



OPEN Estimation of soil free Iron content using spectral reflectance and machine learning algorithms

Wanzhu Ma¹, Hongkui Zhou¹, Hao Hu^{1,2✉}, Zhiqing Zhuo¹, Kangying Zhu¹ & Guangzhi Zhang³

Spectral reflectance technology has emerged as a promising tool for estimating soil properties while offering a rapid, non-destructive, and cost-effective alternative to traditional methods. Free iron is an important soil property, and it reflects the occurrence and evolution of soil. An accurate and efficient determination of soil free iron content is important. To evaluate the feasibility of using spectral reflectance and machine learning methods to estimate soil free iron content, we collected the spectral reflectance of 540 soil samples from 135 locations. We looked at the original spectrum and transforms such as the first derivative (FD), standard normal variate (SNV), and continuum removed (CR). The full spectrum, correlated spectrum, and principal components from principal component analysis (PCA) were considered as model variable selection. We used machine learning algorithms, such as partial least squares (PLS), support vector machine (SVM), random forest (RF), and deep neural network (DNN) algorithms for model construction. We found that FD was a more efficient transform than the original, SNV and CR spectra. The average R^2 , $RMSE$, and $RRMSE$ when using the FD transform for training were 0.797, 5.550 g/kg, and 25.1%, respectively. In testing models, CR had a higher accuracy than the other transforms and its R^2 , $RMSE$, and $RRMSE$ were 0.644, 7.140 g/kg, and 32.7%. Variable selection based on PCA projection improved model accuracy compared to using full and correlated spectra. The average model R^2 , $RMSE$, and $RRMSE$ following PCA were 0.821, 5.260 g/kg, and 23.9% in training and 0.692, 6.744 g/kg, and 30.9% in testing, which had a higher R^2 and lower $RMSE$ and $RRMSE$ than when using the full and correlated spectra. Over-fitting may have occurred in our study when employing the CR transform and RF algorithm. Their models had high accuracy in training and low accuracy in testing. The model R^2 using the DNN showed better performance than those using the PLS and SVM algorithm, but the DNN showed poorer performance in $RMSE$ and $RRMSE$ than that of the model utilizing the SVM and PLS algorithm. The best combination of spectral transform, variable selection, and modeling method was FD + PCA + SVM. The R^2 , $RMSE$ and $RRMSE$ of this combination were 0.876, 4.085 g/kg and 18.8%, respectively, in training; these reached 0.803, 5.203 g/kg and 23.9%, respectively, in testing. Hence, our study showed spectral reflectance and machine learning could be used to estimate soil free iron content rapidly, non-destructively, and economically. Given these valuable findings, the present study benefits soil properties mapping, crop nutrient management and improving environmental issues.

Keywords Free iron content, Soil, Spectral reflectance, Machine learning, Deep neural network

Rapid and reliable assessment of the physical and chemical properties of soil is an important step in agricultural and natural resource management^{1,2}. Iron is an essential trace element for plant growth, as it promotes the synthesis of chlorophyll and maintains photosynthesis³⁻⁵. Meanwhile, iron can also regulate the absorption of nitrogen, phosphorus, and other elements in plants, thereby promoting various metabolic processes required for plant growth and development^{6,7}. Free iron, as one form of iron in soil mainly refers to the forms of iron in soil that are not part of silicate components^{8,9}. Free iron plays an important role in reflecting the impact of forming environments on soil⁹. The free iron content is a vital indicator of pedogenic processes in tropical soils and can be used to understand the soil's weathering history and aid in classification¹⁰. Soil free iron not only

¹Institute of Digital Agriculture, Zhejiang Academy of Agricultural Sciences, 298 Desheng Middle Road, Hangzhou 310021, Zhejiang, China. ²Key Laboratory of Information Traceability for Agricultural Products, Ministry of Agriculture and Rural Affairs of China, Hangzhou 310021, China. ³Zhejiang Institute of Hydraulics and Estuary, Hangzhou 310020, China. ✉email: huh@zaas.ac.cn

significantly affects soil color, but also reflects the occurrence and evolution of soil¹¹. Therefore, accurate and efficient determination of soil free iron content is of great significance.

Traditionally, the determination of soil free iron content is carried out mainly by leaching with sodium hydrosulfite sodium citrate bicarbonate and measuring with an o-phenanthroline colorimetric method. This chemical analysis of solution extraction and colorimetric determination has the advantage of accurate and reliable test results. However, this process is tedious, expensive, laborious, and time-consuming, and may result in environmental pollution. Spectral reflectance technology has emerged in recent years as a promising alternative, given its capability to measure the reflectance of earth surface features, such as soil, water, and vegetation, at hundreds of contiguous and narrow wavelength bands^{12–14}. The availability of such a large pool of spectral information offers an opportunity to estimate soil attributes rapidly, lower costs, and in a non-destructive manner^{15,16}. Some studies have clearly shown that spectral reflectance may be used for estimating several soil properties such as soil organic matter¹⁷, nitrogen¹⁸, phosphorus¹⁹, CEC (cation exchange capacity)²⁰, soil texture²¹, and potassium¹⁹, organic carbon²², and moisture^{23,24}.

Estimating free iron content in soils using the reflectance spectrum offers a rapid, non-destructive, and cost-effective alternative to traditional methods²⁵. The technique leverages the spectral characteristics of iron and robust models to provide accurate and reliable measurements²⁶. However, to the best of our knowledge, most studies have focused on soil total iron, crystalline iron, amorphous iron, goethite, and hematite^{9,11,27–35}. Studies on spectral estimation of soil free iron content are very few. In 2019, Liu and Sun³⁶ reported that FD (first derivative), SD (second derivative), SNV (standard normal variate), and CR (continuum removed) transformed spectra had a significant relationship with soil free iron content at 747, 585, 2187, 1153, and 1380 nm. Their correlation coefficients were 0.324, 0.646, −0.529, 0.655, and −0.467. In 2009, Richter et al.³⁷ proposed a methodology that directly links free iron content using an iron absorption band near 900 nm (Fe-NIR). Based on the Fe-NIR absorption depth, free iron content prediction models were built. The model determination coefficient (R^2) and mean square error were 0.87 and 13.9%, respectively. On modeling methods, Adeline et al.³⁸ used PLSR (partial least squares regression) to build free iron prediction model and reported that the model R^2 was 0.78. In summary, these studies on free iron estimation based on soil reflectance focused on sensitive wavelengths response to soil free iron and using simple and single methods to build the estimation models. Multiple spectral variable selection of soil spectral data and modeling of soil free iron content using machine learning have been still relatively limited.

The objectives of the present study are: (1) to pretreat and transform the soil reflectance using the FD, SNV, CR methods and to analyze the spectral bands sensitive to free iron content; (2) to compare the methods of spectral variable selection methods such as full spectrum, correlated spectrum, and principal component analysis (PCA) and select suitable spectral features as model input variables, and (3) to develop the soil free iron content estimation models using a partial least squares (PLS), support vector machine (SVM), random forest (RF), deep neural network (DNN) approaches. This study can offer a rapid, viable, non-invasive solution for estimating soil properties. Our method has the potential for high-throughput analysis and field applicability. It will benefit soil properties mapping, crop nutrient management, and improving environmental issues.

Materials and methods

Study area and soil sampling points

The study area is located in Zhejiang (27.4°–30.0° N, 118.4°–122.0° E), in southeastern China (Fig. 1a). It has a total area of 16,850 km² with an elevation of 58.45 m above sea level. It has a humid subtropical climate with a yearly average temperature of 18.4 °C, it receives precipitation of about 1395.3 mm, and it has 150 rainy days annually³⁹. Based on the land cover and topography, 135 local sites were selected and soil sampling was carried out in the study area (see Fig. 1a). Each sampling location was defined with GPS, and we took four samples from various soil horizons at 0–80 cm depth using a bamboo shovel. Approximately 1 kg of soil for each sample was collected by gathering five subsamples. In total, 540 soil samples were obtained.

Free iron content measurements

The free iron in the soil samples was extracted using dithionite-citrate-bicarbonate (DCB) treatment⁴⁰. The procedure includes the following steps: (1) sample preparation: air-dry and sieve soil samples to remove large particles and organic matter; (2) reagent preparation: prepare a DCB reagent composed of sodium dithionite, sodium citrate, and sodium bicarbonate; (3) extraction: weigh a specific amount of soil (usually 1–2 g), add the soil to a flask containing the DCB reagent, heat the mixture at 80 °C for 15 min while continuously stirring to ensure a complete reaction, cool the mixture, and separate the soil residue from the solution; via filtering, and (4) iron quantification: measure the concentration of iron in the solution using a spectrophotometer.

Statistics of our soil free iron content dataset are shown in Table 1. The soil iron content for the total, training, and testing samples varies from 4.07 to 60.3 g/kg, 4.49–60.3 g/kg, and 4.07–56.44 g/kg. The median values of soil iron content for the three datasets are 19.96 g/kg, 19.99 g/kg, and 19.82 g/kg. The Q1 (one quarter), and Q3 (three quarters) of the training and testing datasets are very close to those of the total dataset. The data distribution parameters among total, training, and testing data, i.e., skewness and kurtosis are 1.13, 1.16, and 1.06 and 1.09, 1.13, and 1.06. It indicates that training and testing data well represent the total data.

