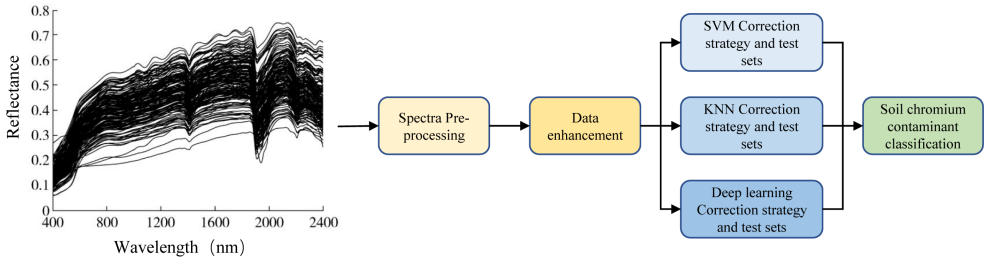
Graphical Abstract

Hyperspectral monitor of soil chromium contaminant based on deep learning network model in the Eastern Junggar coalfield

Yuan Wang,Hongbing Ma,Jingzhe Wang,Li Liu,Matti Pietikäinen,Zipeng Zhang,Xiangyue Chen



Highlights

Hyperspectral monitor of soil chromium contaminant based on deep learning network model in the Eastern Junggar coalfield

Yuan Wang,Hongbing Ma,Jingzhe Wang,Li Liu,Matti Pietikäinen,Zipeng Zhang,Xiangyue Chen

- Hyperspectral data were used to detect Chromium(Cr) pollution.
- A data enhancement method combined with a deep learning method showed the optimal accuracy.
- Our method is expected to provide large-scale accurate, real-time monitoring and guidance in soil pollution.

Hyperspectral monitor of soil chromium contaminant based on deep learning network model in the Eastern Junggar coalfield

Yuan Wang[a,*,1], Hongbing Ma[b], Jingzhe Wang[c,1], Li Liu[d,e], Matti Pietikäinen[e], Zipeng Zhang[f] and Xiangyue Chen[g]

[a]College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China
[b]Electronic Engineering, Tsinghua University, Beijing 100084, China
[c]MNR Key Laboratory for Geo-Environmental Monitoring of Great Bay Area & Guangdong Key Laboratory of Urban Informatics & Shenzhen Key Laboratory of Spatial Smart Sensing and Services, Shenzhen University, Shenzhen 518060, China
[d]College of System Engineering, National University of Defense Technology, Changsha 410073, China
[e]Center for Machine Vision and Signal Analysis, University of Oulu, Oulu 90570, Finland
[f]College of Resource and Environment Sciences, Xinjiang University, Urumqi 830046, China
[g]College of Atmospheric Sciences, Lanzhou University, Lanzhou 730000, China

ARTICLE INFO

Keywords:
soil hyperspectrum
soil heavy metal pollution
data enhancement (DA)
support vector machine (SVM)
k-nearest neighbour (KNN)
deep neural network (DNN)

ABSTRACT

In China, over 10% of cultivated land is polluted by heavy metals, which can affect crop growth, food safety and human health. Therefore, how to effectively and quickly detect soil heavy metal pollution has become a critical issue. This study provides a novel data preprocessing method that can extract vital information from soil hyperspectra and uses different classification algorithms to detect levels of heavy metal contamination in soil. In this experiment, 160 soil samples from the Eastern Junggar Coalfield in Xinjiang were employed for verification, including 143 noncontaminated samples and 17 contaminated soil samples. Because the concentration of chromium in the soil exists in trace amounts, combined with the fact that spectral characteristics are easily influenced by other types of impurity in the soil, the evaluation of chromium concentrations in the soil through hyperspectral analysis is not satisfactory. To avoid this phenomenon, the pretreatment method of this experiment includes a combination of second derivative and data enhancement (DA) approaches. Then, support vector machine (SVM), k-nearest neighbour (KNN) and deep neural network (DNN) algorithms are used to create the discriminant models. The accuracies of the DA-SVM, DA-KNN and DA-DNN models were 95.61%, 95.62% and 96.25%, respectively. The results of this experiment demonstrate that soil hyperspectral technology combined with deep learning can be used to instantly monitor soil chromium pollution levels on a large scale. This research can be used for the management of polluted areas and agricultural insurance applications.

1. Introduction

With the development of global industrialization, the amount of waste gas, wastewater and residue discharged from industrial production are increasing, which has resulted in increasingly serious heavy metal contamination in soil. Once the concentration of heavy metals in the ground is exceeded, it will have a severe negative impact on the growth of local crops and indirectly cause irreversible damage to human health [18]. Chromium is a heavy metal element responsible for environmental pollution. There are two stable states in the soil: chromium (III) and chromium (VI). Chromium (III) has few side effects on plants. In contrast, chromium (VI) is a carcinogen that is corrosive and carcinogenic. The pollution of chromium (VI) mainly comes from two aspects: 1) chromite ore mining and 2) industrial processing, such as leather processing. Therefore, the rapid and accurate detection of industrial chromium pollution in soil, aimed at achieving effective monitoring, has become a concern of environmental scientists worldwide.

Müller defined seven Geoaccumulation Index (Igeo) classes to determine whether heavy metal contamination exists [10]. According to Igeo values, pollution is divided into seven levels. When Igeo is less than 0, the pollution level is 0, indicating that there is no pollution; when Igeo ranges from 0 to 7 (not including the specific cost 7), the pollution level is between 1 and 6, implying that there is pollution. The hyperspectral remote sensing technology developed

*Corresponding author
a123penny@163.com (Y. Wang)
ORCID(s):
[1]These authors contributed equally to this work and should be considered co-first authors.

