Geoderma 402 (2021) 115366

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Soil properties: Their prediction and feature extraction from the LUCAS spectral library using deep convolutional neural networks

Liang Zhong [a,b], Xi Guo [a,b,*], Zhe Xu [b,c], Meng Ding [a,b]

[a] College of Land Resources and Environment, Jiangxi Agricultural University, Nanchang 330045, China
[b] Key Laboratory of Poyang Lake Watershed Agricultural Resources and Ecology of Jiangxi Province, Nanchang 330045, China
[c] College of Forestry, Jiangxi Agricultural University, Nanchang 330045, China

ARTICLE INFO

Handling Editor: Budiman Minasny

Keywords:
Deep learning
Deep convolutional neural network
Feature wavelengths
LUCAS topsoil dataset
Soil properties
Soil spectral library

ABSTRACT

Soil, as a non-renewable resource, should be monitored continuously to prevent its degradation and promote sustainable agriculture. Soil spectroscopy in the visible-near infrared range is a fast and cost-effective analytical technique to predict soil properties. Although traditional machine learning methods are widely used for modeling soil spectral data, large spectral datasets may require better analytical methods for big data. Here, we explored the modeling potential of deep convolutional neural networks (DCNNs) for soil properties based on a large soil spectral library. The European topsoil dataset provided by the Land Use/Cover Area frame Survey (LUCAS) was used for DCNN modeling with the original absorbance spectra. Two single-task 16-layer DCNN models (LucasResNet-16 and LucasVGGNet-16) were used to make regression predictions of seven soil properties and classification predictions of soil texture. The effects of data pre-processing on single-task and multi-task DCNN modeling were assessed. The SHapley Additive exPlanations method was used to interpret the output of a DCNN model (LucasResNet-16). The DCNN models produced accurate predictions for most soil properties, and were superior to a single-task shallow convolutional neural network and traditional machine learning methods. Spectral transformation was effective for predicting some soil properties, while spectral downsampling led to a reduction in the modeling accuracy. The performance of a multi-task DCNN model built on the basis of LucasResNet-16 was improved compared with the performance of the single-task model. Soil organic carbon content, nitrogen content, cation exchange capacity, pH, and calcium carbonate content were well predicted, with the root mean squared error of 19.130 $g \cdot kg^{-1}$, 0.971 $g \cdot kg^{-1}$, 6.614 $cmol(+) \cdot kg^{-1}$, 0.326, and 24.526 $g \cdot kg^{-1}$, respectively. The overall classification accuracy of soil texture was 0.749 (four groups) and 0.566 (12 levels). The position of feature wavelengths differed among the soil properties, for which multiple characteristic peaks were common. This study fully demonstrates the modeling potential of deep learning with soil ultra-spectral data, which could enhance precision agriculture.

1. Introduction

Soil is the loose surface portion of the earth's crust that provides water and nutrients for uptake by plants. Underpinning all agricultural production, soil is the basis of crop growth, and is a natural resource for human survival (Sanchez et al., 2009). However, soil fertility varies across and within regions. If our incomplete knowledge of soil properties is not soon rectified, either excessive or insufficient fertilization will ensue during the plant-growing season. Inappropriate fertilization not only affects crop growth, but also creates potential problems, such as wasted resources, environmental pollution, and land degradation (Lal, 2004; Hartemink, 2015). A prerequisite for precision agriculture is the fast and accurate acquisition of information on soil properties and subsequent development of rational fertilization strategies.

Abbreviations: CEC, Cation exchange capacity; CNN, Convolutional neural network; DCNN, Deep convolutional neural network; 1D, One-dimensional; 2D, Two-dimensional; K, Potassium; LSTM, long short-term memory; LUCAS, Land Use/Cover Area frame statistical Survey; N, Nitrogen; OC, Organic carbon; P, Phosphorous; PLSR, Partial least squares regression; RF, Random forest; RMSE, Root mean squared error; SG0, Zero-order Savitzky-Golay filter with a window width of 50; SG1, First-order Savitzky-Golay filter with a window width of 50; SG2, Second-order Savitzky-Golay filter with a window width of 50; SHAP, SHapley Additive exPlanations; SNV, Standard normal variate; SVM, Support vector machine; Vis-NIR, Visible-near infrared.

