



OPEN ACCESS

EDITED BY

Pu Xia,
University of Saskatchewan, Canada

REVIEWED BY

Shaohua Lei,
Nanjing Hydraulic Research Institute, China
Huanxue Zhang,
Shandong Normal University, China

*CORRESPONDENCE

Huazhou Chen,
✉ hzchengut@foxmail.com

RECEIVED 29 June 2025

REVISED 27 November 2025

ACCEPTED 08 December 2025

PUBLISHED 10 February 2026

CITATION

Hong S, Zhang S, Lin M, Meng F, Hou C and Chen H (2026) Estimation of heavy metals in agricultural soil using near-infrared spectroscopy with the improved evolutionary method.
Front. Environ. Sci. 13:1656095.
doi: 10.3389/fenvs.2025.1656095

COPYRIGHT

© 2026 Hong, Zhang, Lin, Meng, Hou and Chen. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Estimation of heavy metals in agricultural soil using near-infrared spectroscopy with the improved evolutionary method

Shaoyong Hong¹, Siyuan Zhang¹, Miaoquan Lin², Fangxiu Meng³, Can Hou¹ and Huazhou Chen^{3*}

¹School of Artificial Intelligence, Guangzhou Huashang College, Guangzhou, China, ²Mathematics Teaching and Research Group, Rongxian Experimental Senior Middle School, Yulin, China, ³School of Mathematics and Statistics, Guilin University of Technology, Guilin, China

Introduction: Monitoring and real-time detection of soil heavy metal pollution are crucial to ensuring the safety of agricultural food products. Conventional methods for determining heavy metal concentrations require substantial time and costs. This study investigates intelligent modeling approaches for online detection of heavy metal contamination in soils. Based on Internet of Things communication, large-scale near-infrared (NIR) spectral data were collected from distributed sensors for federated analysis.

Methods: In chemometric studies, the improved binary firefly algorithm (IBFA) is proposed for evolutionary variable selection, and the modified method of maximum information coefficient (MMIC) is designed to estimate the nonlinear correlation of unevenly distributed samples. Experimental soil data are collected from the Karst geology on the north side of Guangxi ZAR, China. The NIR calibration model is established by fusion of the IBFA and the MMIC methods (denoted as IBFA + MMIC). The fusion model is applied for quantitative prediction targeting of four heavy metals in the Karst soil samples.

Results: The IBFA + MMIC model can observe correlation coefficients higher than 0.9 and the lowest prediction errors during model training. It is tested with the correlations very close to 0.9, while the testing errors are acceptably low. These results outperform the counterpart models established by other cross combinations of firefly algorithm (FA)/IBFA and maximum information coefficient (MIC)/MMIC.

Discussion: The proposed modeling methodology is effectively validated for quantitative NIR analysis of different heavy metals in Karst soil data. Meanwhile, it provides critical technical support for the federated analytical performance of distributed sensing data, thereby facilitating precision soil management practices.

KEYWORDS

distributed sensing, evolutionary method, firefly algorithm, modified maximum information coefficient, near-infrared spectroscopy, soil heavy metal

1 Introduction

Heavy metal contamination of soil is a grave and growing environmental concern (Zhang X. et al., 2024). Heavy metals enter the soil through fertilizer applications, mining activities, and sewage discharge, gradually accumulating and leading to contamination (Latosinska et al., 2021). Due to its long-term environmental persistence without degradation, this kind of contamination has detrimental effects on the ecosystem, resulting in a cascading impact on soil quality. Then, it is passed along the food chain, ultimately delivering harmful risks to human health (Perkovic et al., 2022). The monitoring and real-time detection of soil heavy metal pollution is essential for protecting the safety of agricultural foods. The conventional method for determining soil heavy metal concentrations is chemical analysis. This involves collecting soil samples from the field and measuring the target heavy metal content under laboratory settings and conditions. However, this process is time-consuming, laborious, and expensive; the number of available soil samples is limited. It does not lend itself to performing high-quality online studies to identify heavy metal variation.

Over the past few decades, near-infrared (NIR) spectral technology, which can obtain spectral responses from the samples and yield detailed spectral information about agricultural objects, has been developing (De Souza et al., 2016; Nawar et al., 2023). Supported by the Fourier transform (FT) technique, the NIR signals are amplified to a magnitude range that is easily distinguished. The FT-NIR sensing signal boasts wide-ranging continuous spectral bands with high resolution, providing rapid and precise estimation of soil sample components (Mortada et al., 2021). Specifically, the FT-NIR spectra of soil (scanning range at 10,000–4,000 cm^{-1}) reflect multiple cumulative properties of soil heterogeneous matters, such as soil minerals, nutrient targets, and moisture (Leenen et al., 2019). With the increasing demand for spectral prospecting, the “big data” concept of FT-NIR sensing of large quantities of soil samples enables the identification of the inherent components and extra substances in soil (including heavy metals).

When spectral sensing technology was initially applied for field detection of soil compositions, many studies focused on spectrally active soil analytes like organic carbon, nutrient components, iron oxides, and clay minerals (Paltseva et al., 2022; Al Maliki et al., 2018). These analytes have a direct relation with the spectral response. The main task is to identify the NIR bands to interpret their spectral features, which is feasible (Zhang et al., 2019). In contrast to the spectrally active analytes, heavy metal concentrations are more difficult to estimate using the spectral data in the NIR region due to their low concentrations in soil (Han et al., 2021). Successful prediction of soil heavy metals is carried out by laboratory experiments under preset ideal conditions (Zukowska et al., 2021). Because of the complex mechanism for the retrieval of soil contaminants, laboratory-based spectroscopy is often used to explore the relationship between an individual heavy metal and the spectrum (Ali et al., 2023).

Compared to laboratory measurements, the Internet of Things (IoT) framework endorses immediate sensing of soil properties with portable devices, which is quicker and less expensive (Dattatreya et al., 2024). Considering the spectral variation caused by field

conditions, several studies focus on the extraction of spectral features from the online measured data (Wang et al., 2022). In combination with machine learning algorithms, calibration models are improved in terms of prediction accuracy and model stability (Zayani et al., 2023). These successful studies have provided a robust foundation for the development of adaptive learning methods in estimating soil metals with NIR technology (Li et al., 2024). In combination with the IoT framework, the spectra with internal signal pre-correction by the instrument are used to detect various alien contaminating elements over a large area of distributed spot-line geographical locations (Chen et al., 2020). Under the “big data” situation, the chemometric methods must be enhanced in fusion with some adaptive learning strategies to fill the gap between laboratory analysis and the IoT-based federal analysis of distributed sensing data. Prompt detection based on analysis of data distributed over a wide range of locations is currently emerging. It has become a new state-of-the-art application of NIR technology to support rapid detection and instant analysis in fields of agricultural and environmental sciences.

The term “distributed sensing technique” refers to real-time monitoring and immediate response by deploying multiple sensors to deal with fertilization, irrigation, defect treatment, and other relevant issues (Acharya and Kogure, 2023). The sensors are usually small or portable with low power support and communicate wirelessly with each other or with a base station (Chamara et al., 2023). Distributed sensing is convenient and useful for simultaneous data collection, good protection of data privacy, reduction of the system complexity, operating with low costs, and provision of a comprehensive view of the environment being monitored (Bourechak et al., 2023). There have been a few initial applications of rapid NIR/FT-NIR detection to environmental targets in food and agriculture. These studies use traditional distributed sensing and a data aggregation algorithm to search for solutions to field problems (Gouda et al., 2024; Chen et al., 2025). However, data aggregation requires substantial computing and communication resources. This scheme has poor performance in responding to instant and precise responses from portable smart devices such as tablets and mobile phones under the IoT framework (Zhou et al., 2024). In a scenario of rapid spectral analysis over a large area of geographical spots, computing resources must encode and decode the transmitted data frequently, which increases the energy consumption of the distributed sensing system (Babbar et al., 2025).

