of simulated marine predators, explores the search space in only one direction, limiting its ability to achieve optimal results and convergence. Moreover, due to the inability to navigate the search space, although the mutation operations can effectively alleviate the problem of getting stuck in local optima, the IWMPA still cannot guarantee reaching the global optimum. In addition, the IWMPA's exploration and development are limited in each stage, making it difficult to modify the model. In future research, it may be beneficial to incorporate the concept of multiple subpopulations, applying different optimization strategies to populations at various stages. This approach would allow the optimization to proceed in multiple directions, addressing the issue of a single optimization direction in the model. Recent advancements in artificial intelligence (AI) methodologies have opened new horizons for the extraction of information from hyperspectral imagery. Innovative models such as the Decoupled-and-Coupled Networks (DC-Net)[46] and SpectralGPT[47], proposed in recent literature, exemplify the significant potential of AI in effectively addressing hyperspectral feature selection challenges. These models, tailored for the intricacies of hyperspectral data, offer sophisticated mechanisms for

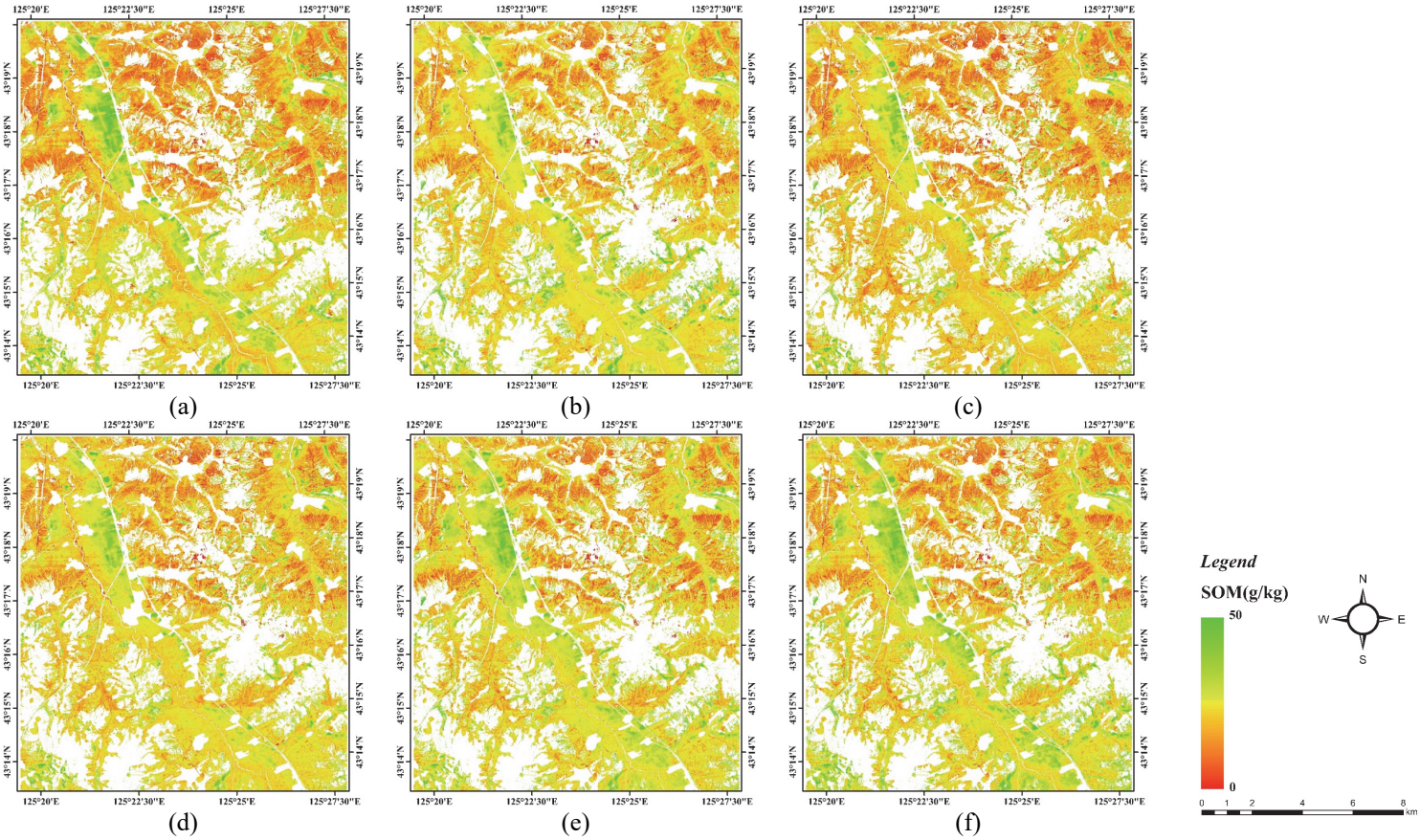

Fig. 6. SOM estimation maps for Yitong Manchu Autonomous County. (a) PCA. (b) VIP. (c) GA. (d) CARS. (e) MPA. (f) IWMPA

enhancing feature extraction and hold great promise for advancing soil analysis and related environmental applications.