Reflectance measurements of soils

To minimize the effects of soil moisture and particle size on the spectral measurements, all samples were air-dried and sieved to 0.25 mm. Before reflectance measurement, the soil sample was put in a culture dish with a diameter of 5.0 cm and a depth of 1.0 cm. The dish was filled with soil sample and the soil surface was flat and flush with the dish. An ASD (Analytical Spectral Devices) FieldSpec 3 portable spectrometer (Malvern Panalytical Ltd, Malvern, UK) was employed to measure the reflectance of the soil samples (see Fig. 1b). The

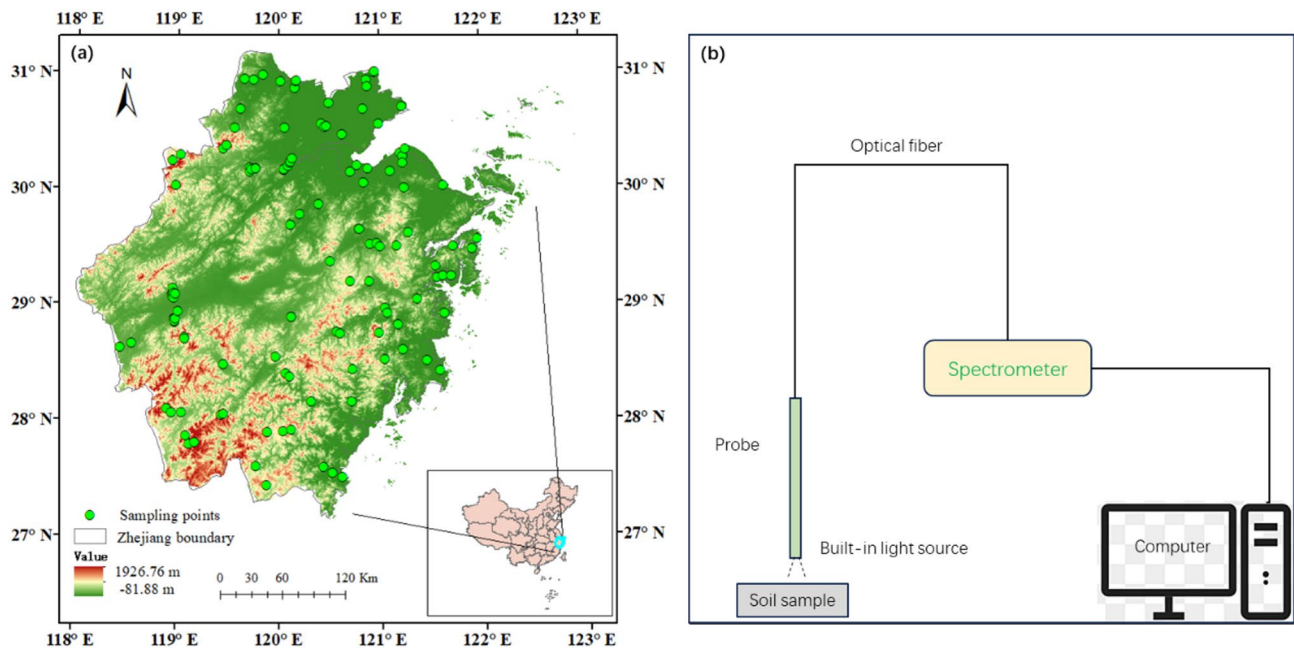


Fig. 1. Location of study area with sampling points (a) and diagram of soil spectral reflectance measurement (b). The map in Fig. 1(a) was generated by ArcGIS 10.8 software (GeoScene Information Technology Co., Ltd., <https://www.esri.com>).

	Number	Mean (g/kg)	Max (g/kg)	Min (g/kg)	Range (g/kg)	Median (g/kg)	Skewness	Kurtosis	Q1 (g/kg)	Q3 (g/kg)
Total	540	21.97	60.3	4.07	56.22	19.96	1.13	1.09	13.55	26.88
Training	360	22.17	60.3	4.49	55.81	19.99	1.16	1.13	13.67	26.92
Testing	180	21.57	56.44	4.07	52.37	19.82	1.06	0.99	13.47	26.62

Table 1. Statistics of our total, training, and testing datasets in our experiment.

light source was a built-in halogen light source probe with a front field of view angle of 25°. The probe should be positioned approximately 2 cm above the soil surface. The spectrometer’s wavelength range was 350–2500 nm. The spectral resolution was 3 nm in 350–1000 nm and 10 nm in 1000–2500 nm. The spectrometer was calibrated with a whiteboard before each sample measurement. To reduce the error, each sample was measured thrice, and for each measurement, 10 spectral curves were averaged to a representative one for subsequent analysis.

Pre-treatment and analysis of raw spectra

The soil spectral reflectance data range was optimized to the 400–2400 nm range for noise reduction. Spectral preprocessing methods such as FD, SNV, and CR transformations on the original spectral data were carried out in our study. The SNV performs a normalization of the spectra that consists in subtracting each spectrum by its own mean and dividing it by its own standard deviation. The CR technique is a method to highlight spectral absorption features. It can be viewed as a way to perform albedo normalization. Since reflectance containing huge data may make analysis complex and difficult, it is necessary to reduce the amount of data and select the appropriate spectral variables to construct a soil free iron content estimation model. We applied Pearson correlation analysis and PCA to reduce the dimensions and amount of spectral data.

Estimation models building

A 10-fold cross-validation (10-fold cross-validation) is used to verify the optimal model selected from different models (the most suitable model). Our dataset was composed of all our soil samples and divided into two parts using a stratified sampling method. The training set was 70% of the total data (i.e., 360 samples). This set was used to develop the estimation models. The testing set consisted of 30% of the total data (i.e., 180 samples). This set was used to test model performance. Based on full original spectral reflectance and its transforms (FD, SNV, and CR), soil free iron content estimation models were constructed. In the model utilizing Pearson correlation, spectral reflectance of original spectra, FD, SNV, and CR with correlation coefficients of higher than 0.400 was selected as model input. In the model employing PCA, principal components (PCs) with eigenvalues higher than 1.00 were selected as input variables.

To build soil free iron content estimation moestimated soil iron condels, we used a PLS, SVM, RF, and DNN. PLS can reduce the dimensionality of the spectral data while preserving the variance related to iron content. SVM can construct hyperplanes in a high-dimensional space to regress based on spectral features. A kernel function with “gaussian” in SVM were determined after 10-fold cross validation. RF is an ensemble method that uses multiple decision trees to improve predictive accuracy. It can combine the output of multiple decision trees to reach a single result.

DNN enables machines to learn complex patterns from data with high accuracy. When adequately trained, DNNs allow machine learning models to reliably interpret spectral data. The DNN structure we used is shown in Fig. 2. The DNN input layer has three kinds of selected features and four hidden layers. Layers 1 through 4 have 256, 128, 64, and 32 neurons, respectively. The output layer contained the soil free iron content data. We used a ReLU activation function after each hidden layer. Moreover, after the first hidden layer, a dropout layer with a ratio of 0.1 was used to prevent overfitting. We trained the network with the Adam optimizer. The maximum number of training rounds was set to 500, and the mini-batch size was 32. The initial learning rate was set to 0.001 and decreased by 10% every 100 rounds. The flowchart of the entire process starting from data collection to analysis and modeling is shown in Fig. 3.

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

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

$$RRMSE = \sqrt{\frac{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i)^2}} \times 100\% \quad (3)$$

Where y_i and $\hat{y}_i$ are a measured and predicted soil free iron content at sample i , respectively. And $\bar{y}$ is the average value of measured soil free iron content. N is the number of observations. The R^2 , root mean square error (RMSE), and relative root mean squared error (RRMSE) were used to evaluate training and testing performance (Eqs. (1)– (3)). R^2 represents the proportion of the variance for a dependent variable that's explained by an independent variable. $RMSE$ measures the average difference between a statistical model's predicted values and the actual values. $RRMSE$ reflects accuracy and allow us to compare the accuracy of different models.