in the 1980s can predict the concentration of heavy metals in soil through its high spectral resolution and continuous spectral bands to achieve rapid noncontact measurements. The principle is to use the fact that soil organic matter, clay minerals, iron and manganese oxides and other main components adsorb heavy metals reflected on the spectral curve to invert the concentration of different heavy metals in the soil. Hang Cheng studied the relationship between spectral information and soil concentration and used the Partial least squares regression model to predict the concentration of Cr, As and Cd with good results, but did not achieve a good result when predicting the concentration of Pb, Cu and Zn; this research further found that the concentration of Cr has a great correlation with SOM, while Cr has a greater relationship with As, Fe, Pb, Cu, and Zn has weak relationship with SOM and Fe, so the prediction of Zn concentration was not good.The prediction of Cr concentrations had good results in this experiment, but its accuracy was still poor compared with the prediction of organic matter, which is a nontrace element [5]. Yongsheng Zhang had a more comprehensive approach to the prediction of heavy metal concentrations. He used spectral information, auxiliary spectral information (SOM, Fe, pH) and combined two kinds of information to predict the concentration of Cd. This demonstrated that the effect of using two types of spectral information is better than using any known information alone to make predictions. Although the strategy of using such information can improve the prediction of heavy metal concentrations, there are many factors that affect heavy metal spectral information, and it is necessary to find a comprehensive collection of information that can effectively improve the forecasting of heavy metal concentrations [13]. Weihong Zhou did not directly measure heavy metal concentrations by using the hyperspectra of soil but rather used rice leaf hyperspectra to estimate CaCl2-extractable concentrations of heavy metals. The study found that the 480 nm band in the original band has a high correlation with Cd, with a correlation coefficient of 0.761, but the correlation accuracy is not ideal when using the Partial least squares regression model model to estimate the Cd concentration through rice leaf hyperspectra [38]. As the composition of soil is very complex, many factors will interfere with the spectral characteristics of the Earth's surface [31].Therefore, the accuracy of directly analysing the concentration of heavy metals in the soil using spectral analysis methods is not satisfactory [1]. For example, the value of R2 associated with the Partial least squares regression model method is not sufficient when scholars attempt to describe the relationship between the concentration of chromium and its spectroscopic characteristics [33, 34]. However, deep learning can accurately extract nonlinear information and perform well in one-dimensional speech signals and two-dimensional image signals, which makes it possible to accurately identify one-dimensional hyperspectral and one-dimensional information to reach the application level. In addition, due to the advantages of rapid, nondestructive, and large-area detection of hyperspectral signals, they have a wide range of application prospects in soil pollution monitoring.

Compared with traditional methods to evaluate the concentration of heavy metals in soils, deep learning is an algorithm-based processing technique that uses artificial neural networks as an architecture with which perform representation learning on data. Deep learning is taken into consideration to address heavy metal pollution because it has been shown to be an excellent method in many fields. In 2006, Geoffrey Hinton proposed a solution to the problem of gradient disappearance in deep network training and thus contributed to the prosperity of deep learning in academia and industry applications [12]. In 2012, DNN technology achieved outstanding results in the field of image recognition, reducing the error rate from 26% to 15% in ImageNet evaluations [11]. Since the spectrum can be regarded as a one-dimensional signal and the picture can be recognized as a two-dimensional signal, the task of knowing the extent of heavy metal pollution can be addressed with deep learning. Meanwhile, DNN was also applied to the DrugeActivity prediction problem of pharmaceutical companies and achieved optimal results. To date, there have been several deep learning frameworks, such as deep neural networks, convolutional neural networks, deep belief networks, and recurrent neural networks, which have been used in computer visioning [9], natural language processing [23], and bioinformatics [19] and have achieved impressive results. Ting Liu proposed a deep autoencoder to assess its performance in near-infrared spectroscopy acquired from different categories of cigarettes. The results of experiments showed that the deep autoencoder model can sufficiently sort through different categories of cigarettes [20]. In addition, Xiaolei Zhang found that convolutional neural network model-based grapevine classification analyses for near infrared spectral data achieve a higher classification accuracy than partial least squares regression–linear discriminant analysis and principal component analysis-logistic regression model methods [36]. Deep convolutional neural networks not only perform well on near infrared spectroscopy but also works well when applied to Raman spectroscopy when facing classification problems. Deep convolutional neural networks using spectral and spatial information to classify bacterial colony species have a better effect than spatial bagging and conventional morphology methods [16]. Furthermore, Alberto Signoroni trained a convolutional neural network to identify bacterial species in hyperspectral images and indicated that deep learning models can be applied to deal with several types of spectral data effectively [30]. As numerous data types are exposed to increasingly better deep learning models with advisable effects, they lay a solid foundation

for future links between the detection of heavy metal concentrations in soil and deep learning. Data enhancement technologies that include a series of techniques used to generate new training samples is an effective means to improve deep learning models. These techniques are implemented by applying random jitter and disturbance methods to the original data without changing the class labels [29]. At present, data enhancement technology has been successfully applied in computer visioning, speech recognising, natural language processing and other fields. It is reasonable to use the data enhancement method for soil hyperspectral applications to solve sample limitation problems, which can trigger insensible results.

In this study, we use soil hyperspectra with machine learning methods to monitor whether the Cr concentration in the soil exceeds the standard. First, the second-order differential of the data set is used for preliminary feature extraction. Then, the dataset is expanded using data enhancement techniques to enhance the generalization ability and robustness of the model. Then, the deep neural network (DNN), support vector machine (SVM), and k-nearest neighbour (KNN) methods are used to perform discriminant analysis on the critical features of spectral data extraction.

The purpose of this study is (1) to explore the feasibility of employing a combination of soil hyperspectral and machine learning methods to detect Cr concentrations in the soil; (2) monitor situations in which Cr exceeds a large area accurately and quickly; and (3) provide a new technology for real-time monitoring and a sufficient basis for governance methods and processes. To the best of our knowledge, this is the first time that spectral data augmentation technologies and neural networks have been integrated to classify heavy metal pollution. Once the soil spectral library is created, the proposed method can provide reliable and immediate information on targeted soils, which is applied to agricultural insurance and environmental protection applications. In our experiment, aimed at addressing the ecological pollution caused by the development and utilization of coal resources, we choose the Eastern Junggar Coalfield as the research area to reveal the trends of soil chromium and decay in mining areas and discuss the feasibility of realizing the quantitative estimation of soil chromium concentrations using hyperspectral remote sensing technology in mining areas.

2. Materials and methods

2.1. Study area and data collection

The Eastern Junggar Coalfield is located in the northern foothills of the Tianshan Mountains and southeast of the Junggar Basin between 88°45′ ∼ 90°20′E, 44°30′ ∼ 45°00′N (Fig. 1) [37]. This area is located in the hinterland of Eurasia and has a typical extreme arid continental climate. Xinjiang's coal forecast resources account for more than 40% of the country's total forecast resources, which rank first in the country. The coal reserves in the Eastern Junggar region reach 390 billion tons, accounting for 17.8% of the total reserves of Xinjiang. Most of its areas are located in the Gobi Desert and surrounding deserts, which have no good ecological foundation. Furthermore, as the mining years continue to increase, soil heavy metal pollution tends to be more serious, resulting in irreversible damage to the ecosystem. Therefore, heavy metal pollution monitoring needs to be rapid and accurate.

Before the soil samples were brought back to the lab, soil sampling occurred from June 22 to July 1, 2014. According to the characteristics of the Eastern Junggar open-pit coal mine, the soil surrounding pits, power plants, ore dressing plants and their surroundings were collected, and GPS was used to coordinate the location of the soil collection points. Five samples were collected in mining areas, reclamation areas, deserts, etc. Using a five-point mixed sampling method to collect soil samples, each sample point was divided into three layers of a collection set with depths of 0-10 cm, 10-20 cm, 20-30 cm; a total of 168 samples were collected, mixed evenly and brought back to the laboratory with sample packages.