V. Conclusion

In this paper, we have proposed a feature selection approach called the improved weighted marine predators algorithm (IWMPA), which is based on the marine predators algorithm (MPA). To enhance the exploration efficiency of the search agents within the feature space, the population is initialized using prior weights and reverse learning. In the development phase of the IWMPA, a mutation operation from GAs is introduced to prevent local optima trapping. The IWMPA incorporates cross-validation for assessing the feature subsets and includes constraints to address the feature inter-correlation. The effectiveness of this method was verified using HyMap-C airborne hyperspectral data from Yitong Manchu Autonomous County in China. The feature subsets extracted by the IWMPA outperformed the common feature selection methods in each inversion model, demonstrating enhanced stability, and the best inversion results were obtained with the XGBoost model (i.e., the highest $R_p^2 = 0.7225$). Compared with the MPA, the features selected by the IWMPA significantly improve the model accuracy for the estimation of SOM, which indicates that the improvement over the MPA is effective. We finally constructed fine distribution trend maps for the SOM content in the study area with the IWMPA and the other comparison methods. These maps show similar trends, which validates the reliability of the proposed IWMPA method.

The burgeoning field of AI presents a transformative potential for soil environmental monitoring. As AI continues to gain momentum, the prospect of integrating sophisticated deep learning methodologies into soil inversion research emerges as a promising avenue. Future endeavors may explore the amalgamation of these advanced AI techniques with hyperspectral analysis to enhance the fidelity and efficacy of soil environmental quality assessments.

Acknowledgment

This research is supported in part by Shanghai Municipal Science and Technology Major Project (grant no. 22511102800), Natural Science Foundation of China (grant nos. 42171335), National Civil Aerospace Project of China (grant no. D040102), International Research Center of Big Data for Sustainable Development Goals (CBAS2022GSP07) and the Open Foundations of Jiangsu Province Engineering Research Center of Airborne Detecting and Intelligent Perceptive Technology (JSECF2023-10).

References

[1] J.L. Smith, J.J. Halvorson, R.I. Papendick. “Using multiple-variable indicator kriging for evaluating soil quality,” *Soil Sci. Soc. Am. J.*, vol. 57, no. 3, pp. 743-749, May. 1993.

[2] A.B. Mcbratney, U. Stockmann, D.A. Angers, B. Minasny, D.J. Field. “Challenges for soil organic carbon research,” *Soil carbon*, pp. 3-16, Jan. 2014.

[3] N. Demirel, Ş. Düzgün, M.K. Emil. "Landuse change detection in a surface coal mine area using multi-temporal high-resolution satellite images," *Int. J. Min., Reclam. Environ.*, vol. 25, no. 4, pp. 342-349, Dec. 2011.

[4] Q.-Q. Zhou, J.-L. Ding, M. Tang, B. Yang. "Inversion of soil organic matter content in oasis typical of arid area and its influencing factors," *Acta Pedol. Sin.*, vol. 55, no. 2, pp. 1-13, Jan. 2018.

[5] H. Tiessen, E. Cuevas, P. Chacon. "The role of soil organic matter in sustaining soil fertility," *Nature*, vol. 371, no. 6500, pp. 783-785, Oct. 1994.

[6] M.E. Parolo, M.C. Savini, R.M. Loewy. "Characterization of soil organic matter by ft-ir spectroscopy and its relationship with chlorpyrifos sorption," *J. Environ. Manage.*, vol. 196, no. pp. 316-322, Jul. 2017.

[7] P. Qiao, M. Lei, S. Yang, J. Yang, G. Guo, X. Zhou. " Comparing ordinary kriging and inverse distance weighting for soil as pollution in beijing," *Environ. Sci. Pollut. Res.*, vol. 25, pp. 15597-15608, Mar. 2018.

[8] P. Su, D. Lin, C. Qian. "Study on air pollution and control investment from the perspective of the enviroental theory model: A case study in china, 2005–2014," *Sustainability*, vol. 10, no. 7, pp. 2181, Jun. 2018.

[9] E. Ben-Dor, A. Banin. "Near-infrared analysis as a rapid method to simultaneously," *Soil Sci. Soc. Am. J.*, vol. 59, no. 2, pp. 364-372, Mar-Apr. 1995.