We conducted a two tailed Pearson correlation analysis in Excel 2022 and a PCA using IBM SPSS Statistics 25.0 (SPSS Inc., NY, USA, 2017). The correlation, scatter, fitted line, and PCA plot were drawn by Origin 2022 (Origin Lab Corporation, MA, USA, 2022). The transform of FD and SNV and PLS algorithm was carried out in

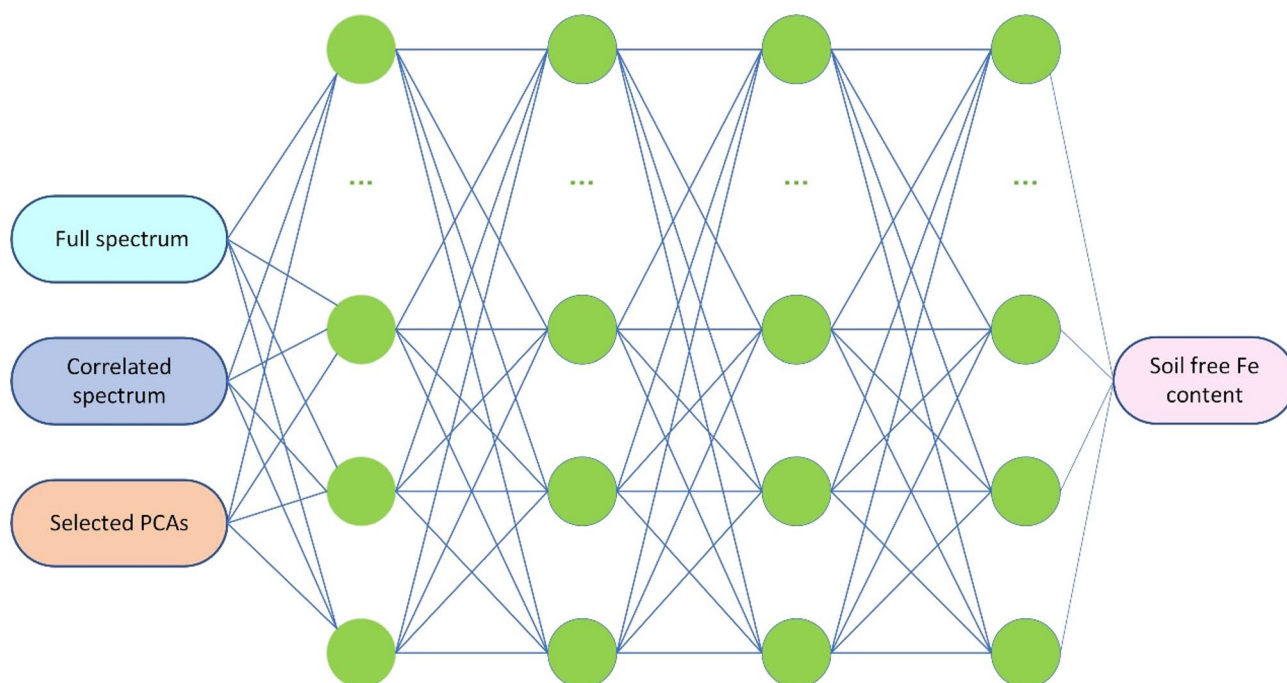


Fig. 2. Deep neural network (DNN) structure employed in our study.

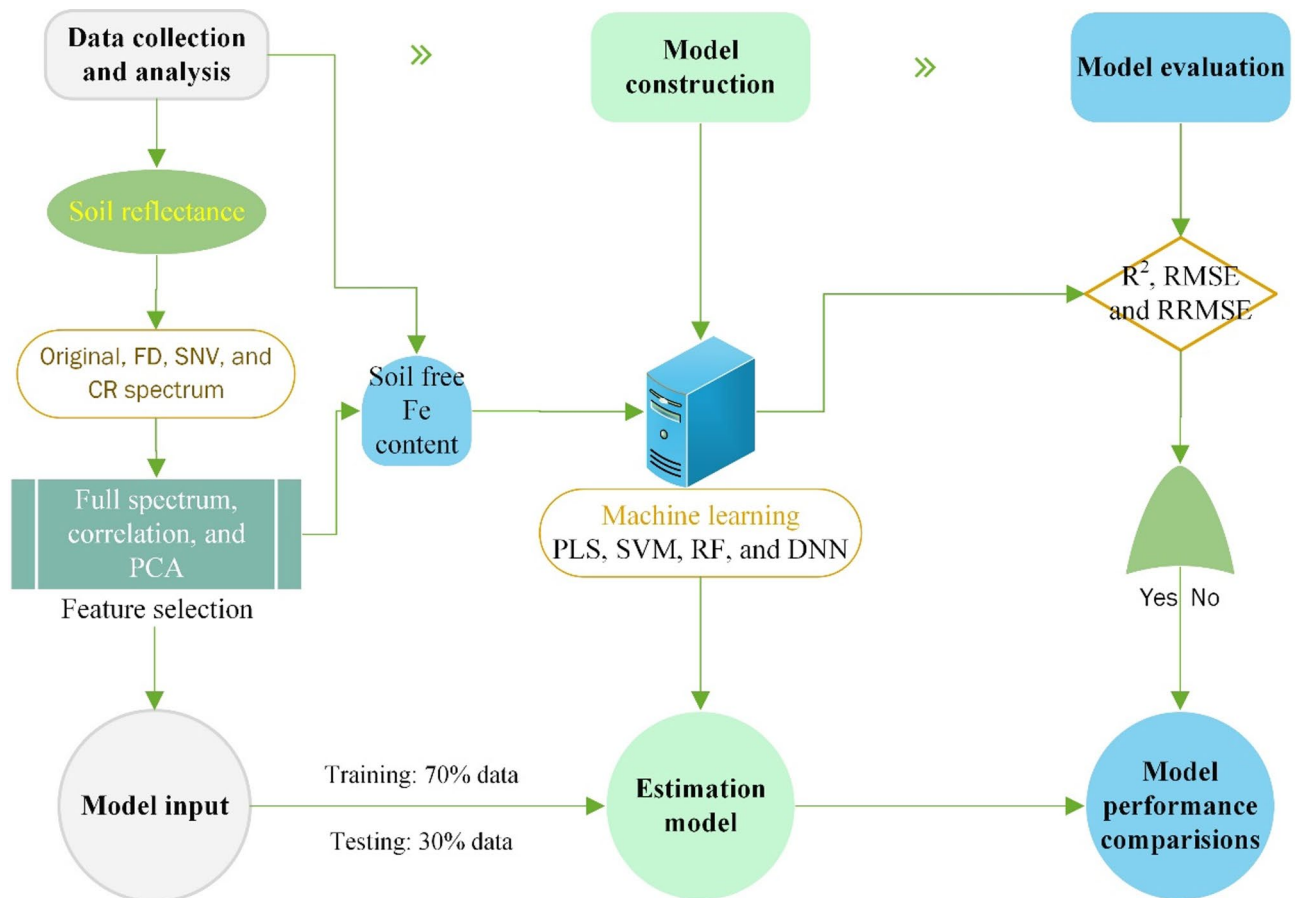


Fig. 3. Workflow of the soil free iron content estimation model.

The Unscrambler X 10.4 (CAMO Software AS., 2016). The SVM, RF algorithm, and DNN were run in MATLAB R2022a (The MathWorks, Inc., CA, USA, 2022). CR transformation of spectral data was processed in ENVI 5.3 (Itt Visual Information Solutions, CO, USA, 2015).

Results

Soil original, FD, SNV, and CR spectra

The soil's original reflectance and its FD, SNV, and CR transformations are shown in Fig. 4. It can be seen that in the visible spectral band, the original spectral reflectance is low, but it increases rapidly; In the near-infrared band, the spectral reflectance is high and the spectral curve changes relatively smoothly (Fig. 4a). Moreover, there are absorption valleys at wavelengths of 1414, 1915, and 2207 nm. The SNV spectral curve has a similar change tendency (Fig. 4b). In the FD spectral curve, the absorption valleys are at 404, 1409, 1898, and 2201 nm. In the CR curve, the absorption valleys are at 488, 1414, 1917, and 2207 nm.

Correlation coefficient of soil iron content in original, FD, SNV, and CR spectra

The correlation coefficient of soil iron content and the original, FD, SNV, and CR spectra within the wavelength of 400–2400 nm is shown in Fig. 5a. In the original spectrum, the spectral reflectance at 400–605 nm and 847–2400 nm had a highly significant relationship with free iron content ($P < 0.01$), and at 400–685 nm and 797–2400 nm was significant relationship ($P < 0.05$) with soil free iron content. Spectral values of the FD spectrum at 447–477 nm, 514–918 nm, 995–1209 nm, 1274–1436 nm, 1584–1995 nm, and 2097–2385 nm had highly significant relationships with soil free iron content. In the SNV spectrum, strong relationships between spectral values and soil free iron content were observed at 449–541 nm, 569–945 nm, 973–1397 nm, 1428–1571 nm, and 1798–2400 nm. Finally, for the CR spectrum, such strong relationships were observed at 401–598 nm, 773–1170 nm, 1354–1567 nm, 1869–2085 nm, and 2134–2399 nm.

The original spectrum at 400–536 nm, the FD spectrum at 560–571 nm and 574–610 nm, the SNV spectrum at 1142–1381 nm, and the CR spectrum at 403–580 nm and 795–1120 nm, all showed very significant correlations with soil free iron content, they were all higher than 0.400. These above spectral data would thus be selected as soil free iron content estimation model input variables.