Evolutionary learning is a type of intelligent algorithm that can ease the computing intensity of encoding and decoding aggregated data. The principle of evolution depends on iterative population processes, which perform with superior individuals selected for reproduction and inferior ones eliminated in swarm selection (Zhang R. et al., 2024). The application of NIR spectroscopy in an IoT scenario partially benefits from this advanced property of evolutionary learning. In recent years, several evolutionary learning methods have become widely used and well-validated for NIR analysis (Zhang et al., 2021; Liu et al., 2024). As a result, studies that employed NIR/FT-NIR spectroscopy adopted evolutionary learning methods for quantitative or qualitative predictions on different soil metal substances in various land-use types. For example, Surawijaya et al. (2023) employed particle swarm optimization (PSO) to explore an optimized design of a

fabricated grating structure, aiming for maximum NIR absorption. This was done to improve the calibration models for predicting silicon (Si) concentration in soil samples. Song et al. (2022) developed a binary grey wolf optimization (GWO) method combined with a PLS model for the determination of arsenic (As) and Pb concentrations in soil. This approach reduced the dimensionality of input data and increased the NIR predictive performance. Among various evolutionary algorithms, the firefly algorithm (FA) conducts global search across the entire space by imitating the brightness attraction mechanism among fireflies. The FA outperforms other evolutionary methods in several aspects. First, compared with other evolutionary algorithms, the FA must only adjust a small number of representative parameters (such as brightness attractiveness, light absorption coefficient, and step factor); it has stronger robustness in global optimization than GA, PSO, or GWO (Chhabra et al., 2021). Second, FA can achieve more flexible optimization according to specific problems, which is conducive to balancing the adaptive search ability and secondary development of the algorithm (Devi et al., 2022). Third, the FA algorithm can find better solutions within fewer iterations, and its optimization convergence speed is faster than other evolutionary algorithms (Tao et al., 2024). Many published articles have used the classical FA as an important chemometric support for NIR quantitative or qualitative analysis. For instance, it is used for the estimation of different ripening stages of apples, for the detection of nutrient freshness of tomato plants, and for the quality determination of grape seed oil (Pourdarbani et al., 2022; Li, 2021). FA has not yet been used as a machine learning tool for NIR detection of soil nutrition or pollutant substances. In addition, the previous works referring to evolutionary learning methods were conducted using laboratory or local field spectra, not for distributed sensing analysis in an IoT scenario involving “big data” over a large area of diverse geographical locations.

This study explored an intelligent modeling method in conjunction with NIR spectral technology to support the distributed sensing of heavy metal contamination of soils over a range of geographically different areas. In order to specify the distributed spectral sensing data for the estimation of soil heavy metal, we used a set of portable *in situ* sensors, which allows for the detection conditions more representative of the real-world field conditions than laboratory settings. Based on the IoT framework, federal analysis of the big spectral data collected from different sensors enabled us to identify the informative features by examining the relations between the NIR data and the targeted heavy metal concentrations. For chemometric studies, we investigate the improvement of evolutionary learning methods, taking FA as an example, to validate the practicability of the selection of spectral features, thereby bolstering a novel study branch of NIR technology for model optimization by the population iterative evolution. Considering the integrated impact of the complexity of soil properties and the diversity of distributed sensing, the extraction of spectral features from the entire variable set can be challenging. Nonetheless, we developed an improved FA method through binary discretization and further combined it with the information correlation method for fusion modeling. The proposed methodology aims to identify the key NIR spectral information for a good prediction of heavy metal concentrations in soil. Implementing intelligent modeling strategies is expected to

effectively monitor heavy metal contamination in the soil environment and provide valuable technical support to assist the efforts in soil management.

2 Materials and methods

2.1 Distributed sensing and data preparation

2.1.1 Site selection and equipment settings

The Guangxi Zhuang Autonomous Region (Guangxi ZAR) is a mountainous and water-rich provincial administrative region in south China (see Figure 1a), which has a mild climate and fertile soil resources and produces a variety of crops, vegetables, and fruits. However, on the north side of Guangxi ZAR, the mountains are mostly Karst geology (Li et al., 2023). The topography of Karst is characterized by ruggedness of the surface, variety of landforms, richness of bio-ecology, and poor nutrition in soils (Zhang Z. et al., 2024). To protect the fertile soils in the Karst circumstance, a total of 14 regular farmland areas were selected on the north side of the Guangxi ZAR for the study of distributed sensing detection of heavy metal concentrations. The 14 areas were distributed at the geographical locations identified in Figure 1b. To balance distributed sensing and precise analysis, we used 12 spectral sensors in each location, which were set at an average squared distance for direct contact with the topsoil surface. The 12 × 14 sensors are all well connected to a cloud platform, from which the sensing operation can be remotely controlled.

The spectral sensing operation was conducted in late summer when the effects of moisture, roughness, and noise interference were minimized. Before spectral sensing, we visited the fields several times and kept in contact with the landowners and managers to acquire basic knowledge about the environmental conditions of the field. No rain fell on the day when we performed our sensing work, nor was any rain recorded 2 days before. Technically, we learned that the drier soil surfaces can be caused by the season's climate factors. The properties of the upper mask of topsoil are mostly affected by boundary reasons other than the soil components. Thus, we removed approximately 1 inch of the upper mask and then collected the soil spectra using sensing devices. The sensors received the FT-NIR spectral data on heavy metal concentrations, as well as the other component properties.

2.1.2 The spectral data and the chemical records

The *in situ* FT-NIR spectra were measured in the fields using a micro FC-025-2 TE spectrometer (ARCOptix Inc., Neuchatel, Switzerland) across the band of 10,000–4,000 cm^{-1} with 4 cm^{-1} resolution. A total of 1,512 discrete variables were recorded for each sensing location. Over the 168 distributed spots, we collected the spectrometer detection results as the spectral sensing responses of 168 topsoil samples. We collected five repeated spectral measurements for each sample and computed the average data from the available spectral data.

Additionally, during the *in situ* measurement of the spectral, soil samples were collected from 1 inch below the surface, to a depth of 4 inches from each position, and delivered to the laboratory for the determination of the heavy metals of Zn, chromium (Cr), copper (Cu), and Pb. In the laboratory, the samples were oven-dried at

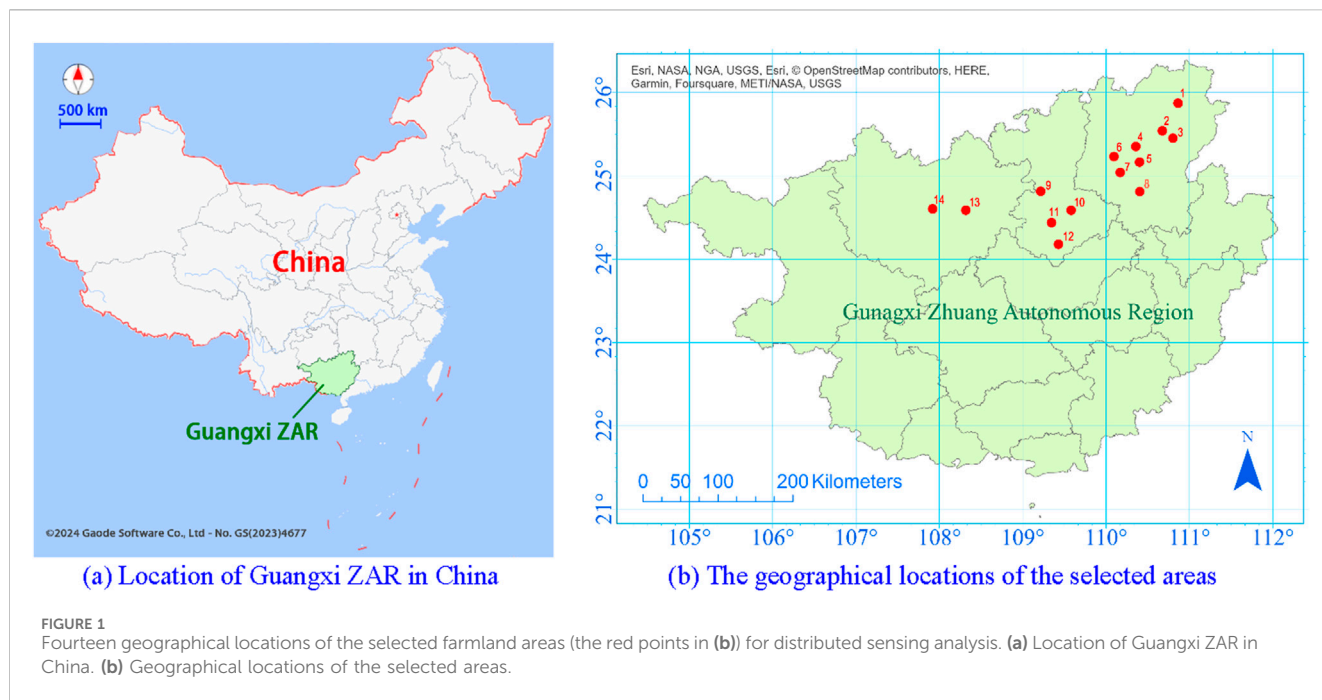


TABLE 1 Descriptive statistics of the four targeted heavy metal concentrations (mg/kg) of the sensing data.

Content	Max	Min	Mean	Median	SD	CV	eKT	Skewness
Zn	594.13	32.33	210.29	124.43	160.19	0.59	0.88	1.25
Cr	148.93	63.69	112.05	16.10	112.61	0.14	0.41	-0.45
Cu	86.79	29.41	60.46	13.30	60.09	0.22	-0.46	0.09
Pb	104.38	14.31	58.29	22.23	51.81	0.38	-0.57	0.62

45 °C, gently crushed, and sieved to a fine earth fraction no larger than 2 mm in diameter. The contents of Zn, Cr, Cu, and Pb were measured using the inductively coupled plasma atomic emission spectroscopy (ICP-AES) method (Daftsis and Zachariadis, 2008). Table 1 shows the descriptive statistics of the targeted heavy metal concentrations for the distributed sensing samples, consisting of the maximum, the minimum, the mean value, the median, standard deviation (SD), coefficient of variation (CV), excess kurtosis (eKT), and skewness. These values are the aggregated amounts of the metals in the distributed samples.