* Corresponding author at: College of Land Resources and Environment, Jiangxi Agricultural University, Nanchang 330045, China.
E-mail address: guoxi@jxau.edu.cn (X. Guo).

https://doi.org/10.1016/j.geoderma.2021.115366
Received 5 February 2021; Received in revised form 19 July 2021; Accepted 19 July 2021

Implementing these strategies can ensure sustainable agricultural production, improve crop yield, and protect the ecological environment (Moran et al., 1997; Wetterlind et al., 2010).

Due to the high cost and low efficiency of traditional field sampling and laboratory testing methods, it is near impossible to achieve large-scale and real-time monitoring of dynamic soil properties (Araújo et al., 2014). In recent years, soil spectroscopy in the visible-near infrared (Vis-NIR) range has established itself as a fast and cost-effective alternative to traditional empirical methods for predicting soil properties (Islam et al., 2003; Guerrero et al., 2015). Many studies have proven that high-precision estimation of soil properties at local and regional scales can be achieved using the Vis-NIR spectroscopy technique (Gogé et al., 2012; Viscarra Rossel et al., 2016; Romero et al., 2018). Furthermore, numerous machine learning models have been calibrated from local soil spectral databases, resulting in independent small-scale models (Stevens et al., 2013).

Site-specific correlations exist between spectral features and soil properties, while the calibration model parameters are often region-specific. As a result, model conversion between different soil sample sets is generally unsuccessful (Grunwald et al., 2018). Therefore, recent research has increasingly used soil spectral libraries at the global (Viscarra Rossel et al., 2016), continental (Stevens et al., 2013), or national (Brodský et al., 2011; Shi et al., 2014) scale to predict soil properties. However, when using a larger regional spectral library, the prediction accuracy of soil properties tends to be diminished. This problem is mainly attributed to the differential nonlinear relationships of soil properties with the spectra, greater variance over wider gradients, and non-standardized spectral analysis leading to larger errors (Stenberg et al., 2010; Nocita et al., 2014; Castaldi et al., 2018).

The Land Use/Cover Area frame Survey (LUCAS) soil spectral library, developed by the European Union as an evolving database, is considered the world's largest unified, open-access dataset of topsoil properties (Panagos et al., 2012; Orgiazzi et al., 2018). So, it is reasonable to anticipate that predicting soil properties using the LUCAS soil spectral library should entail high reliability. The fact is, however, ultraspectral data contain thousands of wavelengths with strong collinearity and complex relationship among them, whereas the processing capabilities of traditional machine learning methods are limited. Therefore, those traditional modeling methods must resort to cumbersome pre-processing work of the spectra before extracting the feature wavelengths related to soil properties (Zhong et al., 2021).

Deep learning, as represented by convolutional neural networks (CNNs), is a family of computational methods that can extract features, layer by layer, via convolution and pooling. Because CNNs are characterized by weight sharing and local connections, the number of calibration parameters needed is reduced, which facilitates model optimization (Lecun et al., 2015). Veres et al. (2015) first applied deep learning to soil spectroscopy, and demonstrated that a one-dimensional (1D) CNN is effective at predicting specific soil properties. Later, some researchers also used 1D CNNs based on the LUCAS soil spectral library. For example, Liu et al. (2018) applied migration learning to predict the clay content of mineral soil samples, while Riese and Keller (2019) were able to classify soil texture into four groups.

Recently, Singh and Kasana (2019) used a 1D long short-term memory (LSTM) model to predict six soil physical and chemical properties from the LUCAS spectral library. Additionally, Padarian et al. (2019a,b) converted the original spectra of the LUCAS database into a two-dimensional (2D) spectrogram, and then used a 2D multi-task CNN to predict six soil properties. Furthermore, Tsakiridis et al. (2020) developed a local multi-channel 1D CNN to predict 10 soil physical and chemical properties from the LUCAS spectral library. They also explained the process by which soil clay content was modeled. While the previous studies relied on a relatively shallow CNN model (<10 layers), the LUCAS spectral library contains nearly 20,000 samples, each having up to 4200 spectral wavelengths. With such big data, a deep convolutional neural network (DCNN) architecture is perhaps more appropriate and effective than a shallow CNN.