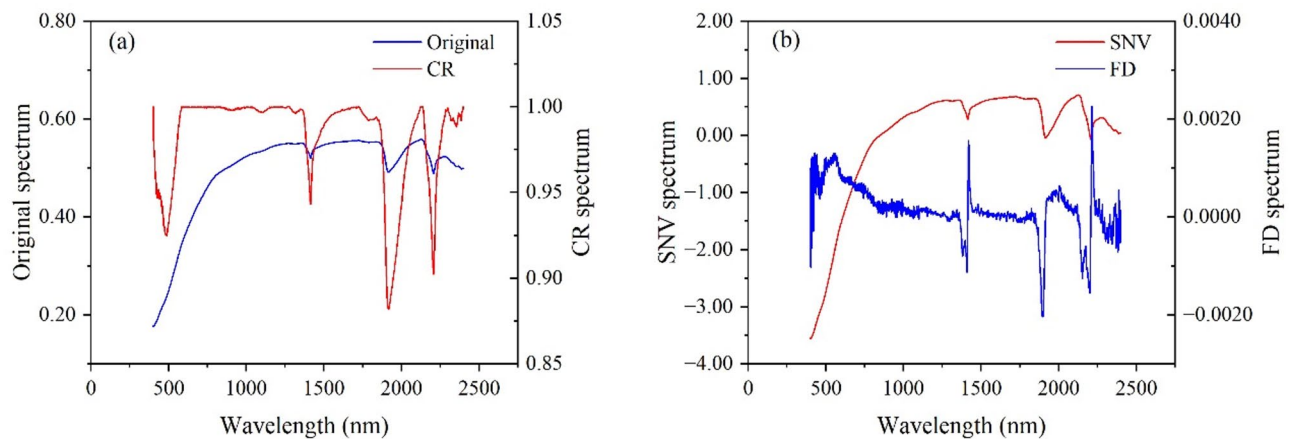


Fig. 4. Original spectrum and CR (a), SNV, and FD (b) spectra with the wavelength ranging from 400 to 2400 nm. *FD* first derivative, *SNV* standard normal variate, *CR* continuum removed.

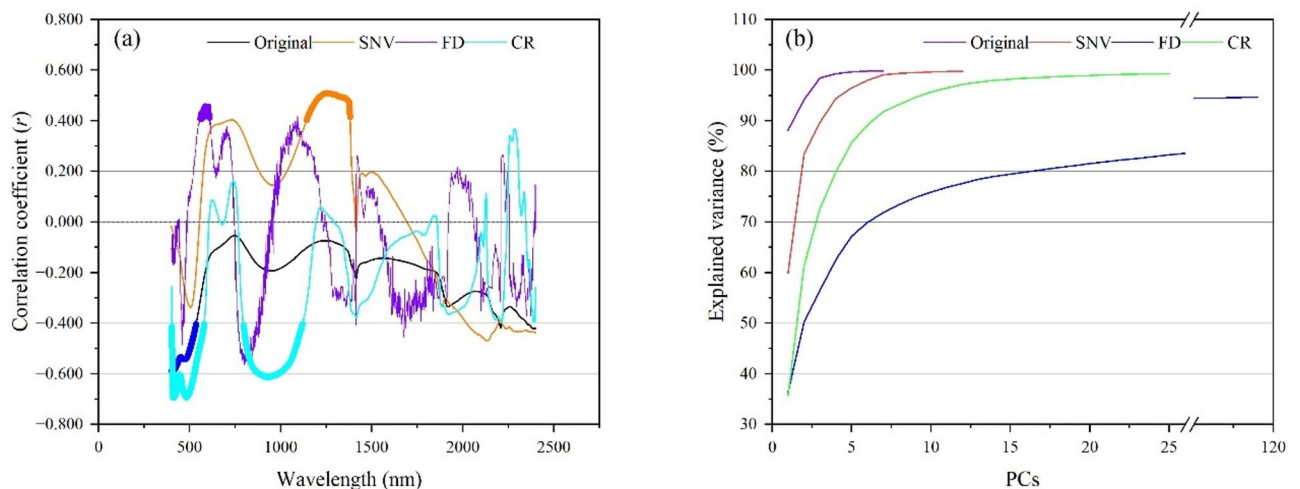


Fig. 5. Correlation coefficient of soil free content and the original and CR, SNV, and FD spectra (a), and explained variance for principal components (PCs) output by a PCA applied to original spectra and CR, SNV, and FD spectra (b). In (a) plot, bolded lines indicate a correlation coefficient value is greater than 0.400. $n = 540$. In (b) plot, violet, red, green, and blue line represent original spectra and CR, SNV, and FD spectra. *FD* first derivative, *SNV* standard normal variate, *CR* continuum removed.

PCA projection of original spectra, FD, SNV, and CR spectra

We conducted PCA projection of the original spectra, FD, SNV, and CR spectra at the wavelengths of 400–2400 nm. The results are shown in Fig. 5b. Seven PCs were extracted from the original spectrum. For FD, SNV, and CR, 119 PCs, 12 PCs, and 25 PCs were obtained, respectively. These PCs can represent most of the information of original spectra, FD, SNV, and CR spectra and are used as the model input variables.

Model performance

The performance of the soil free iron content estimation based on the full spectral spectrum, correlated spectrum, and PCA projection variable selection methods is shown in Tables 2 and 3, and Table 4, respectively. The scatter plot of training and testing data and 1:1 line (i.e., estimated value equal to measured value) is drawn in Figs. 6 and 7, and Fig. 8.

As shown in Table 2, when using the full spectrum, the training model employing the RF algorithm had the highest R^2 (0.954) of all training models. Meanwhile, for testing models, the model utilizing DNN had the highest R^2 (0.718). In terms of *RMSE* and *RRMSE*, the training model with the SVM had the lowest values, while for testing models, the model using the PLS algorithm had the lowest values. For average R^2 , FD had the highest value in the training and testing model. For average *RMSE* and *RRMSE*, CR had the lowest value in training and FD had the lowest in testing. Among the modeling methods, RF had the highest R^2 and lowest *RMSE*, *RRMSE* in training; and DNN had the highest R^2 and lowest *RMSE*, *RRMSE* in testing.

Spectral	Modeling	Training			Testing		
Transform	Method	R^2	RMSE (g/kg)	RRMSE (%)	R^2	RMSE (g/kg)	RRMSE (%)
Original	PLS	0.534	7.985	36.0	0.477	8.141	37.7
	SVM	0.536	8.198	38.9	0.479	8.374	38.8
	RF	0.930	3.908	17.7	0.476	8.437	38.4
	DNN	0.627	7.544	33.0	0.532	8.114	37.0
FD	PLS	0.689	6.525	29.4	0.659	6.563	30.4
	SVM	0.756	6.369	30.3	0.693	6.724	31.2
	RF	0.954	3.685	16.7	0.559	8.156	37.2
	DNN	0.894	4.435	19.2	0.718	6.622	30.2
SNV	PLS	0.576	7.617	34.4	0.587	7.216	33.4
	SVM	0.593	7.795	37.0	0.550	7.794	36.1
	RF	0.933	3.794	17.2	0.479	8.400	38.3
	DNN	0.721	6.642	29.8	0.573	7.779	35.4
CR	PLS	0.617	7.236	32.6	0.574	7.325	34.0
	SVM	0.937	2.921	13.3	0.602	7.451	33.9
	RF	0.944	3.382	15.4	0.628	7.230	32.9
	DNN	0.761	6.200	27.8	0.643	7.383	33.6

Table 2. Performance of soil free iron content models based on full spectrum.

Spectral	Modeling	Training			Testing		
Transform	Methods	R^2	RMSE (g/kg)	RRMSE (%)	R^2	RMSE (g/kg)	RRMSE (%)
Original	PLS	0.605	7.361	33.0	0.549	7.543	35.0
	SVM	0.737	6.153	29.0	0.557	7.768	35.4
	RF	0.883	4.308	19.6	0.335	9.558	43.5
	DNN	0.748	5.929	26.2	0.654	6.841	31.2
FD	PLS	0.569	7.721	33.9	0.415	8.589	39.8
	SVM	0.585	7.799	37.8	0.458	8.786	40.0
	RF	0.893	4.759	21.5	0.147	10.825	49.3
	DNN	0.607	7.554	33.4	0.623	7.742	35.3
SNV	PLS	0.471	8.509	38.4	0.434	8.463	38.2
	SVM	0.546	7.899	37.1	0.478	8.426	38.4
	RF	0.898	4.369	19.9	0.578	7.585	34.6
	DNN	0.666	7.014	31.7	0.626	7.219	32.9
CR	PLS	0.694	6.472	28.8	0.636	6.776	31.4
	SVM	0.702	6.461	30.8	0.636	7.010	31.9
	RF	0.823	3.612	16.4	0.502	8.122	37.0
	DNN	0.738	6.144	27.1	0.646	6.990	31.8

Table 3. Performance of soil free iron content models based on correlated spectrum.