The soil samples brought back to the laboratory were air-dried, ground, passed through a 0.2 nm pore size sieve, sealed and stored. After processing, the soil samples were divided into three parts, which were used to determine the soil chromium concentration, organic matter concentration, and reflectance of each band of the soil hyperspectrum. The concentration of Cr was determined by the Physical and Chemical Testing Center of Xinjiang University. The samples were dissolved in hydrochloric acid, nitric acid and hydrofluoric acid. After steaming to near dryness, they were dissolved by heating with 5% hydrochloric acid and then high-purity water was added until the samples reached 20 ml. Flame atomic absorption spectrometry was used for detection. The organic matter concentration was measured by a Hitachi Z-2000 atomic absorption spectrophotometer [35].

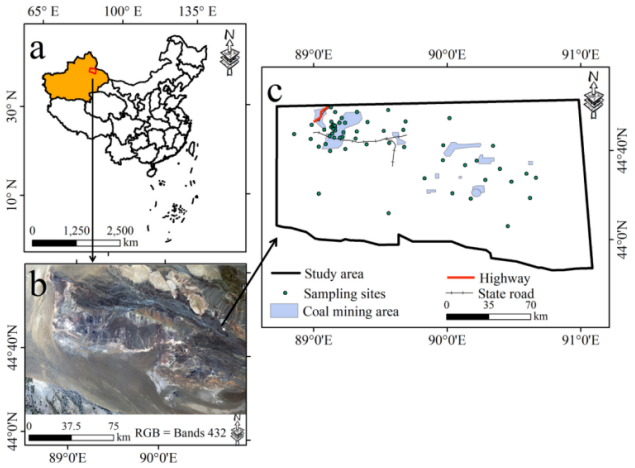


Figure 1: Study areas and location of sampling points.

2.2. Spectral measurement and pretreatment

In this study, the instrument used for obtaining soil sample spectral measurements was an ASD FildSpec3 Spectrometer (Analytical Spectral Devices, Inc., Boulder, CO, USA), and measurements of the soil spectrum were carried out in a dark room. The halogen light source used was 50 W, and it was kept 0.5 m away from the sample at a 30° incident angle; the distance between the probe perpendicular to the soil sample and the sample surface was 10 cm. A whiteboard was used as the standard reference board for diffuse reflection, which was used for calibration in time during measurement. Each sample was measured ten times, and the arithmetic average of 10 spectral values was taken as the actual spectral reflectance of the soil samples.

Due to the correct evaluation of the experimental results of spectral noise, edge bands with low signal-to-noise ratios were removed; the two removed bands were 350-400 nm and 2401-2500 nm. Finally, the reflection spectrum data of the 401-2400 nm bands were used for spectral analysis and research. All the spectral curves were smoothed and denoised using Origin8.0 software to reduce the error of the spectral data during the acquisition process. The convolution smoothing (Savitzky-Golay) method was used in this study [25]. Subsequently, second-derivative pretreatment was employed to extract practical information [15]. The spectra were finally used for resampling, and the resampling interval was 1 nm. In this study, a personal computer with a 2.8 GHz Intel Core i7-4900 MQ, 32 GB RAM, and a Windows 10 operating system was used to conduct the pretreatment. We spent less than 20 minutes finishing the pretreatment, which saves more time than traditional methods.

2.3. Data enhancement

Deep learning can learn more fine-grained knowledge from complex data and effectively filter and combine features with different importance levels, which can yield better performance in the classifier. A spectrum can be viewed as a one-dimensional image, and deep neural networks have been used to classify pharmaceutical tablets and Raman spectra. Image data can be expanded by rotation and other methods [29]. In this experiment, the spectrum enhancement method is primarily used. Compared to the measured value of the simulated spectrum under different illuminations, the model introduces more noise data, discards less valid features, and extracts useful features to increase robustness.

Because traditional models exhibit certain shortcomings in detecting trace elements of heavy metals, this study uses a combination of data augmentation and deep neural network models to achieve more accurate results. There are three ways to enhance data: increase the offset; multiplication; and random changes in slope. In this study, the spectrum of the measured data set is randomly and uniformly multiplied by 0.999-1.001 [3]. This operation is completed ten times to enhance different illumination levels and disturb the data to yield more robust model training.

2.4. Correction strategy

Via multiple divisions, cross-validation significantly reduces the contingency result caused by a random division. Concurrently, via various divisions and many trainings, the model can also encounter a variety of data, thereby im-

proving its generalizability. In the k-fold cross-validation method, it is common to set k to 5 or 10. In the experiment, the entire extended data set (1680) is randomly divided into a training set and a verification set at a ratio of 8:2, and is continuously used for model training; this process is referred to as the fivefold cross-validation method. Based on the criteria of Igeo, the parameter is greater than 73 for pollution and below or equal to 73 for nonpollution. A support vector machine (SVM) or k-nearest neighbour (KNN) method is commonly used for classification due to their mature theories and high accuracies. Deep learning has excellent performance in managing complex problems; thus, this experiment uses SVM, KNN and deep learning for classification.

In this study, SVM is modelled in Python 3.7.4 with Scikit-learn modules, KNN is modelled in Python 3.7.4 with Scikit-learn modules, DNN is modelled in Python 3.7.4 with TensorFlow, Keras, Scikit-learn, and NumPy third-party modules.

2.4.1. SVM

SVM is a supervised, kernel-based nonlinear learning method [32] that uses nuclear techniques to manage complex nonlinear problems with good performance. Using kernel functions, SVM maps a data set to a higher-dimensional feature space than the original space, making the samples linearly separable in the new feature space. The final decision function of SVM is only determined by a few support vectors. The complexity of the calculation depends on the number of support vectors. SVM can extract key sample information and remove many redundant samples; thus, the algorithm is simple and efficient.

2.4.2. KNN

KNN is a supervised learning algorithm that can describe the nonlinear relationship between metal concentrations and spectral data, and it is widely used to solve classification problems [8]. KNNs work as follows: (1) given test samples, a specific distance evaluation method is used to determine the k samples closest to them; and (2) perform prediction classifications based on these k samples. The KNN results strongly depend on the choice of k.

2.4.3. DNN

Deep neural networks (DNNs) are discriminant models. A neural network with at least one hidden layer can be trained using a back-propagation algorithm. A weight update can be solved using the following gradient descent method. By training multilayer neural networks, high-dimensional data can be converted into low-dimensional codes to solve nonlinear problems. The traditional neural network learning algorithm exhibits shortcomings such as overfitting and falling into local optima, which makes the accuracy of classification and regression tasks unsatisfactory. Hinton proposed the "dropout" algorithm, which can randomly ignore half of the feature detectors during training and can prevent overfitting caused by a small training set. In handwriting font recognition, speech recognition and other data sets, DNNs achieve better performance than other methods [11]. Alex used deep convolutional neural networks in 2012 and won the ILSVRC-2012 competition. Since then, deep learning has become a popular research technology in various fields of academia.