Therefore, the purpose of this study was to explore the modeling potential of DCNNs for soil properties when applied to the LUCAS soil spectral library. First, we introduced the conventional architecture and working mechanism of a CNN. Second, we built 16-layer DCNN models and implemented them for regression modeling of seven soil properties and classification modeling of soil texture using only the original absorbance spectra from the soil spectral library. The results were compared with those obtained by traditional machine learning methods and previous research findings. We also assessed the effects of data pre-processing on single-task and multi-task DCNN modeling. Finally, we analyzed the relative importance of feature wavelengths extracted by a DCNN model to soil properties.

2. Materials and methods

2.1. Modeling methods

2.1.1. DCNN

The procedure for CNN modeling of soil properties using spectral data is sketched in Fig. 1a. First, soil spectral data are organized as a matrix to fit the learning architecture of the CNN model. Next, the convolutional layer extracts the features of the input data via multiple convolutional kernels of a certain size and corresponding step size. The pooling layer, also called the downsampling layer, then replaces the values of the original range with the maximum or mean values of a certain size-sampling range. This step lessens the data for processing, while retaining key feature information. Finally, the fully connected layer, coming after the convolutional and pooling layers, non-linearly combines the extracted features to produce the output results. The hyperparameters are mainly the number of neurons used, and the output layer comprises the regression values or classification classes of the predicted soil properties.

The training dataset is calibrated and validated once for every epoch. Specifically, the entire learning process updates the weighted parameter values in an iterative manner through continuous epoch cycles, which minimizes the loss function value and thereby engages in autonomous learning (Petersson et al., 2016). During the model building, the activation function is typically situated behind the convolutional and fully connected layers, where it implements a nonlinear activation function to improve the model's expressive ability. An optimizer calculates and updates the model parameters to better approximate or reach their optimal range and further minimize loss. To prevent model overfitting, the dropout and early-stopping mechanisms respectively mask out a portion of neurons per calibration batch and halt the model prematurely if the loss function is not sufficiently enhanced by certain patience during calibration.

Two single-task DCNN models, this study proposed LucasVGGNet-16 (Fig. 1b) and LucasResNet-16 (Fig. 1d),. To strengthen their effectiveness and comparability, the models were adjusted to the same number of layers (i.e., 16 layers) and a similar number of parameters for calibration (~1.3 million). LucasVGGNet-16 (Simonyan and Zisserman, 2014) consists of 13 convolutional layers, five pooling layers, and three fully connected layers. LucasResNet-16 (He et al., 2016) contains four residual blocks (Fig. 1c), two pooling layers, and three fully-connected layers, with each residual block harboring three convolutional layers. The chief advantage of LucasResNet-16 is conferred by the residual block arising from cross-layer connections that can overcome gradient disappearance in a DCNN. Moreover, a multi-task model, multi-LucasResNet-16 (Fig. 1e), was built on the basis of LucasResNet-16 by sharing convolution and pooling layers. After the "shared layers" extracted the features of soil spectra, the information was directed to different branches, one for each target soil property (Padarian et al., 2019a).

The three DCNN models built in this study have several fixed functions and hyperparameter settings. Sigmoid is used as an activation