When using the correlated spectrum (see Table 3), we found that the training model employing the RF algorithm had the highest R^2 (0.898) and lowest RMSE and RRMSE, meanwhile, our testing model with the DNN had the highest R^2 (0.718) and the lowest RMSE and RRMSE. On average, the original model had the highest R^2 , and the CR had the lowest RMSE and RRMSE in training; in testing, the CR had the highest R^2 , and the lowest RMSE and RRMSE. Among the modeling methods, RF model had the highest R^2 , lowest RMSE, and RRMSE in training; meanwhile, DNN model had the highest R^2 , lowest RMSE, and RRMSE in testing.

When employing PCs (see Table 4), on average, the highest R^2 (0.903) and the lowest RMSE (4.440 g/kg) and RRMSE (19.8%) in training were observed when using the FD transform. For testing, we obtained the highest R^2 with the CR transform and the lowest RMSE and RRMSE with the original spectrum. Comparing the modeling methods, the model with the RF algorithm in training and the model with the DNN in testing had the highest R^2 and the lowest RMSE and RRMSE.

Discussion

Spectral technology may become a promising tool for a rapid, non-destructive, low-cost and relatively accurate evaluation of physical and chemical properties⁴¹. Sensitive and key wavelengths of free iron which mainly contains Fe^{2+} and Fe^{3+} were investigated in this study. We found that 1414 nm, 1915 nm, and 2207 nm were the

Spectral	Modeling	Training			Testing		
Transform	Methods	R^2	RMSE (g/kg)	RRMSE (%)	R^2	RMSE (g/kg)	RRMSE (%)
Original	PLS	0.641	7.02	31.2	0.609	7.041	32.6
	SVM	0.705	6.633	31.4	0.645	6.881	31.4
	RF	0.953	3.244	14.7	0.736	6.334	28.9
	DNN	0.805	5.453	24.6	0.751	5.902	26.9
FD	PLS	0.847	4.579	20.6	0.76	5.539	25.7
	SVM	0.876	4.085	18.8	0.803	5.231	23.8
	RF	0.99	4.497	20.3	0.460	10.325	47.0
	DNN	0.900	4.597	19.4	0.807	5.875	26.8
SNV	PLS	0.642	7.019	31.4	0.625	6.967	32.3
	SVM	0.713	6.408	30.3	0.623	7.264	33.1
	RF	0.950	3.396	15.4	0.639	7.204	32.8
	DNN	0.812	5.203	23.7	0.752	5.955	27.1
CR	PLS	0.790	7.019	31.4	0.732	6.967	32.3
	SVM	0.726	6.408	30.3	0.715	7.264	33.1
	RF	0.960	3.396	15.4	0.703	7.204	32.8
	DNN	0.826	5.203	23.7	0.715	5.955	27.1

Table 4. Performance of soil free iron content models based on principal component analysis (PCA) spectrum.

absorption valleys of free iron in the soil original spectral curve. In the CR spectrum, 488 nm, 1414 nm, 1915 nm, and 2207 nm were considered as the absorption valleys (Fig. 5). Some studies reported that the absorption bands of soil Fe^{3+} and Fe^{2+} were near 380 and 480 nm, respectively, and that they ranged from 600 to 2100 nm (or 920 nm)^{9,11,29}. Mulder (2009)³⁵ reported that the total iron absorption features located at 550 nm and 880 nm in the soil continuum-removed soil spectrum. Our results share some similarities to those of previous studies. The difference may be due to the form and composition of iron in the soil samples.

The pretreated and transformed reflectance spectra of soil were employed to quantify the total amount of iron in the soil. They may help eliminate irrelevant information and noise in spectral data. In 2018, Kayande et al.³¹ used the FD transform to estimate the soil iron oxide content, and their model's $RMSE$ and R^2 were 0.008711 and 0.91, respectively. Rathod et al.⁴² estimated soil total iron content using FD and CR transforms, and their model's R^2 ranged from 0.84 to 0.86. In our study, the FD transform brought our model's average R^2 as high as 0.797; meanwhile, the average R^2 of our model using the original spectrum was 0.725 (Table 5). This indicated that preprocessing and transformation of reflectance may improve the soil iron estimation model accuracy in comparison to using the raw data.

Variable selection refers to the process of choosing the most relevant variables to include in a regression model. They improve model performance and help avoid overfitting. In our study, we selected variables through Pearson correlation analysis and PCA projection. It can be seen that the performance of the model following PCA projection was the best. The correlated method is inferior to the full spectrum methods. In soil iron content estimation model construction, many researchers use linear regression, multiple regression, and PLS algorithms^{15,31,34}. Machine learning regression is a technique for investigating the relationship between independent variables or features and a dependent variable or outcome. Good use of machine learning may improve model accuracy. In 2021, Hu et al.³³ estimated soil iron content using a back propagation neural network and obtained R^2 and $RMSE$ values of 0.955 and 0.336%, respectively; these were better than those obtained using a linear regression model (where the R^2 was 0.859 and the $RMSE$ was 1.07%). In our study, RF and DNN were used to estimate soil free iron content, and the R^2 and $RRMSE$ of the training model were 0.926, 17.5% and 0.720, 28.3%. Thus, our models using the RF algorithm and DNN may be better than those utilizing the PLS algorithm and SVM. However, model performance is not as good during testing as it is during training. Specifically, the RF-based model has the lowest R^2 and the highest $RMSE$ and $RRMSE$, which suggests that it may overfit the data. This is seen in Fig. 6,7,8. Thus, the use of RF algorithms to estimate soil iron content needs further study. DNN is an exciting discipline. It has already transformed the way data is analysed and modeled in estimation of soil properties. In 2023, Lee et al.⁴³ reported DNN-based soil moisture estimated value had better agreement with in situ measured value and their correlation coefficient was 0.9226. Maino et al.⁴⁴ also showed that DNN algorithms can predict clay and sand soil contents from spectrometry data. In our study, DNN-based soil free iron content model's R^2 was 0.720, which was higher than that of the PLS and SVM models. Compared with RF, DNN may effectively avoid underfitting and overfitting of the models by using the algorithm's hyperparameters through various configurations to optimize the models' performance. Additionally, though we tested generalization across training and test sets, we did not validate models across independently collected datasets from different regions or instruments. Model transferability of soil free iron content may be our study limit and needs further study.

In our study, the impact of individual preprocessing steps including reflectance transformation, feature selection, and modeling algorithms on model accuracy were evaluated. The present results showed the effects of each combination, allowing comparison of individual preprocessing effects. But the combined effect of

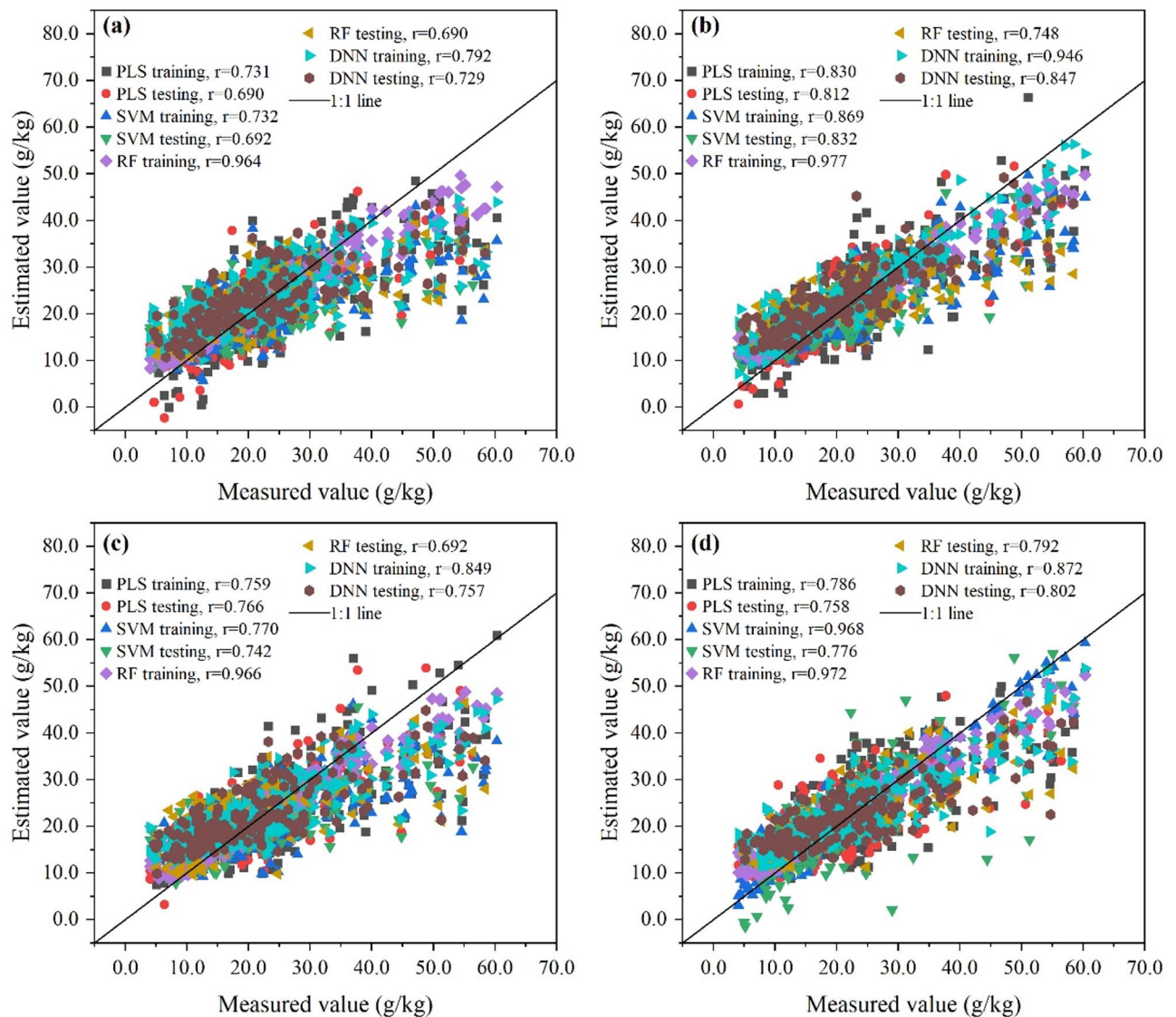


Fig. 6. Measured and estimated soil free iron content based on full spectrum data using the original (a), and FD (b), SNV (c), and CR (d) transformed spectrum. FD first derivative, SNV standard normal variate, CR continuum removed. r linear correlation coefficient.