2.5. Prediction accuracy

In the classification experiment in this study, the related measures of accuracy, sensitivity, and specificity are used to evaluate the results. In all samples, they are divided into positive examples and negative examples. Among the positive examples, correct predictions are referred to as true positives (TPs), and incorrect predictions are referred to as false negatives (FNs). In the negative examples, when the prediction is incorrect, the results are referred to as false positives (FPs), and correct predictions are referred to as true negatives (TNs). The accuracy rate is defined as the number of divided samples divided by the number of all samples. We also use sensitivity and specificity to measure and analyse the spectrum, respectively. The formulations of accuracy, sensitivity and specificity are described as follows:

$$\mathrm{ACC} = (\mathrm{TP} + \mathrm{TN})//(\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}) \tag{1}$$

$$\mathrm{Sensitivity} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \tag{2}$$

$$\mathrm{Specificity} = \mathrm{TN}//(\mathrm{TN} + \mathrm{FP}) \tag{3}$$

3. Results

3.1. Soil spectral analysis

The soil spectrum with Cr concentrations in the range of 401 nm to 2400 nm was compared between soil samples with and without standard concentrations. It is clearly shown that the two sets of original spectra are difficult to distinguish by their shape and intensity (Fig. 2). Therefore, a powerful preprocessing method and a data analysis algorithm are necessary to extract key information and distinguish between polluted and nonpolluted soil spectra. After smoothing and second derivative treatment, apparent peak changes were observed at 562-575 nm, 969-1008 nm, 1402-1512 nm, 1751-1769 nm, and 2252-2391 nm. However, it is not easy to distinguish differences from other bands. Due to the complex spectral features mentioned above, we propose a deep learning model that can detect vital features and other insignificant features from multiple layers.

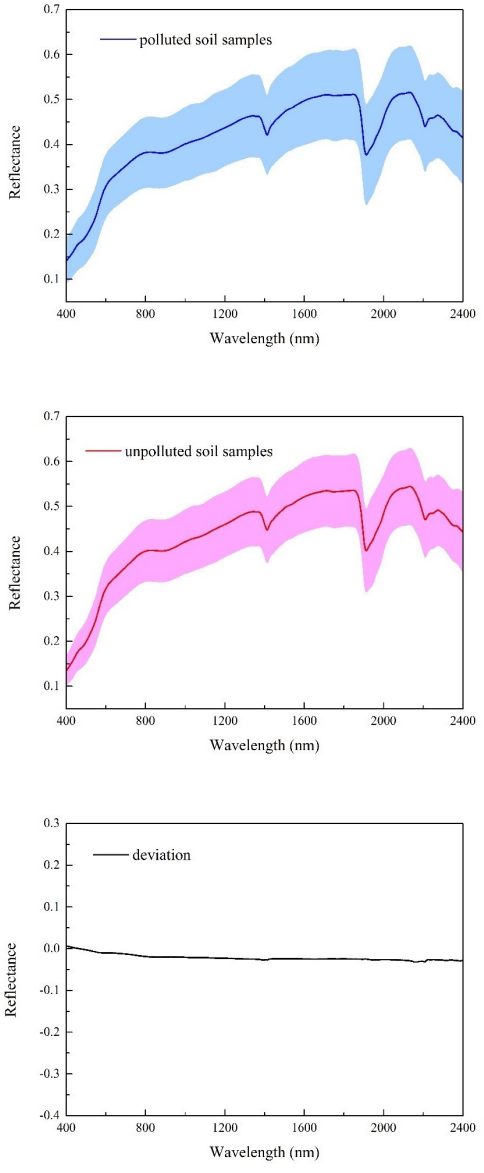


Figure 2: Comparison of the normalized mean spectra between polluted soil samples and unpolluted soil samples (The shaded areas represent the standard deviations of the means).

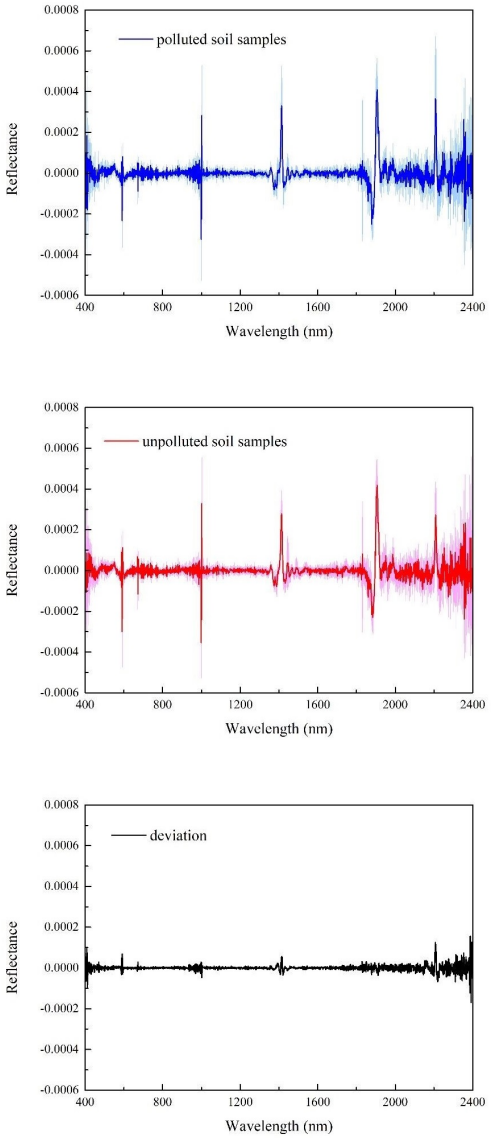


Figure 3: Comparison of the normalized mean spectra that are preprocessing between polluted soil samples and unpolluted soil samples (The shaded areas represent the standard deviations of the means).

3.2. Model evaluation

In the SVM model, the linear kernel, polykernel, and rbf kernel were selected to compare the classification results. For the experiment, experimental parameters and results are shown in Tables 1 and 2, respectively. The judgement accuracy of the SVM model varies with the kernel function; the product of the linear kernel is the worst, and that of the polykernel is the best. The linear kernel makes the characteristics of the map too simple to express the complex spectral shape; thus, the classification result is the worst. Additionally, extreme cases showed that the prediction result of the positive example is 1, and the negative example result is 0, which shows that linear classification is not applicable to the spectral classification of trace elements in complex situations. The development of the rbf function provides an adequate level of sorting, the value of sensitivity is low (47.05%), and the value of specificity (83.44%) is high. The correct rate of pollutants is low; however, the incorrect rate of pollutants is high. These results are caused by the operational nature of the Gaussian kernel, which can map the low-dimensional input space onto the high-dimensional feature space, thus making it easy to learn redundant information when there are fewer pollutant samples and more features. The performance of the polykernel in the experiment is the best, indicating that spectral classification is more

Table 1
Parameter initialization setting.

SVM parameter setting	
C	40
gamma	20
class_weight	balanced
max_iter	2000
verbose	True

Table 2
Comparison of performance for linear kernel,rbf kernel,poly kernel.