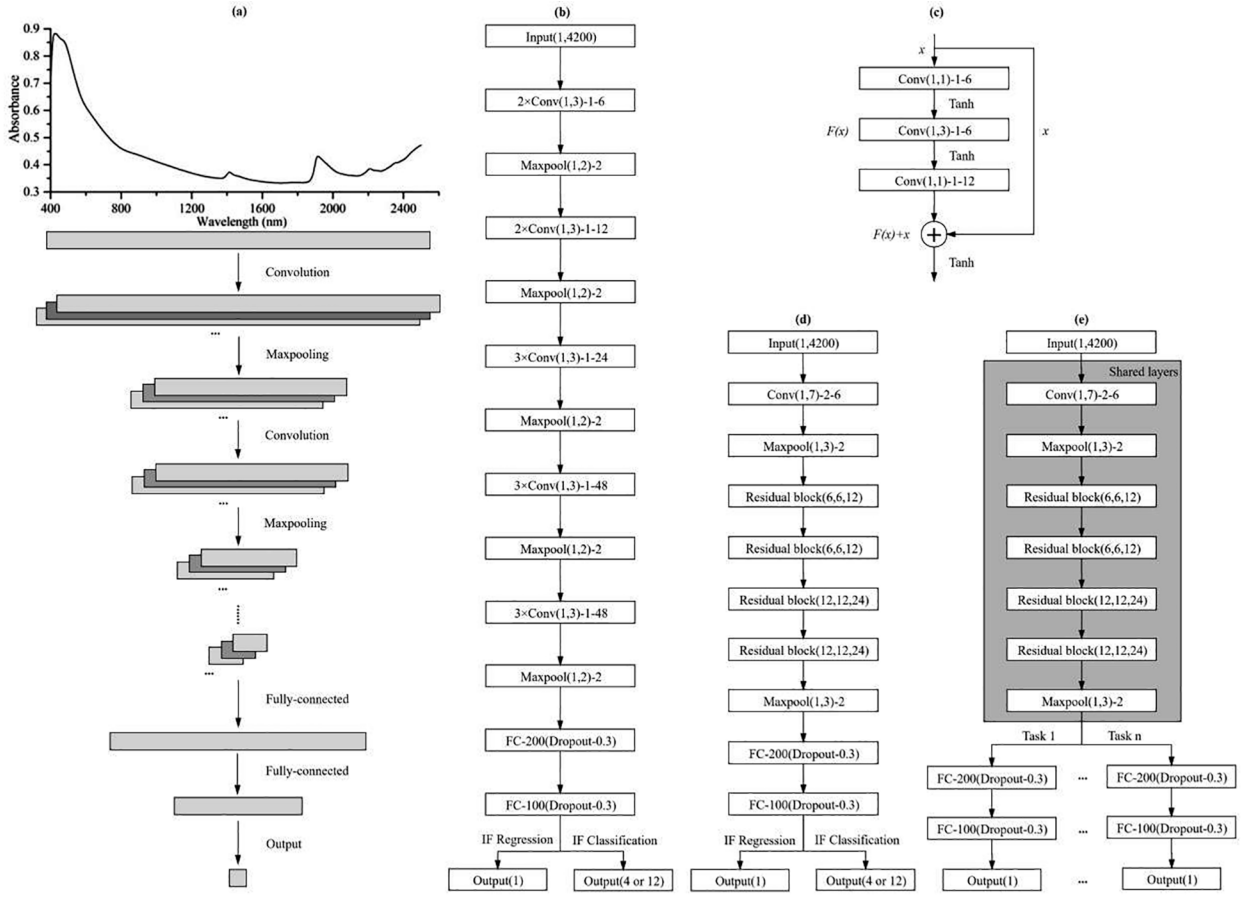


Fig. 1. The convolutional neural network model architecture in its conventional form (a), the LucasVGGNet-16 architecture (b), the residual block (c), the LucasResNet-16 architecture (d), and the multi-LucasResNet-16 architecture (e). Input(1,4200) denotes a spectral matrix with 1 row and 4200 columns; 2 × Conv (1,3)-1–6 denotes two convolutional layers present, each with the convolutional kernel size of (1,3), the step size of 1, and the number of convolutional kernels of 6; Maxpool(1,2)-2 denotes the maximum pooling layer, for which the pooling range is (1,2) and the step size is 2; FC-200(Dropout-0.3) indicates the fully-connected layer consisting of 200 neurons, 30% of which are inactivated at random; Output(1) corresponds to the output value for a given soil property; Residual block(6,6,12) states that the number of convolution kernels in that residual block is 6, 6, and 12 (respectively for its layers), and so on for the other model components.

function for the output layer, while Tanh is used for other layers; the optimizer is Nadam; the batch size is 32; the learning rate is 0.0001; the patience is 100 epochs; the number of neurons in the fully-connected layer is 200 and 100, with dropout introduced into the fully-connected layer to randomly inactivate 30% of the neurons. Once the regression modeling is implemented, the soil property value is normalized by dividing the maximum value before its input to the model. An inverse normalization step is executed to obtain the predicted value by multiplying the maximum value after the model's output. For the classification modeling, one-hot coding is done before inputting the soil texture class into each model, and the output layer is simply the prediction probability of each class in either four groups or 12 levels. The soil texture having the largest value is deemed the predicted class. In this study, the DCNN models were implemented in Python v3.7.3 (https://www.python.org/) using deep learning Keras framework. The code can be found at https://github.com/ZhongL1007/DCNNs.