2.2 Methodologies for evolutionary optimization

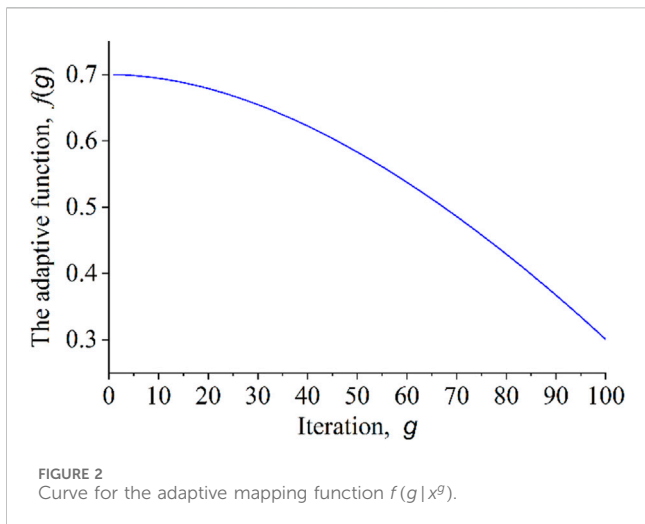
The distribution of information variables is discretized, so that the search for feature variables is in vital demand. In this case, evolutionary learning methods have the advantage of iteratively refining their solutions and are well-suited to assist in an intelligent search of the variables. Furthermore, the fusion or combination of these selected features can probably enhance the NIR model performance by recurrent evolution of the information about the analytes from the diverse distributed sensing data.

2.2.1 Improvement of the FA method

The FA is a classical evolutionary learning method. It was originally developed by simulating the biological characteristics of firefly luminescence. In the swarm gathering principle, the attraction between fireflies decreases as their distance increases. The fireflies spontaneously fly toward the brightest individual to search for a superior position (Zeng et al., 2024). A “0–1” mapping rule can simplify the FA method as a binary processing scheme (Ervural and Hakli, 2023). The binary FA (BFA) method is used to choose some discrete spectral variables at random, obeying the 0–1 encoding rule, to generate a set of discrete NIR feature variables that correspond to candidate responses to the several target metal concentrations.

In the FA story, suppose that there is a total of NP individuals in the firefly population. The population evolves no more than G times. With statistical explanation, we denote the i th individual at the g -th iteration as z_i^g , which is discretized as a vector in p length, that is, $z_i^g = (z_{i1}, z_{i2}, \dots, z_{ip})^g$. For improvement, BFA generates a 0–1 valued regulation index $u_i^g = (u_{i1}, u_{i2}, \dots, u_{ip})^g$ with the same length as z_i^g . The BFA presents the crossover calculation as the dot product of $z_i^g \cdot u_i^g$. By randomization, the individual is decomposed as in binary coding, to produce a new individual for mutation and selection.

For the g th iteration, the candidate wavenumbers that are identified by BFA are used as feature variables to establish



models for the spectral sensing data. The fitness function originally defined by the brightness of firefly individuals is currently re-defined numerically as root mean square error (RMSE), directly linking to the model prediction effect.

To smooth the iteration, a threshold value is set in the fitness function to formulate the evolutionary individual, namely,

$$cz_i^{g+1} = z_i^g + k_i \cdot \Delta z_i^g + l_i \cdot (\theta_i - 0.5), \text{ for } i = 1, 2 \dots NP, \quad (1)$$

where cz_i^{g+1} is the renewed candidate evolutionary individual after the g -th iteration; Δz_i^g represents the stochastic change by iteration; θ_i is the threshold, and k_i and l_i are the tunable regularization parameters defined in the same dimension to z_i^g .

Next, the renewed candidate individual cz_i^{g+1} crosses over with u_i^{g+1} and is further mapped by the fitness function to identify the evolutionary individual for the selection of informative variables. The evolution of the population is determined as follows:

$$z_i^{g+1} = \begin{cases} z_i^g, & f(g|x^g) < f(g|x^{g+1}), \\ cz_i^{g+1}, & f(g|x^g) \geq f(g|x^{g+1}), \end{cases} \quad (2)$$

where $x^g = z_i^g \cdot u_i^g$ and $x^{g+1} = cz_i^{g+1} \cdot u_i^{g+1}$ for $i = 1, 2 \dots NP$.

The primary BFA method is improved by applying the specially designed fitness function $f(\cdot)$ to launch the adaptive optimization for variable selection from the renewed candidate sets of $\{cz_i^g|g, i = 1, 2 \dots NP\}$. The iterative evolution terminates when it reaches the preset maxima and obtains the combination of the selected data. As the variables evolve through iterative learning, we develop a simple and concise model that is optimized with the endorsement of the fusion of these selected variables, so that the model can be improved to have high prediction accuracy. For convenience, the improved BFA method is hereafter denoted as IBFA.

In IBFA, the 0–1 binary coding indicates that the variable is selected only if it is encoded as 1. Thus, the mapping function is important for variable selection. To speed up the convergence of iterations, a special function is proposed for adaptive mapping, which endorses high probabilities of individuals being encoded as 1 in the early iteration stage, and then the probabilities decrease as the iteration continues. Theoretically, this kind of mapping function

should be concave and monotonically decreasing. We define our mapping function as follows:

$$f(g|x^g) = \mu - (1 - \mu) \cdot x^g \cdot \frac{g}{G} \cdot \log\left(1 + \frac{g}{G}\right), \quad (3)$$

where x^g represents the undergoing participants of variables at the g th iteration; G is the preset total number of rounds of iteration; μ is a parameter that denotes the probability of individual variables being encoded as 1. It controls the 0–1 binary division ratio, where $\mu \times 100\%$ of the variables are encoded as 1, and $(1 - \mu) \times 100\%$ as 0. For instance, let $\mu = 0.7$ and $G = 100$, we could have the iterative curve for the adaptive function available for the fixed participation of $x^g \equiv 1.92$ (see Figure 2). The function serves for fast identification of feature variables at a faster iterative convergence.

2.2.2 Information correlation for distributed sensing data

Evolutionary learning methods are advantageous for selecting spectral features. However, evolution optimization methods cannot handle the correlation and collinearity of variables. Confronting the situation of distributed sensing, the spectral data from different locations are inherently related to soil metal contamination, so they are typically correlated with each other. It is necessary to conduct a preliminary correlation analysis on the IoT-based cloud platform for all collected samples.

According to the NIR spectral property, more information is included in the wavelengths with higher correlations to the targeted metal concentration. Correlation analysis is a statistical measure that can tolerate data uncertainties (Mana and Pizzocaro, 2021). To effectively analyze the distributed data, we attempt to set a threshold for the correlation coefficient. This allows us to expand the candidate spectral range in accordance with the known chemical knowledge of the heavy metals. Subsequently, we can select the wavelength variables with higher correlation coefficients that exceed this threshold across the entire sensing dataset. These variables are prospectively available for enhancing models in an IoT environment where data are distributed.

The maximum information coefficient (MIC) is a special index that tells reliable correlations in terms of nonlinear correlation and uneven sample distribution (Luo et al., 2024). In theory, the original MIC algorithm is only concerned with selecting highly correlated variables but does not consider the influence of collinearity among the distributed data. To address this issue, we include a projection subtraction method to reduce the collinear effects on correlation analysis, so that the MIC algorithm is modified. The modified MIC algorithm (denoted as MMIC) is designed specifically for spectral sensing, with the aim of capturing a wide range of functional, interesting associations to provide a scoring indicator similar to the coefficient of determination in a regression model.

In detailed designs, we first make a fast computation of the MIC values between the raw wavenumber variables and the orthogonal variables. Suppose the raw dataset is denoted as $x^{(r)}$ and the orthogonal variable dataset as $x^{(o)}$. A scatter plot is generated based on the two related variable sets, and the plotted map is divided into $n \times m$ small grids. The mutual information is calculated between every two grids over all of the possible one-to-one pair grid combinations, formulated as Equations 4, 5:

$$H_{\max}(x^{(r)}, x^{(o)}, grid) = \max_{j=1,2,\dots,n \times m} \{H(x_j^{(r)}, x_j^{(o)})\}, \quad (4)$$

$$MIC = \frac{H_{\max}(x^{(r)}, x^{(o)}, grid)}{\log\left(\min_j(x_j^{(r)}, x_j^{(o)})\right)}, \quad (5)$$

where $H(\cdot, \cdot)$ represents a function that monitors the frequency of the counted data points appearing in the j -th targeted grid for $j = 1, 2, \dots, n \times m$.