Kernel	Sensitivity(%)	Specificity(%)	Accuracy(%)
linear	1.00	0.00	10.62
rbf	47.05	83.44	79.37
poly	82.35	97.20	95.61

suitable for polynomial fitting and will produce good results. The polynomial kernel increases the feature dimension to a more moderate dimension, which can effectively identify and mitigate information explosion, and can use curves to divide trace elements accurately. When the training data are linearly inseparable, a nonlinear SVM is learned using kernel techniques and soft interval maximization. When the input space is the Hilbert space, the kernel function represents the inner product between the feature vectors obtained by mapping the input data from the input space to the feature space. Using the kernel function, nonlinear SVMs can be learned, which is equivalent to implicitly learning linear SVMs in a high-dimensional feature space. The definition of the kernel function is expressed as equations (4) and (5) [4, 17].

Let χ be the input space and H be the feature space if there is a mapping from χ to H :

$$\Phi(x) : \chi \rightarrow H \tag{4}$$

Therefore, for all $x, z \in X$, the function $K(x, z)$ satisfies the condition:

$$K(x, z) = \Phi(x) \cdot \Phi(z) \tag{5}$$

Then, $K(x, z)$ is the kernel function, where $\Phi(x)$ in equation (4) is the mapping function, and $\cdot$ indicates the inner product multiplication.

Solving the optimization problem with SVMs is equivalent to identifying the maximum separation hyperplane. To obtain an optimal solution, the dual problem can be solved using Lagrangian duality and thus the solution's linearly separable support vector. The optimal solution of the machine is equivalent to solving the objective function of equation (6). When solving the optimal problem of an SVM including a kernel function, the dual objective function to be solved is described by equation (7):

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \left(x_i \cdot x_j\right) - \sum_{i=1}^{N} \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^{N} \alpha_i y_i = 0 \\ & 0 \leqslant \alpha_i \leqslant C, \quad i = 1, 2, \cdots, N \end{aligned} \tag{6}$$

$$W(\alpha) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K\left(x_i, x_j\right) - \sum_{i=1}^{N} \alpha_i \tag{7}$$

In equations (6) and (7), a, a_i, and a_j are Lagrangian multiplier vectors; and (x_i, y_i) belongs to the training data set $T = \{(x_1, y_1), (x_2, y_2) \ldots, (x_n\ y_n)\}$, where $x_i \in R_n, y_i \in \{-1, +1\}, i = 1, 2, \ldots, N$.

In the experiment, a polynomial kernel is used, and the input space corresponds to a feature space via nonlinear transformation so that the hypersurface model in the input space corresponds to the hyperplane model in the feature space H. We extend the original hyperspectral features to higher dimensions using the polynomial kernel function to learn more abundant Cr features. Concurrently, to avoid using the kernel function to expand to infinite dimensions, the highest order term of the polynomial kernel function is set equal to 3 in this experiment to avoid learning redundant Cr features. In the real operation in Python, the inner product $x_i \cdot x_j$ of the input space is transformed into the inner product $\Phi(x_i) \cdot \Phi(x_j)$ in the feature space, and $\Phi(x_i) \cdot \Phi(x_j)$ is calculated to obtain $K(x_i, x_j)$, which is more complex; thus, the toolkit in the software can be directly applied to the matrix to calculate $K(x_i, x_j)$. Nonlinear Cr features can be learned quickly and easily, and the discriminant function can be used to accurately judge whether Cr exceeds the standard.

Next, KNN considers all selected spectra to identify the most distinctive variables and perform classification [8]. In this study, the Euclidean distance is used as a distance measurement, and the model is evaluated by adjusting k in the model. Typically, it is assumed that a large k selects neighbouring points; thus, training examples that are far away from the input instance and are not similar will also contribute to the prediction, making predictions incorrect. However, if k is small, the prediction will be more sensitive to nearby examples, which will lead to noise and prediction errors. The best k is determined using the fivefold cross-validation method. Based on Figure 4, when k was set to 16, the best result was produced, and the accuracy, sensitivity and specificity were 95.62%, 70.58% and 98.60%, respectively (Table 3).

Table 3
Comparison of performance for different k values.

K value	Sensitivity(%)	Specificity(%)	Accuracy(%)
12	47.05	97.20	91.87
14	70.58	98.60	95.62
16	35.29	97.90	91.25
18	35.29	97.20	90.62
20	5.88	1.00	89.99

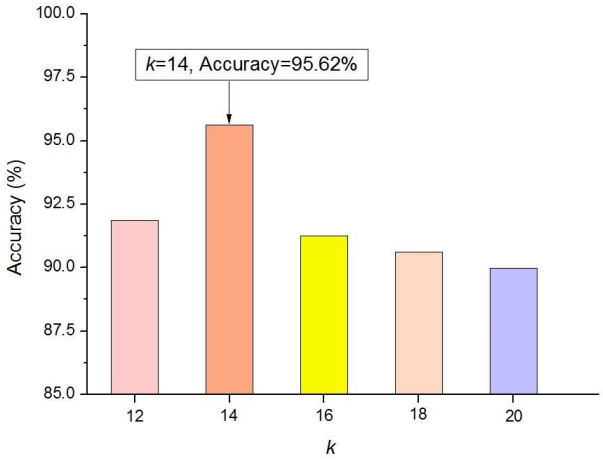


Figure 4: Accuracy of KNN algorithms under different k values.

To evaluate the performance of the deep learning model, we compared its use with three deep neural network models (Figures 5-7). The primary difference between these three models is the model depth. The first DA-DNN model

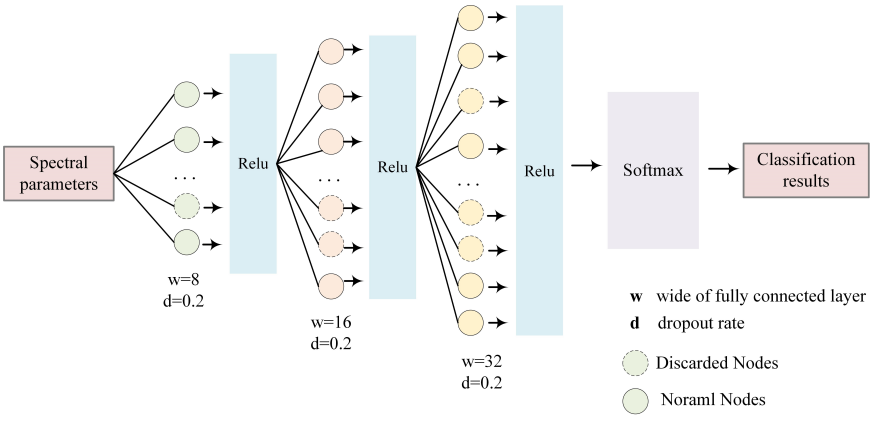


Figure 5: Network structure diagrams of three-layer model (model 1).