2.1.2. Traditional machine leaning methods

Rooted in statistical learning theory, the support vector machine (SVM) maps data to high-dimensional feature space via a kernel function. SVM distinguishes a hyperplane serving as a decision boundary on which the prediction error is minimized (Burges, 1998). Additionally, partial least squares regression (PLSR) combines the advantages of principal component analysis, typical correlations, and multiple linear regression. PLSR is especially advantageous when applied to a multicollinearity predictor variable matrix composed of data. In this case, prediction and observation variables are projected onto a new data space to maximize their covariance (Wold et al., 2001). Furthermore, random forest (RF) is an ensemble learning method that builds multiple decision tree models. The final result of RF is determined by an average or majority voting principle for a single model result, and this method is often used for classification purposes (Breiman, 2001).

We optimized the parameters of the three traditional machine learning models (SVM, PLSR, and RF) using parameter iteration and 5-fold cross-validation (Zhang et al., 2019a). For the SVM model, we used radial basis function as kernel function. We optimized penalty parameter from a list of 1, 10, 50, 100, 200, 500, and 1000, and the parameter gamma from a list of 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, and 0.1. For the PLSR model, we iterated the first 100 principal components and selected the principal components when the accuracy tended to be smooth. For the RF model, we optimized the number of decision trees from a list of 50, 100, and 200, and the maximum depth of the tree from 3 to 10. In addition, the optimal parameters were determined considering the relative influence of model overfitting. The SVM, PLSR, and RF models were run in the corresponding machine learning modules of the Sklearn interface in Python v3.7.3.

2.2. Model assessment

In the regression modeling, the coefficient of determination (R^2; Eq. (1)), RMSE (Eq. (2)), were used to assess the goodness-of-fit and accuracy. The larger the R^2 and the smaller the RMSE, the better the

prediction performance and stability of the fitted model. In the classification modeling, the results were assessed on the basis of overall accuracy and recall. Overall accuracy is defined as the number of correctly classified samples divided by the total number of samples tested; likewise, recall of a class is the number of correctly classified samples divided by the total number of samples of that class.

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

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

where n is the number of samples, y_i is the observed value,$\hat{y}_i$is the predicted value, $\bar{y}$is the mean of the observed values, and *SDy* is the standard deviation of the observed values.

2.3. LUCAS soil spectral library

The topsoil dataset provided by LUCAS is a European-scale soil spectral library, whose aim is to create the first unified and comparable soil database for Europe and support policy development (Tóth et al., 2013; Orgiazzi et al., 2018). In 2009, approximately 20,000 topsoil samples were collected in 25 European countries using standardized sampling procedures. All these samples were analyzed using standard testing methods in laboratories certified by the International Organization for Standardization. Twelve physical and chemical properties of the topsoil samples were determined, of which 10 main soil properties were selected in the present study, namely, organic carbon (OC) content, nitrogen (N) content, phosphorous (P) content, potassium (K) content, cation exchange capacity (CEC), pH measured in water (pH), calcium carbonate ($CaCO_3$) content, clay content (<0.002 mm), silt content (0.02–0.002 mm), and sand content (2–0.02 mm). The absorbance spectra were measured using a FOSS XDS rapid content analyzer (Foss NIR Systems Inc., Laurel, MD, USA) after the samples had been air-dried and sieved (≤2 mm). The spectral range and resolution were 400.0–2499.5 nm and 0.5 nm, respectively. A total of 4200 wavelengths were recorded per sample.