Specifically, a forward-loop encoding procedure is designed to complete the cycling selection of projections to generate the dataset of $x^{(o)}$. For the g th round (going through the total of G iterations), the projection is defined as Equation 6:

$$x_j^{(o)} = x_j^{(r)} - \left((x_j^{(r)})^T x_0^{(r)} \right) \cdot x_j^{(r)} \cdot \left((x_j^{(r)})^T x_j^{(r)} \right)^{-1}, \quad (6)$$

where $x_0^{(r)}$ is the raw variable that has the largest MIC value, and $x_j^{(r)}$ represents the variable in the j th grid. To widen the possible selection range of candidate feature variables, we set a threshold θ to expand the range for candidate features. By the modification of dual factors of projection subtraction and threshold control, the MMIC value is refined as Equation 7:

$$MMIC = \max_{B(nm)} \{MIC\} \cdot (1 + \theta), \quad (7)$$

where $B(nm)$ is the upper bound of the available gridded area, which is usually valued as $(n \times m)^{3/5}$, and θ represents the threshold that is functional to expand the spectral range for IBFA iterative search.

In this way, the MMIC algorithm step-by-step performs correlation judgment by a series of dynamic MMIC values that are calculated by the embodiment of projection subtraction during the iterative data refining for the distributed sensing acquisition. Technically, we fuse the MMIC algorithm with the IBFA evolutionary method to identify the informative wavenumbers/variables.

3 Results and discussion

For model establishment, training, and optimization, the 168 soil samples detected at different spot locations are partitioned into two functional sets, one for modeling and the other for testing. We used a usual partitioning ratio of 3:1 to determine the number of samples for the modeling set and for the testing set. The WSPXY method (Tian et al., 2019) is employed to make the division, so that the selected modeling samples are representative. Accordingly, the determination coefficient (R) and the root mean square error (RMSE) for modeling are denoted as R_M and $RMSE_M$ and as R_T and $RMSE_T$ for testing.

3.1 Data pre-processing and model evaluation mode

The raw spectral sensing data are accompanied by noisy interference. To reduce the influence of multiple kinds of noise, the data were subjected to various pre-processing methods using the MATLAB toolbox. The pre-processing algorithms included

TABLE 2 Comparison of eight different solo and combined pre-processing algorithms based on 5-fold cross-validation for the spectral sensing data using classical PLS regression.

Method	R_{CV}				$RMSE_{CV}$			
	Zn	Cr	Cu	Pb	Zn	Cr	Cu	Pb
Raw	0.688	0.671	0.691	0.713	41.77	25.20	12.97	11.91
SG	0.806	0.730	0.754	0.779	36.77	23.07	11.36	10.48
DWT	0.824	0.728	0.752	0.777	34.61	22.18	10.60	10.43
MSC	0.777	0.710	0.744	0.756	38.24	22.51	11.58	10.90
SNV	0.796	0.721	0.732	0.769	37.31	22.08	11.87	10.64
SG + MSC	0.779	0.758	0.759	0.784	36.60	21.88	10.74	9.73
SG + SNV	0.799	0.735	0.790	0.812	36.26	20.59	11.26	9.49
DWT + MSC	0.849	0.768	0.783	0.817	34.13	20.48	11.41	10.34
DWT + SNV	0.834	0.765	0.804	0.820	33.28	20.88	10.33	9.87

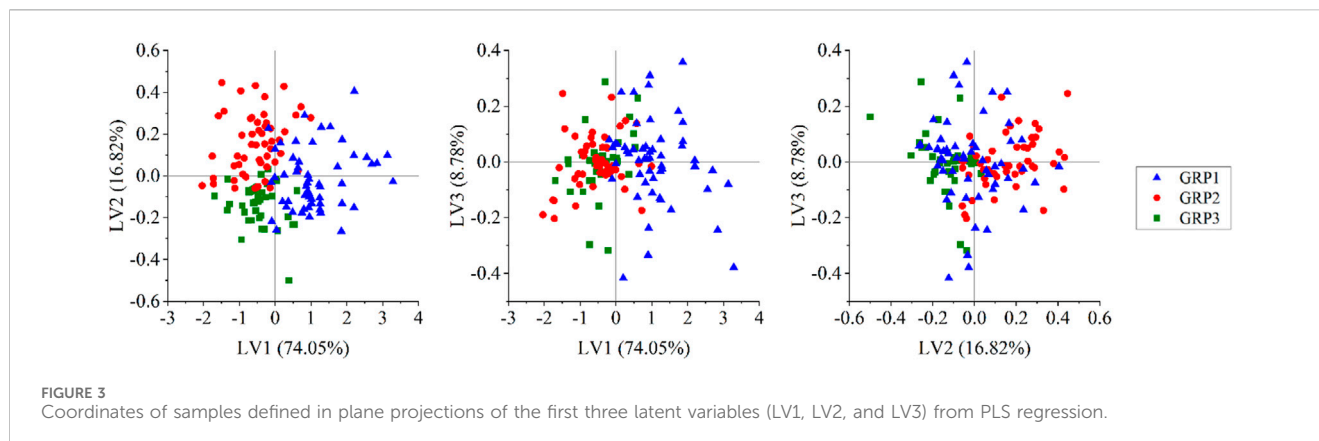
Savitzky–Golay (SG), discrete wavelet transformation (DWT), multiplicative scatter correction (MSC), standard normal variate (SNV), and their pair combinations.

To ensure the pre-processed models are not affected by multivariate aggregation, we employed the classical PLS method to perform pre-modeling training, with the computing mode by 5-fold cross-validation over the 168 soil samples. The cross-validation accuracy was evaluated by the determination coefficient (R_{CV}) and the root mean square error ($RMSE_{CV}$). In the PLS process, the participant variables are the latent variables developed by a linear combination of the initial wavenumbers. The model runs and evaluates itself for the number of latent variables changing from 1 to 20. The optimal number was selected with the release of R_{CV} and $RMSE_{CV}$. To ensure model stability, the procedure was repeated 20 times, and the average results of R_{CV} and $RMSE_{CV}$ are observed for each of the heavy metal targets. After discarding the less significant results, the best pre-processing observations of the data are shown in Table 2 for comparative selection of the best and refined noise reduction method.

Table 2 indicates that the best refined pre-processing methods are different for each heavy metal concentration. Aiming to identify the minimum $RMSE_{CV}$ with a corresponding appropriate high value of R_{CV} , we learned that the pair-combination methods have better pre-processing results for the distributed sensing soil samples. The best pre-processing method is DWT + SNV, which released higher R_{CV} values and lower $RMSE_{CV}$ than the other pre-processing methods for Zn, Cu, and Pb. Although the DWT + SNV method did not observe the best values for Cr (the minimum $RMSE_{CV}$ or the highest R_{CV}), the observed values are very close to the best values.

3.2 Multivariate analysis by classical PLS

The PLS method performs variable decomposition to generate a series of latent variables obeying the principal component analysis (PCA) principle (Cook, 2022). The latent variables in the front of the queue have major data information when sorted in descending order according to the variance contribution rate. Then, we can select a



few latent variables for tuning the regression model and observing the optimal RMSE values for prediction. The first three latent variables (LV1, LV2, and LV3) account for most of the total variance of the dataset. They explained more than 99% of the total variance in spectral behaviors of the four targeted heavy metal concentrations. The contributions of LV1, LV2, and LV3 are 74.05%, 16.82%, and 8.78%, respectively. In contrast, the remaining latent variables (LVs) contribute much less (<1% in all). Thus, we illustrate the LV biplot graph based on the LV1, LV2, and LV3 values (shown in Figure 3). In Figure 3, the samples are divided into three groups (denoted as GRP1, GRP2, and GRP3). The first group is totally consistent with the testing set. The remaining two sets are evenly divided from the modeling set so that they have similar numbers as the first group. The figures show that the distributed samples could be aggregated for uniform modeling according to their spectral properties with the designated groups. Indeed, with more than 90% variance contribution, the LV1–LV2 projection scatter plot showed three main clusters of the designated sample partitions with obvious differences in the loadings, which indicates that the FT-NIR spectral signals over the distributed spots are able to quantify metal concentrations. As they inherit most of the spectral information, LV1 and LV2 are selected as the most important feature variables that represent the soil metal properties. Some other latent variables are also included for model refinement.

Aiming at the four targeted heavy metals of Zn, Cr, Cu, and Pb, we established and trained the calibration models based on the key latent variables, including LV1 and LV2. We identified the optimal calibration models that demonstrated commendable prediction capabilities both for modeling and for testing. Some extracted latent variables other than LV1 and LV2 are tested. Table 3 shows the prediction results from the optimal model endorsed with their supporting latent variables. All the selected latent variables originate from the raw wavenumber variables based on the full range of 10,000 cm^{-1} to 4,000 cm^{-1} . With interactive comparison among the targeted elements, we can find that the LV-based optimized models effectively predict Zn and Cu concentrations. Their models exhibit high correlation coefficients exceeding 0.88 in the modeling process, with the RMSE_M lower than 13% of their respective mean values. In contrast, the models for Cr and Pb are less accurate, but they still reach the acceptable performance levels. The testing results are generally inferior to

the modeling outputs, providing practical confirmation of the data-driven machine learning mechanism. In addition, we noted that the selected latent variables for each of the targets fall within the first 10 principal components, which indicates that the PLS models established for the FT-NIR prediction of soil heavy metals can be optimized using informative principal variables.