[10] L. Chen, J. Lai, K. Tan, X. Wang, Y. Chen, J.J.S.O.T.T.E. Ding. "Development of a soil heavy metal estimation method based on a spectral index: Combining fractional-order derivative pretreatment and the absorption mechanism," *Sci. Total Environ.*, vol. 813, no. pp. 151882, Mar. 2022.

[11] K. Tan, W. Ma, L. Chen, H. Wang, Q. Du, P. Du, B. Yan, R. Liu, H. Li. "Estimating the distribution trend of soil heavy metals in mining area from hymap airborne hyperspectral imagery based on ensemble learning," *J. Hazard. Mater.*, vol. 401, pp. 123288, Jan. 2021.

[12] K. Tan, H. Wang, L. Chen, Q. Du, P. Du, C. Pan. "Estimation of the spatial distribution of heavy metal in agricultural soils using airborne hyperspectral imaging and random forest," *J. Hazard. Mater.*, vol. 382, pp. 120987, Jan. 2020.

[13] J. Benediktsson, J. Sveinsson. "Feature extraction for multisource data classification with artificial neural networks," *Int J Remote Sens*, vol. 18, no. 4, pp. 727-740, Nov. 1997.

[14] Y. Fenghua. "Retrieving nutrient information of japonica rice based on uanned aerial vehicle hyperspectral remote sensing," *Shenyang Nongye Daxue Xuebao*, vol. 10, no. 4,pp. 150-157, Jul. 2017.

[15] S. Zhang, Q. Shen, C. Nie, Y. Huang, J. Wang, Q. Hu, X. Ding, Y. Zhou, Y. Chen. "Hyperspectral inversion of heavy metal content in reclaimed soil from a mining wasteland based on different spectral transformation and modeling methods," *Spectrochim. Acta, Part A*, vol. 211, pp. 393-400, Mar. 2019.

[16] M. Wu, S. Dou, N. Lin, R. Jiang, B.J.R.S. Zhu. "Estimation and mapping of soil organic matter content using a stacking ensemble learning model based on hyperspectral images," *Remote Sens.*, vol. 15, no. 19, pp. 4713, Sept. 2023.

[17] G. Song, Q. Wang, J. Jin. "Fractional-order derivative spectral transformations improved partial least squares regression estimation of photosynthetic capacity from hyperspectral reflectance," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, no. pp. 1-10, Apr. 2023.

[18] S.B. Serpico, L. Bruzzone. "A new search algorithm for feature selection in hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1360-1367, Jul. 2001.

[19] S. Li, H. Wu, D. Wan, J. Zhu. "An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine," *Knowl Based Syst*, vol. 24, no. 1, pp. 40-48, Feb. 2011.

[20] B.-C. Kuo, H.-H. Ho, C.-H. Li, C.-C. Hung, J.-S. Taur. "A kernel-based feature selection method for svm with rbf kernel for hyperspectral image classification," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 7, no. 1, pp. 317-326, May. 2013.

[21] J. Zhao, M. Karimzadeh, A. Masjedi, T. Wang, X. Zhang, M.M. Crawford, D.S. Ebert. "Featureexplorer: Interactive feature selection and exploration of regression models for hyperspectral images," in *VIS*, Vancouver, BC, Canada, 2019, pp. 161-165.

[22] S. Wold, H. Martens, H. Wold. "The multivariate calibration problem in chemistry solved by the pls method, " Matrix Pencils, Pite Havsbad, Sweden, 1982, pp. 286-293.

[23] B. Tan, Y. Liang, L. Yi, H. Li, Z. Zhou, X. Ji, J. Deng. "Identification of free fatty acids profiling of type 2 diabetes mellitus and exploring possible biomarkers by gc–ms coupled with chemometrics," *Metabolomics*, vol. 6, no. pp. 219-228, Jun. 2010.

[24] Z. Aiwu, D. Nan, K. Xiaoyan, G. Chaofan. "Hyperspectral adaptive band selection method through nonlinear transform and information adjacency correlation," *Infrared Laser Eng.*, vol. 46, no. 5, pp. 538001-0538001 (0538009), Jun. 2017.

[25] H. Li, Y. Liang, Q. Xu, D. Cao. "Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration," *Anal. Chim. Acta*, vol. 648, no. 1, pp. 77-84, Aug. 2009.

[26] J.H. Holland. "Genetic algorithms," *SciAm*, vol. 267, no. 1, pp. 66-73, Jul. 1992.

[27] X. Shi, J. Song, H. Wang, X. Lv, Y. Zhu, W. Zhang, W. Bu, L. Zeng. "Improving soil organic matter estimation accuracy by combining optimal spectral preprocessing and feature selection methods based on pxrf and vis-nir data fusion," *Geoderma*, vol. 430, no. pp. 116301, Feb. 2023.