multiple preprocessing steps (e.g., FD + CR vs. no SNV + CR under different modeling conditions) has not been investigated in our study. It may be our study limit and could be analyzed more systematically in future studies. Meanwhile, we acknowledge that reflectance transformation, feature selection, and modeling algorithms contain more methods than those used in our study. More data mining and modeling methods need further research.

Many factors, such as soil organic matter, water content, and texture, affect the soil reflectance, and thus affect the estimation model accuracy. Heller Pearlshien and Ben-Dor²⁹ showed the spectral responses of soil organic matter and iron oxides overlapped in the VIS–NIR spectral region. This should be taken into account when iron oxide and organic matter contents in the soil are to be spectrally estimated. Francos²⁷ also found that soil physical and chemical properties can have overlapping spectral features. Soil organic matter content interferes with the assessment of iron content. The estimation model R^2 and $RMSE$ were improved from 0.61, 0.81 to 0.83, 0.55 after soil organic matter was removed. In our study, the estimated soil free iron content value tended to be less than the true value when the true value was higher than 40 g/kg (Fig. 6,7,8). This may be because soil reflectance was influenced by organic matter. Moreover, when the soil free iron content is higher, and thus has a greater influence on reflectance. Additionally, clayey soils tend to have different reflectance properties compared to sandy soils due to differences in particle size and mineralogy. Soil moisture affects reflectance by altering the absorption and scattering properties of the soil, thereby potentially obscuring the spectral signatures of iron oxides. The effects of soil properties on reflectance require further investigation in the context of soil free iron content estimation.

In addition, excessive iron may cause environmental pollution to soil and water bodies. Specially, soil iron could transfer into plant food production. This may have great effect on growth and development of human

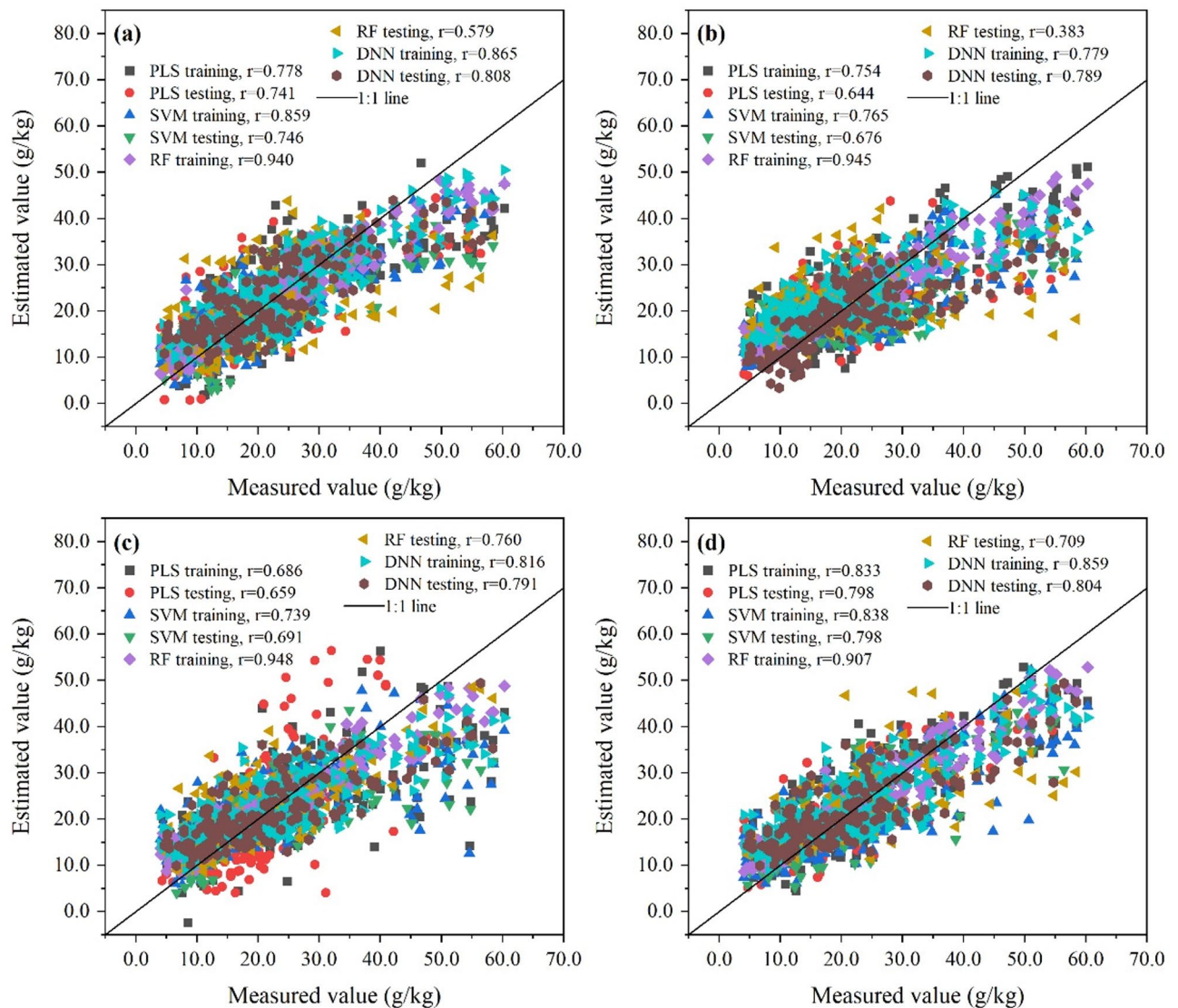


Fig. 7. Measured and estimated soil free iron content based on correlated spectrum data using the original (a), and FD (b), SNV (c), and CR (d) transformed spectrum. *FD* first derivative, *SNV* standard normal variate, *CR* continuum removed. r linear correlation coefficient.

beings. Effect of soil iron on agronomy community needs more study. Meanwhile, by integrating spectral reflectance data with machine learning models, a rapid, non-destructive, and cost-effective approach for predicting free iron content is given. This allows for improved spatial mapping of this critical soil chemical property at higher resolutions and frequencies than conventional laboratory-based analyses. What's more, our model outputs can be linked with geospatial data (e.g., GPS, remote sensing imagery), enabling the generation of spatial distribution maps of soil free iron over large agricultural regions. This is particularly useful for precision agriculture and land degradation assessments. Besides this, incorporating multi-source remote sensing data (e.g., UAV, satellite) to develop regional-scale free iron mapping tools, including soil qualitative factors such as parent material or classification units to enhance model robustness, and exploring domain adaptation and transfer learning techniques to improve cross-site model transferability could be future research recommended.