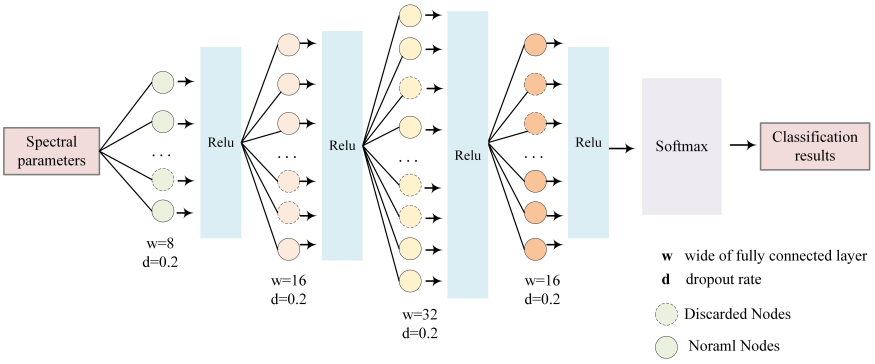


Figure 6: Network structure diagrams of four-layer model (model 2).

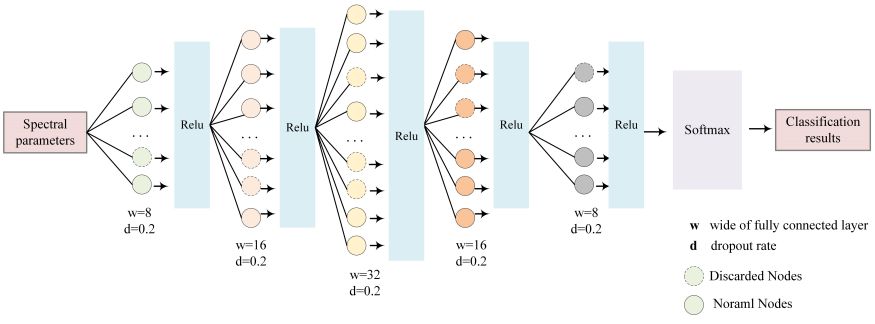


Figure 7: Network structure diagrams of five-layer model (model 3).

(Model 1) is the shallowest model with three hidden layers. Every hidden layer is equipped with a "relu" activation function and "dropout" regularization terms to optimize the network structure, in which the activation function can learn the nonlinear characteristics of the spectrum. The role of dropout is to prevent overfitting. The number of neurons in the first model's three hidden layers is 8, 16, and 32. There is also an input layer (F1) and an output layer in the model. The second DNN model (Model 2) includes three hidden layers, an input layer and an output layer. Compared to the first model, a hidden layer with 16 neurons is added, and a "relu" activation function and a dropout regularization term are also present to learn various features of the data by deepening the neural network. The third DNN model (Model 3) includes five hidden layers, an input layer and an output layer. Apart from the structure of the first model, a hidden layer with 16 neurons and a hidden layer with eight neurons are added, and each hidden layer

is followed by a "relu" activation function and a dropout regularization term. The accuracy of the three-layer neural network is found to be the highest of the tested models. As the number of network layers increases, the accuracy rate decreases and becomes stable, demonstrating that a deeper network makes the model tend towards overfitting because the deeper network has learned too much useless information (Table 4), thus decreasing the accuracy rate.

Table 4
Comparison of performance for DNN models with different layers.

DNN model	Sensitivity(%)	Specificity(%)	Accuracy(%)
Model 1	88.23	97.20	96.25
Model 2	0.00	1.00	89.37
Model 3	0.00	1.00	89.37

4. Discussion

In this paper, we study the preprocessing and statistical learning methods of analysing soil hyperspectra combined with second-order differential and data enhancement and a deep learning method for the screening of excessive Cr concentrations. Through the second derivative processing and analysis of hyperspectral characteristics in the polluted and nonpolluted groups, we found that the soil hyperspectra between the two groups had different specificities. Due to the differences in Cr concentrations, the absorption peak position and peak intensity of the two groups of soil spectra change. These changes indicate that individual components in contaminated and noncontaminated soil change with varied Cr concentrations. Based on the situation, using near-infrared spectroscopy coupled with deep learning methods and statistical methods may be able to determine whether the Cr concentration in the soil exceeds the standard.

From the literature, it is known that near-infrared spectroscopy can reflect many vibration modes of considerable substances, such as iron oxides, clay minerals and organic matter components, and their composition or quantity may be related to the concentration of Cr. Iron oxides have a relatively large influence on the concentration of Cr [39]. Furthermore, the corresponding prominent bands are shown in Table 5. In addition, clay minerals and organic matter also have varying degrees of influence on heavy metals. Table 5 shows the corresponding bands of the primary substance.

Due to the high dimensionality of hyperspectral data, processing hyperspectral data will cause problems, such as requiring a large number of calculations and producing results with low accuracy. Traditional spectroscopy studies need to exclude useless variables to prevent collinearity in data processing from affecting the final prediction results. Some studies use PCA to extract the most useful hyperspectral information. This method shows good results in situations with fewer impurities. However, there will be many impurities that affect the results of trace elements in soil hyperspectra, which produce undesirable effects. Choosing appropriate bands from all available bands to obtain useful information requires expert knowledge and rich practical experience, and it is not easy to achieve automated and rapid detection. SVM and KNN are often used for quantitative analysis of spectra, but SVM easily learns too many features when processing hyperspectral data, which leads to the occurrence of overfitting. KNN is computationally expensive, and when the sample size is small, the classification accuracy is not very high.

Deep learning is a powerful data analysis algorithm that is driven by big data. The primary learning method of deep learning is to establish a connection between the input, hidden and output layers through a neural network and select an activation function for each hidden layer to simulate a complex nonlinear process. Because deep learning has been applied to a variety of complex data structures, such as image signals and voice signals, the experiment in this study attempts to process multidimensional soil hyperspectra with deep learning. Because the number of experimental hyperspectral data is insufficient, a deep neural network model with more parameters will not have good generalizability or robustness. To overcome this problem, this experiment uses data enhancement technology that is commonly used in deep learning. Based on its spectral characteristics, the spectrum is appropriately disturbed to expand the dataset. With data enhancement technology and deep learning, hyperspectral data can yield good results with regard to predicting whether the heavy metal Cr concentration exceeds the standard. Combined with data enhancement, SVM and KNN achieve satisfactory performance when detecting excessive Cr in soil.