[28] Y. Zhang, L. Wei, Q. Lu, Y. Zhong, Z. Yuan, Z. Wang, Z. Li, Y.J.E.P. Yang. "Mapping soil available copper content in the mine tailings pond with combined simulated annealing deep neural network and uav hyperspectral images," *Environ. Pollut.*, vol. 320, no. pp. 120962, Mar. 2023.

[29] A. Faramarzi, M. Heidarinejad, S. Mirjalili, A.H. Gandomi. "Marine predators algorithm: A nature-inspired metaheuristic," *Expert Syst. Appl.*, vol. 152, no. pp. 113377, Aug. 2020.

[30] M.A. Al-Betar, M.A. Awadallah, S.N. Makhadmeh, Z.a.A. Alyasseri, G. Al-Naymat, S. Mirjalili. "Marine predators algorithm: A review," *Arch. Comput. Methods Eng.*, vol. 30,no. pp. 3405-3435, April. 2023.

[31] M. Ramezani, D. Bahmanyar, N. Razmjooy. "A new improved model of marine predator algorithm for optimization problems," *Arab J Sci Eng*, vol. 46, no. 9, pp. 8803-8826, May. 2021.

[32] L. Yang, Q. He, L. Yang, S. Luo. "A fusion multi-strategy marine predator algorithm for mobile robot path planning," *Appl. Sci.*, vol. 12, no. 18, pp. 9170, Sep. 2022.

[33] D.S. Abd Elminaam, A. Nabil, S.A. Ibraheem, E.H. Houssein. "An efficient marine predators algorithm for feature selection," *IEEE Access*, vol. 9, pp. 60136-60153, Apr. 2021.

[34] D. Yousri, M. Abd Elaziz, D. Oliva, A. Abraham, M.A. Alotaibi, M.A. Hossain. "Fractional-order comprehensive learning marine predators algorithm for global optimization and feature selection," *Knowl Based Syst*, vol. 235, no. pp. 107603, Jan. 2022.

[35] Y. Shang, X. Zheng, J. Li, D. Liu, P. Wang. "A comparative analysis of swarm intelligence and evolutionary algorithms for feature selection in svm-based hyperspectral image classification," *Remote Sens.*, vol. 14, no. 13, pp. 3019, Jun. 2022.

[36] S. Wang, K. Guan, C. Zhang, D. Lee, A.J. Margenot, Y. Ge, J. Peng, W. Zhou, Q. Zhou, Y.J.R.S.O.E. Huang. "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.,* vol. 271, no. pp. 112914, Mar. 2022.

[37] Y. Wang, X. Zhang, W. Sun, J. Wang, S. Ding, S.J.S.O.T.T.E. Liu. "Effects of hyperspectral data with different spectral resolutions on the estimation of soil heavy metal content: From ground-based and airborne data to satellite-simulated data," *Sci. Total Environ.*, vol. 838, no. pp. 156129, Sept. 2022.

[38] L. Zhao, K. Tan, X. Wang, J. Ding, Z. Liu, H. Ma, B.J.R.S. Han. "Hyperspectral feature selection for som prediction using deep reinforcement learning and multiple subset evaluation strategies," *Remote Sens.*, vol. 15, no. 1, pp. 127, Dec. 2022.

[39] K. Pearson. "Liii. On lines and planes of closest fit to systems of points in space," *Lond. Edinb. Dublin philos. mag. j. sci.*, vol. 2, no. 11, pp. 559-572, 1901.

[40] P. Geladi, B.R. Kowalski. "Partial least-squares regression: A tutorial," *Anal. Chim. Acta*, vol. 185, no. pp. 1-17, 1986.

[41] A.J. Smola, B. Schölkopf. "A tutorial on support vector regression," *Stat Comput*, vol. 14, no. pp. 199-222, Aug. 2004.

[42] L. Breiman. "Random forests," *Mach Learn*, vol. 45, no. pp. 5-32, Oct. 2001.

[43] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou. "Xgboost: Extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1-4, Dec. 2015.

[44] D. Ou, K. Tan, J. Lai, X. Jia, X. Wang, Y. Chen, J. Li. "Semi-supervised dnn regression on airborne hyperspectral imagery for improved spatial soil properties prediction," *Geoderma*, vol. 385, no. pp. 114875, Mar. 2021.

[45] G. Zheng, R. Dongryeol, J. Caixia, H. Changqiao. "Estimation of organic matter content in coastal soil using reflectance spectroscopy," *Pedosphere*, vol. 26, no. 1, pp. 130-136, Feb. 2016.

[46] D. Hong, J. Yao, C. Li, D. Meng, N. Yokoya, J. Chanussot. "Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Oct. 2023.

[47] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia. "Spectralgpt: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1-18, Apr. 2024.