Conclusions

In summary, we concluded soil free iron content estimation using spectral reflectance and machine learning. The original spectrum and its FD, SNV, and CR transforms were preprocessed. Full spectrum, correlated spectrum, and PCA were considered as modeling variable selection. PLS, SVM, RF, and DNN were used in our models. Results showed that the FD transform was a more efficient spectral reflectance transform than the SNV and CR transform and original spectrum. The average R^2 , $RMSE$, and $RRMSE$ when using the FD transform for training were 0.797, 5.550 g/kg, and 25.1%, respectively. When using SNV, the training model had a low accuracy and the average R^2 , $RMSE$, and $RRMSE$ were 0.710, 6.305 g/kg, and 28.9%. In testing models, CR had a higher accuracy than the other transforms and its R^2 , $RMSE$, and $RRMSE$ were 0.644, 7.140 g/kg, and 32.7%. Variable

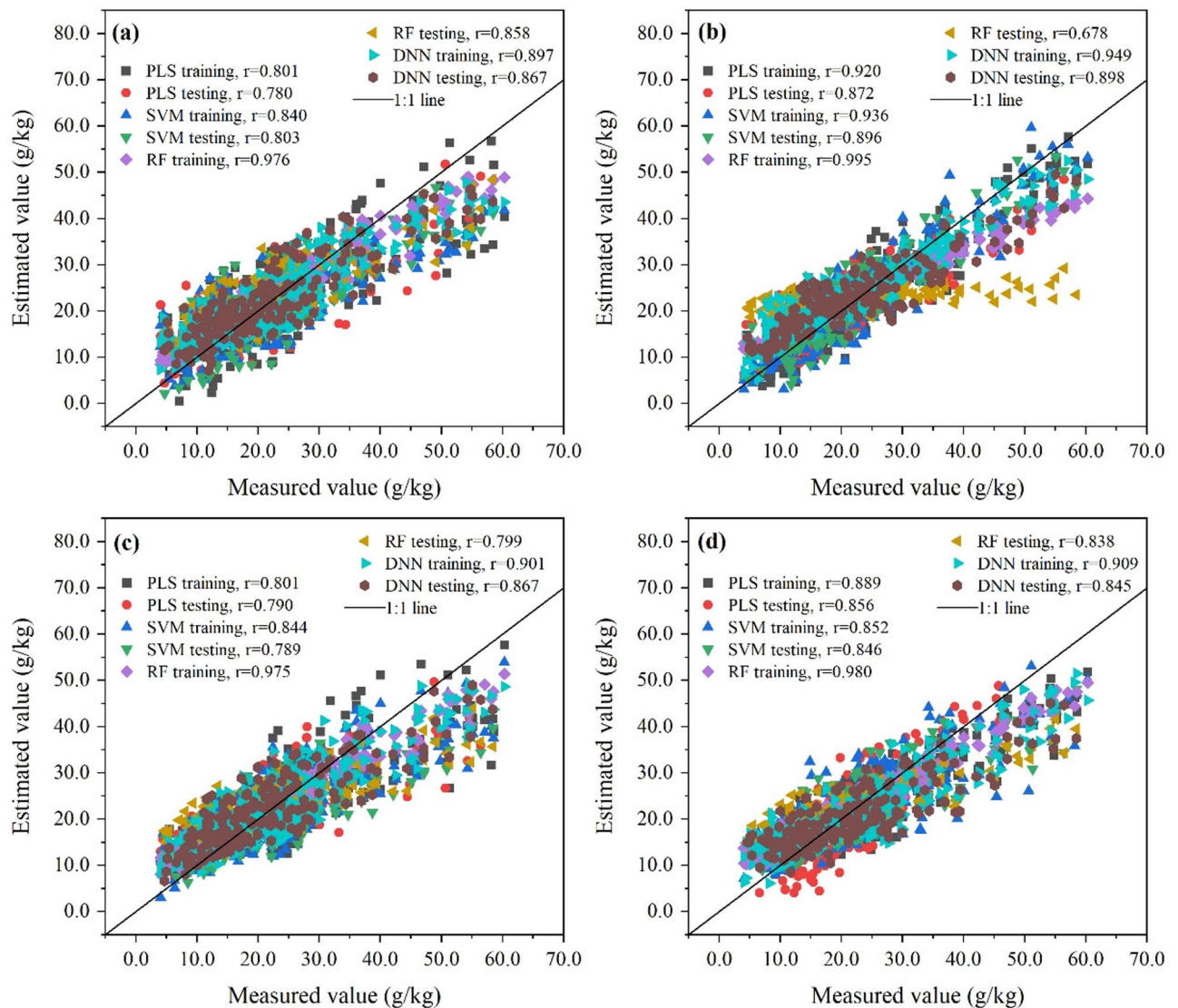


Fig. 8. Measured and estimated soil free iron content based on principal component analysis (PCA) spectrum data using the original (a), and FD (b), SNV (c), and CR (d) transformed spectrum. FD first derivative, SNV standard normal variate, CR continuum removed. r linear correlation coefficient.

selection based on PCA projection improved model accuracy compared to using full and correlated spectra. The average model R^2 , $RMSE$, and $RRMSE$ following PCA were 0.821, 5.260 g/kg, and 23.9% in training and 0.692, 6.744 g/kg, and 30.9% in testing, which had a higher R^2 and lower $RMSE$ and $RRMSE$ than when using the full and correlated spectra without PCA projection. Overfitting may have occurred in our study when employing the CR transform and RF algorithm. Their models had high accuracy in training and low accuracy in testing. The model R^2 using the DNN showed better performance than those using the PLS and SVM algorithm, but the DNN showed poorer performance in $RMSE$ and $RRMSE$ than that of the model utilizing the SVM and PLS algorithm (Table 4). The best combination of spectral transform, variable selection, and modeling method was FD + PCA + SVM. The R^2 , $RMSE$ and $RRMSE$ of this combination were as high as 0.876, 4.085 g/kg and 18.8%, respectively, in training; these reached 0.803, 5.203 g/kg and 23.9%, respectively, in testing. With rapid, non-destructive, low-cost, and relatively high accuracy, our model could produce reliable estimations of soil free iron content. With the continued advancements in spectral transforms, variable selection, and machine learning, reflectance spectroscopy has great potential to revolutionize soil properties mapping, crop nutrient management and improving environmental issues.

Spectrum transform, variable selection, and modeling method	Training			Testing		
	R ²	RMSE (g/kg)	RRMSE (%)	R ²	RMSE (g/kg)	RRMSE (%)
Original	0.725	6.145	27.9	0.567	7.578	34.7
FD	0.797	5.550	25.1	0.592	7.581	34.7
SNV	0.710	6.305	28.9	0.579	7.523	34.4
CR	0.793	5.371	24.4	0.644	7.14	32.7
Full	0.750	5.890	26.8	0.577	7.607	34.9
Correlated	0.698	6.379	29.0	0.517	8.015	36.6
PCA	0.821	5.260	23.9	0.692	6.744	30.9
PLS	0.640	7.089	31.8	0.588	7.261	33.6
SVM	0.701	6.427	30.4	0.603	7.414	33.9
RF	0.926	3.863	17.5	0.520	8.282	37.7
DNN	0.720	6.330	28.3	0.646	6.976	31.9

Table 5. Performance of soil free iron content models based on spectrum transforms, variable selection, modeling methods *. * values in this table were calculated from the data presented in Tables 2, 3 and 4.

Data availability

The data that support the findings of this study are not openly available due to reasons of sensitivity and are available from the corresponding author upon reasonable request.