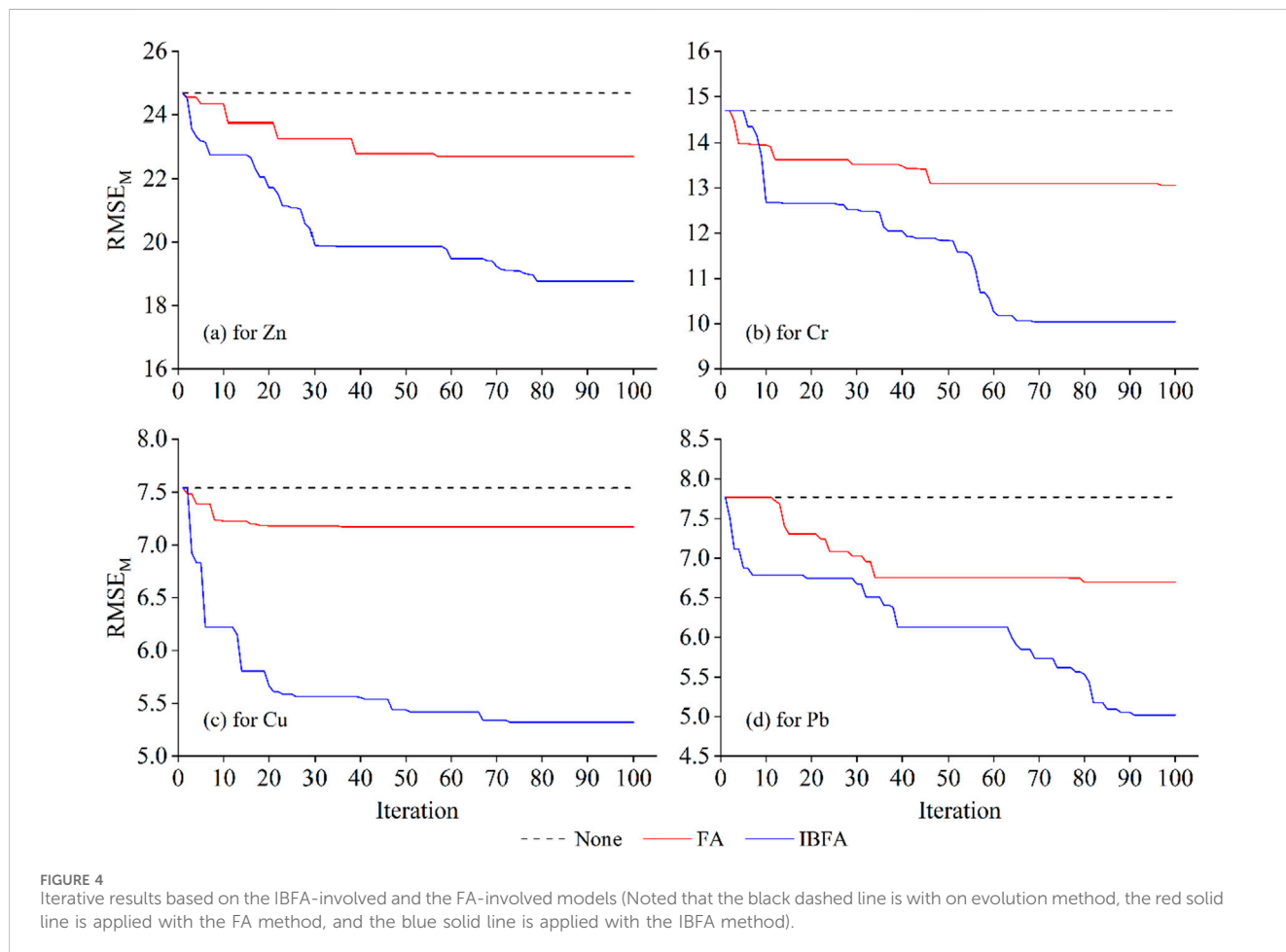
3.3 IBFA evolutionary learning model

With binary transform and by-threshold improvement strategies, the IBFA evolutionary method was validated for its efficacy in selecting informative variables. In practice, we simultaneously set the inherent evolutionary parameters for both IBFA and FA algorithms. Based on the population that has a total of 60 individuals, the generation of renewed candidate individuals was regulated by the fitness function (see Formula 1), where we set the regularization parameters k_i and l_i as changing from 0 to 1, and the threshold was set to an adjustment of $\theta_i \in [0.97, 1]$. The iteration process for evolution was regulated with the upper limitation of $G = 100$. In the selection step (see Formula 2), the binary factor u_i was randomly generated as 0–1 arrays for each round of iteration but always maintaining the same dimension consistent with z_i . With the control of the fitness defined by the function of $f(g)$ (see Formula 3), the variables selected by IBFA were regarded as informative and further delivered to re-establish the PLS models.

Figure 4 depicts the iterative optimization processes of IBFA and FA, showcasing the predictive RMSE_M derived from PLS modeling based on the chosen informative variables. For convenient comparison, the PLS prediction with none of the evolutionary optimization is also shown in the figure. The comparative experiments have substantiated that the involvement of IBFA and FA enhances the efficiency of wavenumber selection. As a result, the swarm-evolved PLS models outperformed the counterpart conventional PLS model. The lowest RMSE_M is obtained in the IBFA intervened model after ~ 70 evolutions by iteration. This revealed that the adaptive mapping function defined by Formula 3 appreciably improved the firefly evolutionary effect in combination with the binary discretization strategy. The gradient convergence of the IBFA model appears slightly later than that of FA, which corroborates the fact that local optimization is postponed when using a specially designed fitness function instead of the

TABLE 3 Quantification results by the classical PLS model accompanied by LV selections on multivariate features.

Content	Selected LVs	Modeling			Testing		
		R_M	$RMSE_M$	Relative percentage	R_T	$RMSE_T$	Relative percentage
Zn	LV1, LV2, LV3, LV5, and LV8	0.881	24.68	11.7%	0.878	29.92	14.2%
Cr	LV1, LV2, LV4, and LV5	0.869	14.69	13.1%	0.859	17.13	15.3%
Cu	LV1, LV2, LV3, LV4, and LV6	0.886	7.54	12.5%	0.867	8.36	13.8%
Pb	LV1, LV2, LV5, LV7, LV8, and LV10	0.879	7.77	13.3%	0.862	8.78	15.1%



random mapping. With the best iterative optimization by IBFA, the PLS models were improved to have the lowest $RMSE_M$ values of 18.768 mg/kg, 10.043 mg/kg, 5.322 mg/kg, and 5.019 mg/kg for Zn, Cr, Cu, and Pb, respectively.

The IBFA iteration runs with 0–1 binary crossover and selection over the raw variables of the FT-NIR wavenumbers. We identified the feature variables out of the total 1,512 raw wavenumbers for the FT-NIR data modeling on the Zn, Cr, Cu, and Pb elements (see Figure 5). Figure 5 shows that the concatenation of the four sets of feature variables for Zn, Cr, Cu, and Pb merged in a total of 493 located wavenumbers. The placements of the 493 wavenumbers are uniformly distributed in the full range of 10,000–4,000 cm^{-1} . These feature

variables are suitable for optimizing the PLS model to stable prediction results. This validates that the IBFA variable evolution technique is feasible to enhance the PLS model that works with the selected pre-processing method of DWT + SNV. Both the distribution of latent variables and the IBFA iterative function are fused to perform variable computation on the spectral detection of soil heavy metal contents. This is an indication that evolution techniques could be largely attributed to the finding and identification of heavy metal contamination at the rough topsoil surface, even though that soil is in “Karst fertile” status.