This experiment also shows that when the same hyperspectral data undergoes preprocessing, deep learning achieves the highest accuracy when monitoring hyperspectral heavy-metal pollution in soil due to its strong fit. In the proposed experiment, three marked absorption zones (1413, 1922, and 2200 nm) were found in the spectrum. Wavelengths of

Table 5
Soil absorbance in the visible-near-infrared regions

Soil constituent	Wavelength	Reference
Fe oxides		
goethite	420	Sherman and Waite(1985) [27]
	427	Scheinost et al(1998) [26]
	434	Rossel et al(2010) [24]
	480	Sherman and Waite(1985) [27]
	650	Rossel et al(2010) [24]
	920	Sherman and Waite(1985) [27]
haematite	404	Rossel et al(2010) [24]
	444	Rossel et al(2010) [24]
	529	Rossel et al(2010) [24]
	650	Rossel et al(2010) [24]
	884	Rossel et al(2010) [24]
	510	Sherman and Waite(1985) [27]
	531	Scheinost et al(1998) [26]
water	940	Hunt(1977) [14]
	1135	Hunt(1977) [14]
	1380	Hunt(1977) [14]
	1455	Hunt(1977) [14]
	1915	Rossel et al(2010) [24]
hydroxyl	700	Rossel et al(2010) [24]
	930	Rossel et al(2010) [24]
	1400	Hunt(1977) [14]
	1929	Ben-Dor et al(1997) [2]
	1932	Ben-Dor et al(1997) [2]
	2200	Clark et al.(1990) [7],Post and Noble(1993) [22]
Clay minerals		
Kaolin doublet	1395	Oinuma and Hayashi(1965) [21]
smectite	2206	Oinuma and Hayashi(1965) [21]
illite	2340	Post and Noble(1993) [22]
carbonate	2336	Rossel et al(2010) [24]
organics		
aromatics	1650	Clark et al(1990) [7], Clark(1999) [6]
amine	1000	Clark et al(1990)[7],Clark(1999) [6]
	2060	Hunt(1977) [14]
Alkyl asymmetric-symmetric doublet	853	Clark et al(1990) [7],Clark(1999) [6]
amides	2033	Clark et al(1990)[7],Clark(1999) [6]
methyls	2310	Shonk et al(1991) [28]

1400 and 1900 nm were primarily water absorption bands, which belong to the first-order frequency doubling zone. The O-H lattice structure water of clay minerals and soil-adsorbed water has a greater impact at 1400 and 1900 nm. There are certain wave bands that have a strong correlation with Cr. For example, the high-signal vibration at 432 nm corresponding to goethite (a-FeOOH) shows that the Cr concentration in the soil has a strong correlation with the increase in the a-FeOOH concentration. Absorptions near 1004, 853, 2035 and 2310 nm in the spectra are related to the vibration of organic matter. Goethite in nature must contain organic matter such as amides, and these substances cause the redox reaction of Cr in the soil. For example, Cr^{3+} can be oxidized on the surface of true iron ore; NO_3^- has a reducing effect and can oxidize Cr^{3+} to Cr^{6+} on the surface of goethite. In deep learning, neural networks can extract more abstract and deeper information in hyperspectra to improve data mining. In the heavy metal pollution and non-pollution of the soil, each hidden layer can extract different degrees of information to make different combinations: a hidden layer extracts the smaller wavelength range from 401 to 2400 nm for identification, such as the influence of 1004, 853 and 2035 nm on Cr concentration; another hidden layer is analysed for a broader range of wavelengths, such as features around three pronounced absorption peaks at 1413, 1922, and 2200 nm; and the third hidden layer

can be combined with the results of the first and second layers for waveband discrimination. Because heavy-metal elements in the soil spectra are strongly affected by other substances, a dropout layer is added after each layer of the neural network in this experiment, and certain information is randomly discarded to prevent the model from overfitting. In this experiment, more preprocessing methods have not been tried to affect the deep learning model, and the deep learning model has not designed a more complex network structure to compare model performance. The next step will explore how different preprocessing methods and different complexities of the model influence deep learning effects.

5. Result

In this study, we proved that soil near-infrared spectroscopy combined with deep learning methods or statistical learning methods is used for screening Cr pollution. Due to the low concentration of Cr in the soil, the amount of useful information in the band after the second derivative treatment increased. In contrast, a small number of bands and the Cr concentration show strong correlations. The experiment proves that Fe oxides and organic matter have a significant effect on the concentration of Cr in some specific wavebands, and the features of each level can be abstractly extracted through deep learning to achieve valid recognition. The high classification accuracy of DA-SVM, DA-DNN and DA-KNN illustrates the potential advantages of using near-infrared hyperspectral spectroscopy to identify heavy metal pollution.

The main goal of Cr concentration detection is to identify pollution in a timely manner to avoid heavy metal pollution in the environment and prevent irreversible harm to the human body caused by Cr absorption through soil and water. High Cr concentrations in different areas should mandate different pollution control strategies, which can be achieved by the detection and treatment of high Cr pollution areas based on different spectral characteristics. In recent years, with the launch of hyperspectral satellites and development of portable hyperspectral instruments equipped with unmanned aerial vehicles, this research is expected to provide large-scale accurate, real-time monitoring and guidance for areas with heavy metals exceeding the standard.

References

[1] Baveye, P.C., Laba, M., 2015. Visible and near-infrared reflectance spectroscopy is of limited practical use to monitor soil contamination by heavy metals. Journal of Hazardous Materials 285, 137–139. URL: https://www.sciencedirect.com/science/article/pii/S0304389414009650, doi:https://doi.org/10.1016/j.jhazmat.2014.11.043.

[2] Ben-Dor, E., Inbar, Y., Chen, Y., 1997. The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400–2500 nm) during a controlled decomposition process. Remote Sensing of Environment 61, 1–15. URL: https://www.sciencedirect.com/science/article/pii/S0034425796001204, doi:https://doi.org/10.1016/S0034-4257(96)00120-4.

[3] Bjerrum, E., Glahder, M., Skov, T., 2017. Data augmentation of spectral data for convolutional neural network (cnn) based deep chemometrics. In: arXiv. 1710.01927 .

[4] Chen, Q., Zhao, J., Fang, C., Wang, D., 2007. Feasibility study on identification of green, black and oolong teas using near-infrared reflectance spectroscopy based on support vector machine (svm). Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 66, 568–574. URL: https://www.sciencedirect.com/science/article/pii/S1386142506002216, doi:https://doi.org/10.1016/j.saa.2006.03.038.

[5] Cheng, H., Shen, R., Chen, Y., Wan, Q., Shi, T., Wang, J., Wan, Y., Hong, Y., Li, X., 2019. Estimating heavy metal concentrations in suburban soils with reflectance spectroscopy. Geoderma 336, 59–67. URL: https://www.sciencedirect.com/science/article/pii/S0016706117319651, doi:https://doi.org/10.1016/j.geoderma.2018.08.010.

[6] Clark, R.N., 1999. Spectroscopy of rocks and minerals, and principles of spectroscopy. Remote Sensing for the Earth Sciences Manual of Remote Sensing 3, 3–52.

[7] Clark, R.N., King, T.V.V., Klejwa, M., Swayze, G.A., Vergo, N., 1990. High spectral resolution reflectance spectroscopy of minerals. Journal of Geophysical Research solid earth 95, 12653–12680.

[8] Cui, X., Zhao, Z., Zhang, G., Chen, S., Zhao, Y., Lu, J., 2018. Analysis and classification of kidney stones based on raman spectroscopy. Biomedical Optics Express 9, 4175.