To analyze the effects of data pre-processing on DCNN modeling, we applied spectral pre-processing to the original absorbance spectra based on previous studies (Ng et al., 2019; Padarian et al., 2019a; Tsakiridis et al., 2020; Tziolas et al., 2020; Zhang et al., 2019a). First, only spectral transformations were performed using six common methods: i) a zero-order Savitzky-Golay filter with a window width of 50 (Abs-SG0); ii) SG0 followed by the standard normal variate (Abs-SG0-SNV); iii) a first-order Savitzky-Golay filter with a window width of 50 (Abs-SG1); iv) SG1 followed by the standard normal variate (Abs-SG1-SNV); v) a second-order Savitzky-Golay filter with a window width of 50 (Abs-SG2); vi) SG2 followed by the standard normal variate (Abs-SG2-SNV). Second, spectral dimensionality reductions were performed on the basis of spectral transformations. Due to the presence of noise, the 400–499.5 nm and 2450–2499.5 nm wavelength ranges were removed. Then the data were downsampled by retaining one value every 10 nm, thus leaving 195 data points after pre-processing.

Based on the LUCAS database, we aimed to model soil properties using only the spectral data without any prior information on the samples. Therefore, we implemented the modeling for organic and mineral soil samples together to explore the potential of DCNNs for predicting soil properties. In this way, a DCNN model with broader applicability could be obtained.

2.4. Spectral feature extraction

Proposed by Lundberg and Lee (2017), SHapley Additive exPlanations (SHAP) is an interpretable method that takes the classical Shapley values from game theory (Lipovetsky and Conklin, 2001). These values are linked to a local interpretation to derive a unified method for interpreting the output of machine learning models, enabling SHAP to rank the relative importance of features. The SHAP value indicates the relative contribution (positive or negative) of each feature in a given sample to the model's output. By calculating the SHAP value for each wavelength in the DCNN model, the respective relative contribution is obtainable, providing a basis to extract the spectral features. Recently, Padarian et al. (2020) demonstrated the successful use of SHAP values for interpreting digital soil mapping models. Haghi et al. (2021) identified and interpreted the important wavelengths that were used by the Cubist and CNN models to predict soil properties. Here, we calculated the SHAP values by calling the DeepExplainer module in the SHAP interface. Suppose the i-th sample is x_i and the j-th feature of the i-th sample is x_{ij}; the model predicted value for this sample set would be$\hat{y}_i$, for which $\phi 0$ is the base value of the entire model (i.e., usually the mean of the target variables of all samples), with the SHAP value $f(x_{ij})$ for x_{ij} obeying Eq. (3):

$$\hat{y}i = \phi 0 + f(xi1) + f(xi2) + \ldots + f(xij) \quad (3)$$

3. Results

3.1. Descriptive statistics of soil properties

The LUCAS dataset contained 19,036 samples of spectral wavelengths, spanning 23 countries. A total of 10 soil properties were selected for analysis: OC content, N content, P content, K content, CEC, pH, $CaCO_3$ content, clay content, silt content, and sand content. In screening them for outliers, we found that soil K content was missing for one sample and soil particle size was missing for 1097 samples. After these outliers were eliminated, 75% and 25% of the dataset was randomly assigned to serve as training and testing sets, respectively. At each epoch, 20% of the training samples were randomly selected as the validation set and the remaining 80% as the calibration set. The descriptive statistics of the 10 soil properties in the two sets of samples are summarized in Table 1. The range of each soil property was large. The mean and standard deviation of the soil properties were similar in the training and testing sets, indicating their relatively uniform distribution.

Soil texture was classified into four groups and 12 levels in terms of the contents of clay, silt, and sand (Table 2). In the four groups of classification, the sample sizes of sand, loam, clay loam, and clay were 1220, 6924, 4949, and 4846, respectively. Among the 12 levels of classification, all soil texture classes were found in the dataset. Sandy clay and heavy clay had relatively few samples, whereas the others had a sample size of > 500, with sandy loam being most frequent ($n = 4625$). The ratio of training to testing sample sizes of each soil texture class was close to 3:1. The triangular diagram of soil texture shows the distribution of the training and testing samples (Fig. 2).

3.2. Comparison of DCNNs with traditional regression methods

Regression predictions were carried out for seven soil properties (OC, N, P, K, CEC, pH, and $CaCO_3$) using four different models (LucasResNet-16, LucasVGGNet-16, SVM, and PLSR). The testing set accuracy of the models was compared in terms of R^2 and RMSE (Table 3). When predicting the soil properties, LucasResNet-16 or LucasVGGNet-16 yielded considerably better prediction accuracy than SVM or PLSR. Despite their similar accuracy, compared with LucasVGGNet-16, the overall prediction performance of LucasResNet-16 was slightly better for most soil properties. DCNN modeling performed well in predicting soil OC content, N content, CEC, pH, and $CaCO_3$ content ($R^2 = 0.763$–0.960). Soil K content was adequately predicted ($R^2 = 0.591$), and soil P content was poorly predicted ($R^2 = 0.367$).