Received: 21 February 2025; Accepted: 26 June 2025

Published online: 04 July 2025

References

- Awais, M. et al. AI and machine learning for soil analysis: an assessment of sustainable agricultural practices. *Bioresour. Bioprocess.* **10**, 90. <https://doi.org/10.1186/s40643-023-00710-y> (2023).
- Oliver, D. P., Bramley, R. G. V., Riches, D., Porter, I. & Edwards, J. Review: soil physical and chemical properties as indicators of soil quality in Australian viticulture. *Aust J. Grape Wine Res.* **19**, 129–139. <https://doi.org/10.1111/ajgw.12016> (2013).
- Patra, A. et al. Academic Press., in *Frontiers in Plant-Soil Interaction* (eds Tariq Aftab & Khalid Rehman Hakeem) 531–546 (2021). <https://doi.org/10.1016/B978-0-323-90943-3.00009-2>
- Therby-Vale, R., Lacombe, B., Rhee, S. Y., Nussaume, L. & Rouached, H. Mineral nutrient signaling controls photosynthesis: focus on iron deficiency-induced chlorosis. *Trends Plant. Sci.* **27**, 502–509. <https://doi.org/10.1016/j.tplants.2021.11.005> (2022).
- Bhat, M. A. et al. Soil and mineral nutrients in plant health: A prospective study of iron and phosphorus in the growth and development of plants. *Curr. Issues Mol. Biol.* **46**, 5194–5222. <https://doi.org/10.3390/cimb46060312> (2024).
- Ning, X. et al. Research progress on iron absorption, transport, and molecular regulation strategy in plants. *Front. Plant. Sci.* **14**, 1190768. <https://doi.org/10.3389/fpls.2023.1190768> (2023).
- Rout, G. R. & Sahoo, S. Role of iron in plant growth and metabolism. *Rev. Agric. Sci.* **3**, 1–24. <https://doi.org/10.7831/ras.3.1> (2015).
- Degórski, M. The relationships between different forms of iron and aluminium in soils as indicators of soil-cover development on india's Cherrapunji spur (Meghalaya Plateau). *Geogr. Pol.* **84**, 61–73. <https://doi.org/10.7163/GPol.2011.1.4> (2011).
- Sánchez-Marañón, M., Molinero-García, A., Delgado, R., García del Moral, L. F. & Martín-García, J. M. Spectral analysis of Fe oxidation in the early stages of weathering and soil formation. *Catena* **222**, 106850. <https://doi.org/10.1016/j.catena.2022.106850> (2023).
- Mendes, W. S. et al. Free iron oxide content in tropical soils predicted by integrative digital mapping. *Soil. Till Res.* **219** <https://doi.org/10.1016/j.still.2022.105346> (2022).
- Viscarra Rossel, R. A., Bui, E. N., de Caritat, P. & McKenzie, N. J. Mapping iron oxides and the color of Australian soil using visible–near-infrared reflectance spectra. *J. Geophys. Res. Earth Surf.* **115**, F04031. <https://doi.org/10.1029/2009jf001645> (2010).
- Gutierrez, M., Reynolds, M. P. & Klatt, A. R. Association of water spectral indices with plant and soil water relations in contrasting wheat genotypes. *J. Exp. Bot.* **61**, 3291–3303. <https://doi.org/10.1093/jxb/erq156> (2010).
- Ma, S. et al. Application of the water-related spectral reflectance indices: A review. *Ecol. Indic.* **98**, 68–79. <https://doi.org/10.1016/j.ecolind.2018.10.049> (2019).
- Vasava, H. B. & Das, B. S. Assessment of soil properties using spectral signatures of bulk soils and their aggregate size fractions. *Geoderma* **417** <https://doi.org/10.1016/j.geoderma.2022.115837> (2022).
- Sarathjith, M. C., Das, B. S., Wani, S. P. & Sahrawat, K. L. Variable indicators for optimum wavelength selection in diffuse reflectance spectroscopy of soils. *Geoderma* **267**, 1–9. <https://doi.org/10.1016/j.geoderma.2015.12.031> (2016).
- Nanni, M. R. & Dematté, J. A. M. Spectral reflectance methodology in comparison to traditional soil analysis. *Soil. Sci. Soc. Am. J.* **70**, 393–407. <https://doi.org/10.2136/sssaj2003.0285> (2006).
- Sun, W., Liu, S., Zhang, X. & Li, Y. Estimation of soil organic matter content using selected spectral subset of hyperspectral data. *Geoderma* **409** <https://doi.org/10.1016/j.geoderma.2021.115653> (2022).
- Vibhute, A. D., Kale, K. V., Gaikwad, S. V. & Dhupal, R. K. Estimation of soil nitrogen in agricultural regions by VNIR reflectance spectroscopy. *SN Appl. Sci.* **2** <https://doi.org/10.1007/s42452-020-03322-9> (2020).
- Shao, Y. & He, Y. Nitrogen, phosphorus, and potassium prediction in soils, using infrared spectroscopy. *Soil. Res.* **49** <https://doi.org/10.1071/sr10098> (2011).
- Wan, M. et al. Rapid Estimation of soil cation exchange capacity through sensor data fusion of portable XRF spectrometry and Vis-NIR spectroscopy. *Geoderma* **363** <https://doi.org/10.1016/j.geoderma.2019.114163> (2020).
- Tiruneh, G. A. et al. Use of soil spectral reflectance to estimate texture and fertility affected by land management practices in Ethiopian tropical Highland. *PLoS One*. **17**, e0270629. <https://doi.org/10.1371/journal.pone.0270629> (2022).
- Ribeiro, S. G. et al. Soil organic carbon content prediction using soil-reflected spectra: A comparison of two regression methods. *Remote Sens.* **13** <https://doi.org/10.3390/rs13234752> (2021).

23. Loshelder, J. I. & Coffman, R. A. Soil moisture content from spectral reflectance using visible, near-infrared, and short-wave infrared light. *J. Irrig. Drain. Eng.* **149** <https://doi.org/10.1061/JIJDH.IRENG-10055> (2023).
24. McGuirk, S. L. & Cairns, I. H. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. V-3-447–453 (2022). <https://doi.org/10.5194/isprs-annals-V-3-2022-447-2022> (2022).
25. Bartholomeus, H., Epema, G. & Schaepman, M. Determining iron content in mediterranean soils in partly vegetated areas, using spectral reflectance and imaging spectroscopy. *Int. J. Appl. Earth Obs Geoinf.* **9**, 194–203. <https://doi.org/10.1016/j.jag.2006.09.001> (2007).
26. Ben-Dor, E. Quantitative remote sensing of soil properties. *Adv. Agron.* **75**, 173–243. [https://doi.org/10.1016/S0065-2113\(02\)75005-0](https://doi.org/10.1016/S0065-2113(02)75005-0) (2002).
27. Francos, N., Gedulter, N. & Ben-Dor, E. Estimation of iron content using reflectance spectroscopy in a complex soil system after a loss-on-ignition pre-treatment. *J. Soil. Sci. Plant. Nutr.* **23**, 6866–6873. <https://doi.org/10.1007/s42729-023-01416-y> (2023).
28. Jiang, X. et al. Remotely sensed Estimation of total iron content in soil with harmonic analysis and BP neural network. *Plant. Methods.* **17**, 116. <https://doi.org/10.1186/s13007-021-00812-8> (2021).
29. Heller Pearlshien, D. & Ben-Dor, E. Effect of organic matter content on the spectral signature of iron oxides across the VIS–NIR spectral region in artificial mixtures: an example from a red soil from Israel. *Remote Sens.* **12**, 1960. <https://doi.org/10.3390/rs12121960> (2020).
30. Xia, W. et al. Influence of environmental factors on biodegradation of dispersed diesel oil in seawater. *Environ. Forensics.* **10**, 354–358. <https://doi.org/10.1080/15275920903347453> (2009).
31. Kayande, K. S., Deshmukh, R., Janse, P. & Kayte, J. Hyper spectral analysis of soil iron oxide using Fieldspec4 spectroradiometer. *Int. J. Comput. Sci. Eng.* **6**, 395–399. <https://doi.org/10.26438/jcse/v6i11.395399> (2018).
32. Badura, I. & Dąbski, M. Reflectance spectroscopy in geology and soil sciences: literature review. *Quaest Geogr.* **41** <https://doi.org/10.2478/quageo-2022-0031> (2022).
33. Hu, C., Guo, H., Mei, H. & Zhu, J. Prediction of iron content in soil based on microspectrophotometry analysis. *Forensic Sci. Int.* **318**, 110600. <https://doi.org/10.1016/j.forsciint.2020.110600> (2021).
34. Sahwan, W., Lucke, B., Sprafke, T., Vanselow, K. A. & Bäumler, R. Relationships between spectral features, iron oxides and colours of surface soils in Northern Jordan. *Eur. J. Soil. Sci.* **72**, 80–97. <https://doi.org/10.1111/ejss.12986> (2020).
35. Mulder, V. L. & Bartholomeus, H. in *6th EARSeL SIG IS workshop IMAGING SPECTROSCOPY: Innovative tool for scientific and commercial environmental applications*. 1–7 (2009). <https://edepot.wur.nl/173177>
36. Liu, J. & Sun, Z. *IOP Conf. Ser. : Earth Environ. Sci.* **300**, 022147 (IOP Publishing Ltd) <https://doi.org/doi:https://doi.org/10.1088/1755-1315/300/2/022147> (2019).
37. Richter, N. et al. Free iron oxide determination in mediterranean soils using diffuse reflectance spectroscopy. *Soil. Sci. Soc. Am. J.* **73**, 72–81. <https://doi.org/10.2136/sssaj2008.0025> (2009).
38. Adeline, K. R. M., Gomez, C., Gorretta, N. & Roger, J. M. Predictive ability of soil properties to spectral degradation from laboratory Vis–NIR spectroscopy data. *Geoderma* **288**, 143–153. <https://doi.org/10.1016/j.geoderma.2016.11.010> (2017).
39. Zhejiang Climate Center. *Climate Impact Assessment*, (2022). <http://www.weather.com.cn/images/zhejiang1/qhbh/yqhyxpj/2024/01/31/61366E7C17529BFDD1A79F0555F039E8.pdf>>
40. Coffin, D. E. A method for the determination of free iron in soils and clays. *Can. J. Soil. Sci.* **43**, 7–17. <https://doi.org/10.4141/cjss63-002> (1963).
41. Cozzolino, D. & Morón, A. The potential of near-infrared reflectance spectroscopy to analyse soil chemical and physical characteristics. *J. Agric. Sci.* **140**, 65–71. <https://doi.org/10.1017/s0021859602002836> (2003).
42. Rathod, P. H., Muller, I., Van der Meer, F. D. & de Smeth, B. Analysis of visible and near infrared spectral reflectance for assessing metals in soil. *Environ. Monit. Assess.* **188**, 558. <https://doi.org/10.1007/s10661-016-5568-9> (2015).
43. Lee, S. J. et al. Estimation of high-resolution soil moisture in Canadian croplands using deep neural network with sentinel-1 and sentinel-2 images. *Remote Sens.* **15** <https://doi.org/10.3390/rs15164063> (2023).
44. Maino, A. et al. A deep neural network for predicting soil texture using airborne radiometric data. *Radiat. Phys. Chem. Oxf. Engl.* **221**, (2024). <https://doi.org/10.1016/j.radphyschem.2024.111767>

Author contributions

W.M. and H.H. wrote the main manuscript text and prepared all the figures and tables. H.Z. and Z.Z. assisted in processing the materials, K.Z. and G.Z. were responsible for guiding the experimental plan. All the authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.