[9] Dong, Z., Li, G., 2020. Cascade r-cnn: Delving into high quality object detection. In CVPR .

[10] Ghrefat, H.A., Abu-Rukah, Y., Rosen, M.A., 2011. Application of geoaccumulation index and enrichment factor for assessing metal contamination in the sediments of kafrain dam, jordan. Environmental monitoring and assessment 178, 95–109.

[11] Hinton, G., Srivastave, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. In: arXiv:1207.0580 .

[12] Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313, 504–507.

[13] Hong, Y., Shen, R., Cheng, H., Chen, S., Chen, Y., Guo, L., He, J., Liu, Y., Yu, L., Liu, Y., 2019. Cadmium concentration estimation in peri-urban agricultural soils: Using reflectance spectroscopy, soil auxiliary information, or a combination of both? Geoderma 354, 113875. URL: https://www.sciencedirect.com/science/article/pii/S0016706119300849, doi:https://doi.org/10.1016/j.geoderma.2019.07.033.

[14] Hunt, G., 1977. Spectral signatures of particulate minerals in the visible and near infrared. Geophysics 42, 501–513.

[15] Jin, X., Li, S., Zhang, W., Zhu, J., Sun, J., 2020. Prediction of soil-available potassium content with visible near-infrared ray spectroscopy of different pretreatment transformations by the boosting algorithms. Applied Sciences 10, 1520.

[16] Krauss, S.D., Roy, R., Yosef, H.K., Lechtonen, T., El-Mashtoly, S.F., Gerwert, K., Mosig, A., 2018. Hierarchical deep convolutional neural networks combine spectral and spatial information for highly accurate raman-microscopy-based cytopathology. Journal of Biophotonics 11, e201800022.

[17] Li, H., 2019. Statistical learning methods. Tsinghua University Press.

[18] Li, Z., Ma, Z., van der Kuijp, T.J., Yuan, Z., Huang, L., 2014. A review of soil heavy metal pollution from mines in china: Pollution and health risk assessment. Science of The Total Environment 468-469, 843–853. URL: https://www.sciencedirect.com/science/article/pii/S0048969713010176, doi:https://doi.org/10.1016/j.scitotenv.2013.08.090.

[19] Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M., 2019. Deep learning for generic object detection: A survey. International Journal of Computer Vision 128, 261–318.

[20] Liu, T., Li, Z., Yu, C., Qin, Y., 2017. Nirs feature extraction based on deep auto-encoder neural network. Infrared Physics and Technology 87, 124–128. URL: https://www.sciencedirect.com/science/article/pii/S1350449517302268, doi:https://doi.org/10.1016/j.infrared.2017.07.015.

[21] Oinuma, K., Hayashi, H., 1965. Infrared study of mixed-layer clay minerals. American Mineralogist 50, 1213–1227.

[22] Post, J.L., Noble, P.N., 1993. The near-infrared combination band frequencies of dioctahedral smectites, micas, and illites. Clays and Clay Minerals 41, 639–644.

[23] Provilkov, I., Emelianenko, D., Voita, E., 2019. Bpe-dropout: Simple and effective subword regularization. In: arXiv:1910.13267 .

[24] Rossel, R.V., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158, 46–54. URL: https://www.sciencedirect.com/science/article/pii/S0016706109004315, doi:https://doi.org/10.1016/j.geoderma.2009.12.025.

[25] Savitzky, A., Golay, M., 1964. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry 36, 1627–1639.

[26] Scheinost, A., Chavernas, A., V.Barrón, Torrent, J., 1998. Use and limitations of second-derivative diffuse reflectance spectroscopy in the visible to near-infrared range to identify and quantify fe oxide minerals in soils. Clays and Clay Minerals 46, 528–536.

[27] Sherman, D.M., Waite, T.D., 1985. Electronic spectra of fe3+ oxides and oxyhydroxides in the near infrared to ultraviolet. American Mineralogist 70, 1262–1269.

[28] Shonk, J.L., Gaultney, L.D., Schulze, D.G., Scoyoc, G.E.V., 1991. Spectroscopic sensing of soil organic-matter content. Transaction of the Asae 34, 1978–1984.

[29] Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. Journal of Big Data 6.

[30] Signoroni, A., Savardi, M., Pezzoni, M., Guerrini, F., Arrigoni, S., Turra, G., 2018. Combining the use of cnn classification and strength-driven compression for the robust identification of bacterial species on hyperspectral culture plate images. IET Computer Vision 12, 941–949.

[31] Song, L., Jian, J., Tan, D.J., Xie, H.B., Luo, Z.F., Gao, B., 2015. Estimate of heavy metals in soil and streams using combined geochemistry and field spectroscopy in wan-sheng mining area, chongqing, china. International Journal of Applied Earth Observation and Geoinformation 34, 1–9. URL: https://www.sciencedirect.com/science/article/pii/S0303243414001469, doi:https://doi.org/10.1016/j.jag.2014.06.013.

[32] Vapnik, V., 1998. The Nature of Statistical Learning Theory. Wiley.

[33] Wu, Y., Chen, J., Ji, J., Gong, P., Liao, Q., Tian, Q., Ma, H., 2007. A mechanism study of reflflectance spectroscopy for investigating heavy metals in soils. Soil Sci. Soc. Am. J. 71, 918–926.

[34] Wu, Y., Chen, J., Wu, X., Tian, Q., Ji, J., Qin, Z., 2005. Possibilities of reflectance spectroscopy for the assessment of contaminant elements in suburban soils. Applied Geochemistry 20, 1051–1059. URL: https://www.sciencedirect.com/science/article/pii/S088329270500048X, doi:https://doi.org/10.1016/j.apgeochem.2005.01.009.

[35] Xia, J., 2014. The study of remote sensing dynamic monitoring for coalfield fire area in shuixigou, xinjiang. IOP Conference Series: Earth and Environmental Science 17, 012097.

[36] Zhang, X., Xu, J., Lin, T., Ying, Y., 2018. Convolutional neural network based classification analysis for near infrared spectroscopic sensing. in: ASABE Annu. Int. Meet., American Society of Agricultural and Biological Engineers , 1.

[37] Zhang, Z., Ding, J., Wang, J., Ge, X., 2020. Prediction of soil organic matter in northwestern china using fractional-order derivative spectroscopy and modified normalized difference indices. Catena 185, 104257. URL: https://www.sciencedirect.com/science/article/pii/S0341816219303996, doi:https://doi.org/10.1016/j.catena.2019.104257.

[38] Zhou, W., Zhang, J., Zou, M., Liu, X., Li, J., 2019. Feasibility of using rice leaves hyperspectral data to estimate cacl2-extractable concentrations of heavy metals in agricultural soil. Sentific Reports 9, 16084.

[39] Zhu, C., 2004. Study of process and mechanism of chromium adsorption on natural biomineralized goethite.