A recent study shows that a distributed sensing of soil surface reduces the FT-NIR reflectance, which may be a drawback of the

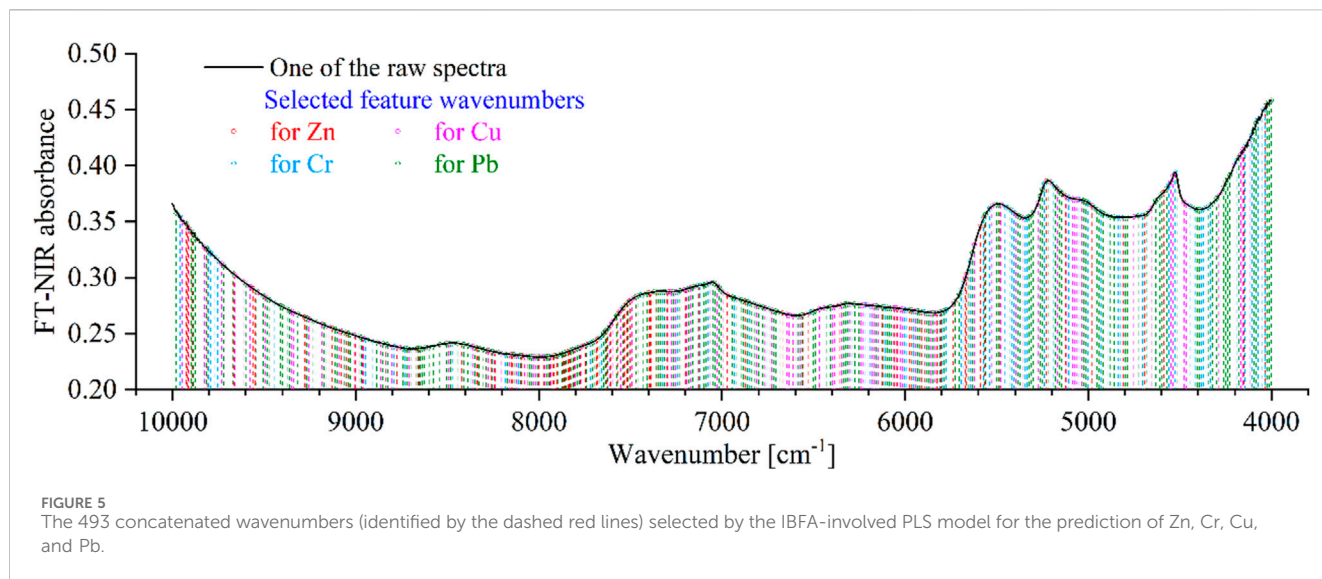


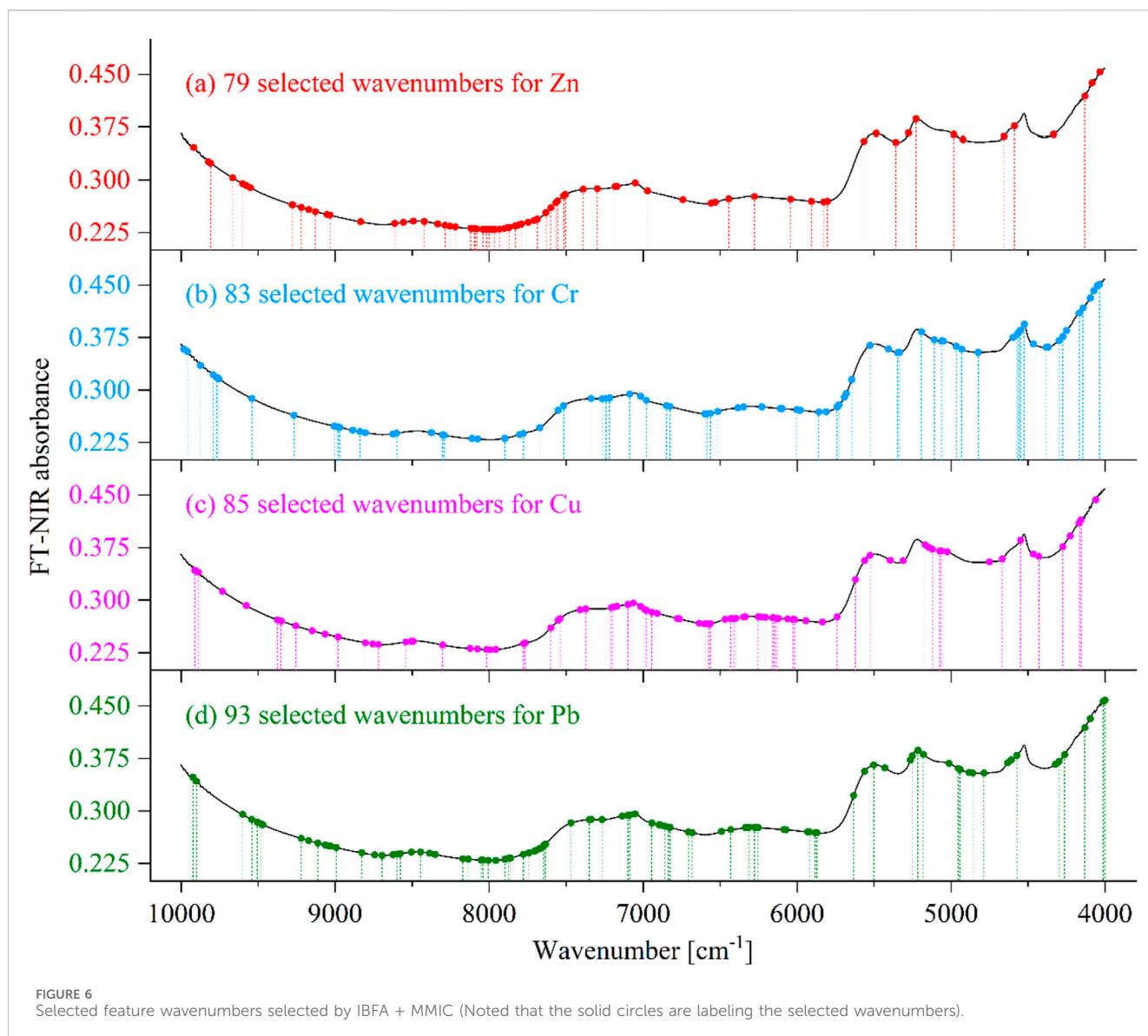
TABLE 4 Prediction results of the adaptive machine learning models by a combination of MIC and FA methods and their improvements.

Method	Content	Modeling		Number of variables	t value ^(#)	Testing	
		R _M	RMSE _M			R _T	RMSE _T
FA + MIC	Zn	0.915	23.072	230	2.49	0.874	27.324
	Cr	0.902	13.124	268	2.58	0.872	15.650
	Cu	0.890	6.560	286	2.45	0.863	8.359
	Pb	0.856	6.786	279	2.52	0.838	7.729
FA + MMIC	Zn	0.922	18.396	153	2.52	0.890	24.631
	Cr	0.916	9.713	148	2.57	0.884	13.269
	Cu	0.906	5.145	150	2.48	0.875	7.033
	Pb	0.864	5.665	158	2.60	0.853	6.668
IBFA + MIC	Zn	0.928	16.802	122	2.48	0.898	20.691
	Cr	0.921	8.668	128	2.52	0.887	10.980
	Cu	0.918	4.611	131	2.53	0.882	6.417
	Pb	0.898	4.778	143	2.58	0.868	5.529
IBFA + MMIC	Zn	0.938	15.130	79	2.74	0.899	18.161
	Cr	0.922	6.993	83	2.63	0.892	9.545
	Cu	0.927	4.262	85	2.65	0.895	4.759
	Pb	0.914	4.633	93	2.84	0.876	4.824

^(#)The t value calculated during the t-test for significance analysis.

contribution of the loss of reflectance of metal elements (Lovynska et al., 2024). Moreover, simultaneous data analysis on a cloud platform may lead to uneven and heterogeneous effects in multivariate computation, thereby affecting the *in situ* spectral sensing output and making the evolutionary module less effective (Russell-Pavier et al., 2023). We can easily observe that the IBFA iteration has enhanced the PLS prediction results to an optimized level, and they remain stable to the end of the 100 rounds of

evolutionary iteration. However, intensive training of the IBFA can lead the model into a local optimization trap, which means that the refined models with 493 selected feature variables are conditionally good and occasionally stable. The models have potential for further enhancement by studying the mutual correlation of information between the endorsed variables. Thus, the proposed MMIC method is employed to improve model stability and to seek a higher R value or a lower RMSE.



3.4 MMIC improvement for distributed data

To improve model precision, we combine the MMIC method with the IBFA-optimized PLS model for global refinement of feature selection. As the original MIC method accounts for data collinearity, the MMIC method fuses the projection subtraction in variable space and scores the association relationship between variables. In practice, the MMIC algorithm is applied after the IBFA, with the aim of capturing the features from the full range of available IBFA-produced variable space. The MMIC scoring indicator is dynamically evaluated by iterative steps to ensure the selected variables make an interesting contribution to model improvement.

In a deep analysis of the distributed spectral sensing data in a specified area, the MMIC-combined IBFA model was targeted to balance the prediction of the four metal elements of Zn, Cr, Cu, and Pb in the global view; the results are shown in Table 4. To examine the practical advantages of the IBFA-combined-MMIC (IBFA + MMIC) model, the original FA method and the MIC algorithm are alternatively applied for comparison. Table 4 shows that the IBFA +

MMIC model produced fairly good prediction results, with the R_M values uniformly higher than 0.9 during the model training process. The other models (FA + MIC, FA + MMIC, IBFA + MIC) performed less well than the IBFA + MMIC model. The statistical significance of the presented models was validated by using a t-test. Aiming at 95% confidence level, the t values of these models were calculated and are listed in Table 4. In practice, the critical values of the t-test for each optimal model situation are different as the degrees of freedom vary, but all critical values are less than 2. The calculated t values of all optimal models for the prediction of Zn, Cr, Cu, and Pb were obviously larger than the critical t-test values, which validated that the selected feature variables are significant for model optimization. Specifically, the IBFA + MMIC model has the largest calculated t values (2.74 for Zn, 2.63 for Cr, 2.65 for Cu, and 2.84 for Pb) of the models, which indicates that the proposed fusion model of IBFA + MMIC performs best for both the validation samples and the testing samples. With this optimal modeling result, we confirmed the improvement of model prediction accuracy for estimating Zn, Cr, Cu, and Pb in soil.