Scatter plots of the measured and predicted values for the seven soil properties in the LucasResNet-16 model are shown in Fig. 3. Low biases

Table 1
Descriptive statistics of 10 soil properties in the LUCAS dataset.

Soil properties	Valid samples	Training					Testing				
		Samples	Min	Max	Mean	Standard deviation	Samples	Min	Max	Mean	Standard deviation
OC/g•kg^{-1}	19,036	14,277	0.00	586.80	50.26	91.66	4759	0.00	577.00	49.23	90.24
N/g•kg^{-1}	19,036	14,277	0.00	36.20	2.94	3.76	4759	0.00	38.60	2.89	3.75
P/mg•kg^{-1}	19,036	14,277	0.00	1366.40	30.01	33.42	4759	0.00	431.90	30.27	31.10
K/mg•kg^{-1}	19,035	14,276	0.00	7342.00	196.97	236.86	4759	0.00	3059.30	197.29	204.94
CEC/cmol(+)•kg^{-1}	19,036	14,277	0.00	227.70	15.84	14.46	4759	0.00	234.00	15.51	14.56
pH	19,036	14,277	3.40	9.75	6.20	1.36	4759	3.21	10.08	6.19	1.35
$CaCO_3$/g•kg^{-1}	19,036	14,277	0.00	909.00	51.68	125.72	4759	0.00	944.00	51.36	124.10
Clay/%	17,939	13,454	0.00	79.00	18.89	13.02	4485	0.00	76.00	18.86	12.97
Silt/%	17,939	13,454	0.00	92.00	38.31	18.82	4485	1.00	88.00	37.98	18.34
Sand/%	17,939	13,454	1.00	99.00	42.80	26.08	4485	1.00	99.00	43.14	26.17

OC, organic carbon; N, nitrogen; P, phosphorus; K, potassium; CEC, cation exchange capacity.

Table 2
Classification of soil texture in the LUCAS dataset.

Four groups of classification	12 levels of classification	All samples	Training samples	Testing samples
Sand	Sand and loamy sand	1220	908	312
Loam	Sandy loam	4625	3448	1177
	Loam	991	751	240
	Silty loam	1308	1004	304
Clay loam	Sandy clay loam	524	391	133
	Clay loam	2019	1505	514
	Silty clay loam	2406	1808	598
Clay	Sandy clay	24	18	6
	Loamy clay	1744	1297	447
	Silty clay	2320	1752	568
	Clay	710	534	176
	Heavy clay	48	38	10
Total		17,939	13,454	4485

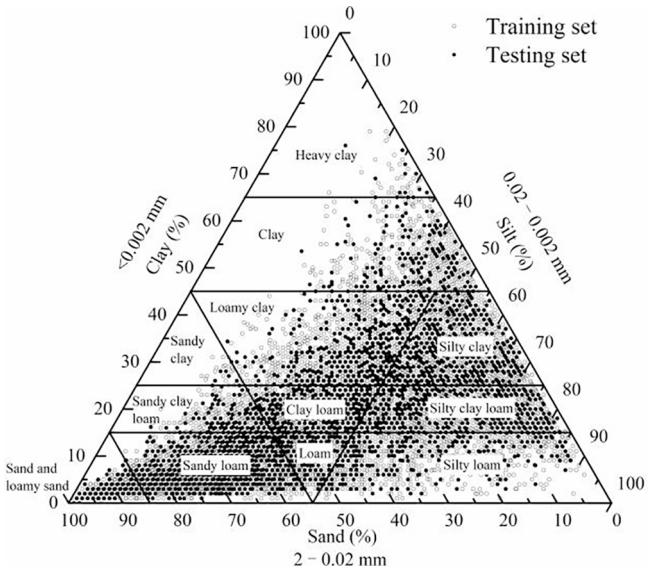


Fig. 2. Distribution of soil texture in the LUCAS dataset.

were observed for soil OC content, N content, CEC, pH, and $CaCO_3$ content, while high biases were observed for soil P and K contents. The predicted values tended to underestimate the measured values of all properties, and the prediction error was greater for higher measured values where there were fewer data.