In other aspects, it is found that the IBFA + MMIC eventually selects less informative variables than the other counterpart methods for model optimization. The applied feature variables are identified in the full spectral detection region (See Figure 6). Successively, when using the selected feature variables for model testing, the model revealed appreciable testing results in terms of correlations (R_T) and errors ($RMSE_T$). The optimized model results are better than those of the previous IBFA-PLS models without MMIC improvement (with higher R_M and lower $RMSE_M$ values). These results indicate that the proposed IBFA + MMIC method is advantageous for the analysis of distributed spectral sensing data. Considering the differences in circumstances, the distributed data vary in dependence; each module must effectively test different emerging data from different located spots and each of the metals investigated to select the most suited model parameters for evolutionary iterative estimation of metal concentrations.

4 Conclusion

This study proposed an intelligent modeling approach for the distributed sensing of soil heavy metal contamination across diverse geographical regions. Large-scale NIR spectral data from IoT-distributed sensors underwent federated analysis. For chemometric innovation, the IBFA and MMIC methods were integrated for iterative optimization of the calibration model, targeting the protection of fertile Karst soils in Guangxi ZAR, China. The IBFA + MMIC fusion model outperformed other preset models in quantifying four heavy metals (Zn, Cr, Cu, and Pb). IBFA extracted informative variables from raw full-range spectral wavenumbers for each metal (see Figure 6), and MMIC improved the fusion of distributed samples. During training, the IBFA + MMIC model achieved $R_M > 0.9$ and low $RMSE_M$; in testing, it maintained R_T close to 0.9 and exhibited lower $RMSE_T$ than the FA + MIC, FA + MMIC, and IBFA + MIC models (see Table 4).

In conclusion, soil heavy metal contamination is a complex challenge that demands a comprehensive, multifaceted mitigation strategy. This study explored the correlations between NIR spectral data and target heavy metal concentrations, successfully identifying informative spectral features. In chemometric research, we investigated improvements to evolutionary learning methods, validating the practicality of spectral feature selection and establishing a novel research paradigm for optimizing NIR technology-driven models via iterative evolution. Specifically, the proposed IBFA + MMIC fusion model is engineered to extract critical NIR spectral information, facilitating accurate quantitative prediction of diverse heavy metal contents in soil and providing valuable technical support for soil pollution monitoring. However, IBFA is somewhat sensitive to parameter tuning during optimization, with parameter initialization influencing iterative results. Moreover, while MMIC captures multi-variable correlations, it cannot identify synergistic interactions in variable combinations—limitations that restrict the generalized application of the IBFA + MMIC fusion algorithm. Future work will involve developing a more stable IBFA optimization approach via parameter uncertainty analysis and repeated sample modeling, as well as exploring novel projection methods to uncover variable interactions. This will further improve the fusion model's

predictive performance and generalization, enabling NIR spectroscopic technology (aided by intelligent algorithm advancements) to effectively monitor soil heavy metal contamination and support soil management initiatives.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material; further inquiries can be directed to the corresponding author.

Author contributions

SH: Funding acquisition, Writing – original draft, Data curation, Validation, Writing – review and editing, Methodology. SZ: Writing – original draft, Methodology, Software, Visualization, Formal Analysis. ML: Visualization, Investigation, Data curation, Writing – original draft. FM: Formal analysis, Writing – review and editing, Investigation. CH: Software, Writing – review and editing, Validation. HC: Conceptualization, Resources, Project administration, Writing – review and editing, Funding acquisition, Supervision.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was financially supported by the National Natural Science Foundation of China (62365008) and Special Project in Key Areas of Ordinary Universities and Colleges in Guangdong Province (2023ZDZX4069).

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or

claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Acharya, A., and Kogure, T. (2023). Application of novel distributed fibre-optic sensing for slope deformation monitoring: a comprehensive review. *Int. J. Environ. Sci. Technol.* 20, 8217–8240. doi:10.1007/s13762-022-04697-5
- Al Maliki, A., Owens, G., Hussain, H. M., Al-Dahaan, S., and Al-Ansari, N. (2018). Chemometric methods to predict of Pb in urban soil from Port Pirie, South Australia, using spectrally active of soil carbon. *Commun. Soil Sci. Plant Anal.* 49, 1370–1383. doi:10.1080/00103624.2018.1464178
- Ali, J. K., Ghaleb, H., Arangadi, A. F., Le, T. P. P., Moraetis, D., Pavlopoulos, K., et al. (2023). Comprehensive assessment of the capacity of sand and sandstone from aquifer vadose zone for the removal of heavy metals and dissolved organics. *Environ. Technol. Innov.* 29, 102993. doi:10.1016/j.eti.2022.102993
- Babbar, H., Rani, S., Soni, M., Keshta, I., Prasad, K. D. V., and Shabaz, M. (2025). Integrating remote sensing and geospatial AI-enhanced ISAC models for advanced localization and environmental monitoring. *Environ. Earth Sci.* 84, 118. doi:10.1007/s12665-025-12121-7
- Bourechak, A., Zedadra, O., Kouahla, M. N., Guerrieri, A., Seridi, H., and Fortino, G. (2023). At the confluence of artificial intelligence and edge computing in IoT-Based applications: a review and new perspectives. *Sensors* 23, 1639. doi:10.3390/s23031639
- Chamara, N., Bai, G., and Ge, Y. (2023). AICropCAM: deploying classification, segmentation, detection, and counting deep-learning models for crop monitoring on the edge. *Comput. Electron. Agric.* 215, 108420. doi:10.1016/j.compag.2023.108420
- Chen, H., Liu, X., Chen, A., Cai, K., and Lin, B. (2020). Parametric-scaling optimization of pretreatment methods for the determination of trace/quasi-trace elements based on near infrared spectroscopy. *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* 229, 117959. doi:10.1016/j.saa.2019.117959
- Chen, H., Li, L., Xie, J., Meng, F., Xu, L., and Xie, H. (2025). A fuzzy-refined neural network architecture for rapid assessment of seawater pollution by hyperspectral imaging. *J. Clean. Prod.* 512, 145704. doi:10.1016/j.jclepro.2025.145704
- Chhabra, A., Singh, G., and Kahlon, K. S. (2021). Performance-aware energy-efficient parallel job scheduling in HPC grid using nature-inspired hybrid meta-heuristics. *J. Ambient. Intell. Humaniz. Comput.* 12, 1801–1835. doi:10.1007/s12652-020-02255-w
- Cook, R. D. (2022). A slice of multivariate dimension reduction. *J. Multivar. Anal.* 188, 104812. doi:10.1016/j.jmva.2021.104812
- Daftis, E. J., and Zachariadis, G. A. (2008). Analytical performance of a multi-element ICP-AES method for Cd, Co, Cr, Cu, Mn, Ni and Pb determination in blood fraction samples. *Microchim. Acta* 160, 405–411. doi:10.1007/s00604-007-0791-2
- Dattatreya, S., Khan, A. N., Jena, K., and Chatterjee, G. (2024). Conventional to modern methods of soil NPK sensing: a review. *IEEE Sens. J.* 24, 2367–2380. doi:10.1109/JSEN.2023.3334243
- De Souza, A. M., Filgueiras, P. R., Coelho, M. R., Fontana, A., Winkler, T. C. B., Valderrama, P., et al. (2016). Validation of the near infrared spectroscopy method for determining soil organic carbon by employing a proficiency assay for fertility laboratories. *J. Near Infrared Spectrosc.* 24, 293–303. doi:10.1255/jnirs.1219
- Devi, K. G., Mishra, R. S., and Madan, A. K. (2022). A dynamic adaptive firefly algorithm for flexible job shop scheduling. *Intell. Autom. Soft Comput.* 31, 429–448. doi:10.32604/iasc.2022.019330
- Ervural, B., and Hakli, H. (2023). A binary reptile search algorithm based on transfer functions with a new stochastic repair method for 0-1 knapsack problems. *Comput. Ind. Eng.* 178, 109080. doi:10.1016/j.cie.2023.109080
- Gouda, M., Abu-hashim, M., Nassrallah, A., Khalil, M. N., Hendawy, E., Benhasher, F. F., et al. (2024). Integration of remote sensing and artificial neural networks for prediction of soil organic carbon in arid zones. *Front. Environ. Sci.* 12 (1–15), 1448601. doi:10.3389/fenvs.2024.1448601
- Han, A., Lu, X., Qing, S., Bao, Y., Bao, Y., Ma, Q., et al. (2021). Rapid determination of low heavy metal concentrations in grassland soils around mining using Vis-NIR spectroscopy: a case Study of Inner Mongolia, China. *Sensors* 21, 3220. doi:10.3390/s21093220
- Latosinska, J., Kowalik, R., and Gawdzik, J. (2021). Risk assessment of soil contamination with heavy metals from municipal sewage sludge. *Appl. Sci.* 11, 548. doi:10.3390/app11020548
- Leenen, M., Welp, G., Gebbers, R., and Paetzold, S. (2019). Rapid determination of lime requirement by mid-infrared spectroscopy: a promising approach for precision agriculture. *J. Plant Nutr. Soil Sci.* 182, 953–963. doi:10.1002/jpln.201800670
- Li, Y. (2021). Rapid quality discrimination of grape seed oil using an extreme machine learning approach with near-infrared (NIR) spectroscopy. *Spectroscopy* 36, 14–20.
- Li, Y., Wang, S., Peng, T., Zhao, G., and Dai, B. (2023). Hydrological characteristics and available water storage of typical karst soil in SW China under different soil-rock structures. *Geoderma* 438, 116633. doi:10.1016/j.geoderma.2023.116633
- Li, Y., Du, W., Wang, X., and Yu, H. (2024). A Novel adaptive Robust NIR modeling method based on sparse Bayesian Learning. *IEEE Trans. Ind. Inf.* 20, 8499–8511. doi:10.1109/TII.2024.3367007
- Liu, Y., Sun, H., Zhao, C., Ai, C., and Bian, X. (2024). Extreme learning machine combined with whale optimization Algorithm for spectral quantitative analysis of complex samples. *J. Chemom.* 38, 3590. doi:10.1002/cem.3590
- Lovynska, V., Bayat, B., Bol, R., Moradi, S., Rahmati, M., Raj, R., et al. (2024). Monitoring heavy metals and metalloids in soils and vegetation by remote sensing: a review. *Remote Sens.* 16, 3221. doi:10.3390/rs16173221
- Luo, J., Chen, Y., Zhu, Z., Wei, C., Sun, L., Zhang, H., et al. (2024). Maximal information coefficient and geodetector coupled quantification model: a new data-driven approach to coalbed methane reservoir potential evaluation. *J. Pet. Explor. Prod. Technol.* 14, 2937–2951. doi:10.1007/s13202-024-01880-x
- Mana, G., and Pizzocaro, M. (2021). The least informative distribution and correlation coefficient of measurement results. *Metrologia* 58, 015012. doi:10.1088/1681-7575/abcbe9
- Mortada, B., Medhat, M., Sabry, Y. M., Sadek, M., Shebl, A., Hassan, K., et al. (2021). Ultra-Compact fourier transform near-infrared MEMS spectral sensor for smart industry and IoT. *IEEE J. Sel. Top. Quantum Electron.* 27, 2–9. doi:10.1109/JSTQE.2021.3091375
- Nawar, S., Mohamed, E. S., Essam-Eldeen Sayed, S., Mohamed, W. S., Rebouh, N. Y., and Hammam, A. A. (2023). Estimation of key potentially toxic elements in arid agricultural soils using Vis-NIR spectroscopy with variable selection and PLSR algorithms. *Front. Environ. Sci.* 11 (1–13), 1222871. doi:10.3389/fenvs.2023.1222871
- Palteva, A. A., Deeb, M., Di Iorio, E., Circelli, L., Cheng, Z., and Colombo, C. (2022). Prediction of bioaccessible lead in urban and suburban soils with Vis-NIR diffuse reflectance spectroscopy. *Sci. Total Environ.* 809, 151107. doi:10.1016/j.scitotenv.2021.151107
- Perkovic, S., Paul, C., Vasic, F., and Helming, K. (2022). Human health and soil health risks from heavy metals, micro(nano)plastics, and antibiotic resistant bacteria in agricultural soils. *Agronomy-Basel* 12, 2945. doi:10.3390/agronomy12122945
- Pourdarbani, R., Sabzi, S., Rohban, M. H., Garcia-Mateos, G., Molina-Martinez, J. M., Paliwal, J., et al. (2022). Metaheuristic algorithms in visible and near infrared spectra to detect excess nitrogen content in tomato plants. *J. Near Infrared Spectrosc.* 30, 197–207. doi:10.1177/09670335221098527
- Russell-Pavier, F. S., Kaluvan, S., Megson-Smith, D., Connor, D. T., Fearn, S. J., Connolly, E. L., et al. (2023). A highly scalable and autonomous spectroscopic radiation mapping system with resilient IoT detector units for dosimetry, safety and security. *J. Radiol. Prot.* 43, 011503. doi:10.1088/1361-6498/acab0b
- Song, H., Lei, B., Guang, P., Guo, C., Zhou, Y., Han, X., et al. (2022). Rapid determination of As and Pb concentrations in soil based binary Grey Wolf Optimization and partial least squares regression. *Eurasian Soil Sci.* 55, 1313–1322. doi:10.1134/S1064229322090071
- Surawijaya, A., Chandra, Z., Sulthoni, M. A., Idris, I., and Adiono, T. (2023). Modeling and Simulation of Si Grating Photodetector Fabricated Using MACE Method for NIR Spectrum. *Electron.* 12, 663. doi:10.3390/electronics12030663
- Tao, R., Zhou, H. L., Meng, Z., and Liu, Z. T. (2024). An integrated firefly algorithm for the optimization of constrained engineering design problems. *Soft Comput.* 28, 3207–3250. doi:10.1007/s00500-023-09305-3
- Tian, H., Zhang, L., Li, M., Wang, Y., Sheng, D., Liu, J., et al. (2019). WSPXY combined with BP-ANN method for hemoglobin determination based on near-infrared spectroscopy. *Infrared Phys. Technol.* 102, 103003. doi:10.1016/j.infrared.2019.103003
- Wang, Y., Zhang, X., Sun, W., Wang, J., Ding, S., and Liu, S. (2022). 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.* 838, 156129. doi:10.1016/j.scitotenv.2022.156129
- Zayani, H., Fouad, Y., Michot, D., Kassouk, Z., Baghdadi, N., Vaudour, E., et al. (2023). Using machine-learning algorithms to predict soil organic carbon content from combined remote sensing imagery and laboratory Vis-NIR spectral datasets. *Remote Sens.* 15, 4264. doi:10.3390/rs15174264
- Zeng, S., Wang, Y., Wen, Y., Yu, X., Li, B., and Wang, Z. (2024). Firefly forest: a swarm iteration-free swarm intelligence clustering algorithm. *J. King Saud. Univ. Inf. Sci.* 36, 102219. doi:10.1016/j.jksuci.2024.102219

Zhang, Y., Li, M., Zheng, L., Qin, Q., and Lee, W. S. (2019). Spectral features extraction for estimation of soil total nitrogen content based on modified ant colony optimization algorithm. *Geoderma* 333, 23–34. doi:10.1016/j.geoderma.2018.07.004

Zhang, Y., Chen, H., Chen, W., Xu, L., Li, C., and Feng, Q. (2021). Near Infrared feature waveband selection for fishmeal quality assessment by frequency adaptive binary differential evolution. *Chemom. Intell. Lab. Syst.* 217, 104393. doi:10.1016/j.chemolab.2021.104393

Zhang, R., Meng, Z., Wang, H., Liu, T., Wang, G., Zheng, L., et al. (2024a). Hyperscale data analysis oriented optimization mechanisms for higher education management systems platforms with evolutionary intelligence. *Appl. Soft Comput.* 155, 111460. doi:10.1016/j.asoc.2024.111460

Zhang, X., Zhang, S., Liu, S., Ren, D., and Zhang, X. (2024b). Study on the migration behaviour of heavy metals at the improved mine soil-plant

rhizosphere interface. *Environ. Technol.* 45, 4691–4703. doi:10.1080/09593330.2023.2283061

Zhang, Z., Zhang, F., Yang, X., and Zhang, J. (2024c). The occurrence and distributions characteristics of microplastics in soils of different land use patterns in Karst Plateau, Southwest China. *Sci. Total Environ.* 906, 167651. doi:10.1016/j.scitotenv.2023.167651

Zhou, X., Lei, X., Yang, C., Shi, Y., Zhang, X., and Shi, J. (2024). Handling data heterogeneity for IoT devices in federated learning: a knowledge fusion approach. *IEEE Internet Things J.* 11, 8090–8104. doi:10.1109/JIOT.2023.3319986

Zukowska, G., Bik-Malodzinska, M., Myszyra, M., Pawlowski, A., and Pawlowska, M. (2021). Effect of sewage sludge and mineral wool on water retention and heavy metal content in medium agronomic category soil. *Environ. Prot. Eng.* 47, 101–110. doi:10.37190/epe210407