3.3. Comparison of DCNNs with traditional classification methods

Classification predictions were made for four groups and 12 levels of soil texture using four different models (LucasResNet-16, LucasVGGNet-16, SVM, and RF). Table 4 provides information on the overall classification accuracy of the models for the calibration and testing sets. In both cases (four groups and 12 levels), LucasResNet-16 and LucasVGGNet-16 were considerably more accurate than either SVM or RF. While LucasVGGNet-16 performed slightly better than LucasResNet-16, there was greater divergence between the calibration and testing of LucasVGGNet-16, indicating more pronounced overfitting. Since there was less risk of overfitting by LucasResNet-16, this model had higher applicability for soil texture classification. The overall testing set accuracy of LucasResNet-16 for the four groups of soil texture was 0.749. After subdividing soil texture into the 12 levels, the overall testing set accuracy of LucasResNet-16 was reduced to just 0.566.

Using a confusion matrix, the recall was calculated for the four groups of soil texture classified by the LucasResNet-16 model (Fig. 4a). Loam had the highest accuracy with a recall of 0.84, followed by clay with a recall of 0.75, below which were clay loam (0.65) and sand (0.63). Moreover, sand was easily mistaken for loam (0.36), and likewise loam for clay loam (0.11), clay loam for either loam (0.22) or clay (0.13), while clay was easily mistaken for clay loam (0.22). The recall was also calculated for the 12 levels of soil texture classified by the LucasResNet-16 model (Fig. 4b). Compared with the other soil texture classes, sand and loamy sand, sandy loam, and clay were better classified with a recall of 0.61, 0.75, and 0.64, respectively. Clay loam, silty clay loam, loamy clay, and silty clay had a slightly lower recall of 0.54–0.56, while the recall of silty loam was even lower, at 0.47. However, it was more difficult to classify loam (0.09) and sandy clay loam (0.13), with neither sandy clay nor heavy clay classifiable at all. Furthermore, mistaken classifications into similar soil texture classes were prone to occur on the abscissa, especially as the following five classes: sandy loam, clay loam, silty clay loam, loamy clay, and silty clay.

The prediction results of the LucasResNet-16 model for soil texture are presented in triangular diagrams (Fig. 5), conveying the distribution of correctly versus incorrectly predicted samples. For the four groups of soil texture, the incorrectly predicted samples were mainly distributed near the boundary of each class, with clay mostly near the clay loam, loam concentrated near the clay loam and sand, and sand often found near the loam (Fig. 5a). For the 12 levels of soil texture, many samples with incorrect predictions could also be found near the boundary of each soil texture class and their intersection (Fig. 5b).

3.4. Effects of data pre-processing on DCNN modeling

We tested the results of the single-task LucasResNet-16 and multi-task multi-LucasResNet-16 models based on the original spectra (4200 data points) and downsampled spectra (195 data points) with different spectral transformations. The performance of the best spectral pre-

Table 3
The testing set accuracy of four different models for seven soil properties.

Soil properties	Assessment indicators	LucasResNet-16	LucasVGGNet-16	SVM	PLSR
OC	R^2	0.952	0.953	0.906	0.906
	RMSE	19.837	19.612	27.618	27.732
N	R^2	0.935	0.934	0.820	0.859
	RMSE	0.957	0.960	1.589	1.409
P	R^2	0.367	0.386	−2.160	0.282
	RMSE	24.743	24.365	55.283	26.347
K	R^2	0.591	0.494	−2.043	0.394
	RMSE	131.071	145.717	357.512	159.548
CEC	R^2	0.763	0.772	0.531	0.712
	RMSE	7.089	6.953	9.969	7.816
pH	R^2	0.938	0.936	0.833	0.876
	RMSE	0.334	0.340	0.550	0.473
$CaCO_3$	R^2	0.960	0.955	0.843	0.903
	RMSE	24.908	26.458	49.110	38.649

R^2, coefficient of determination; RMSE, root mean squared error; SVM, support vector machine; PLSR, partial least squares regression.

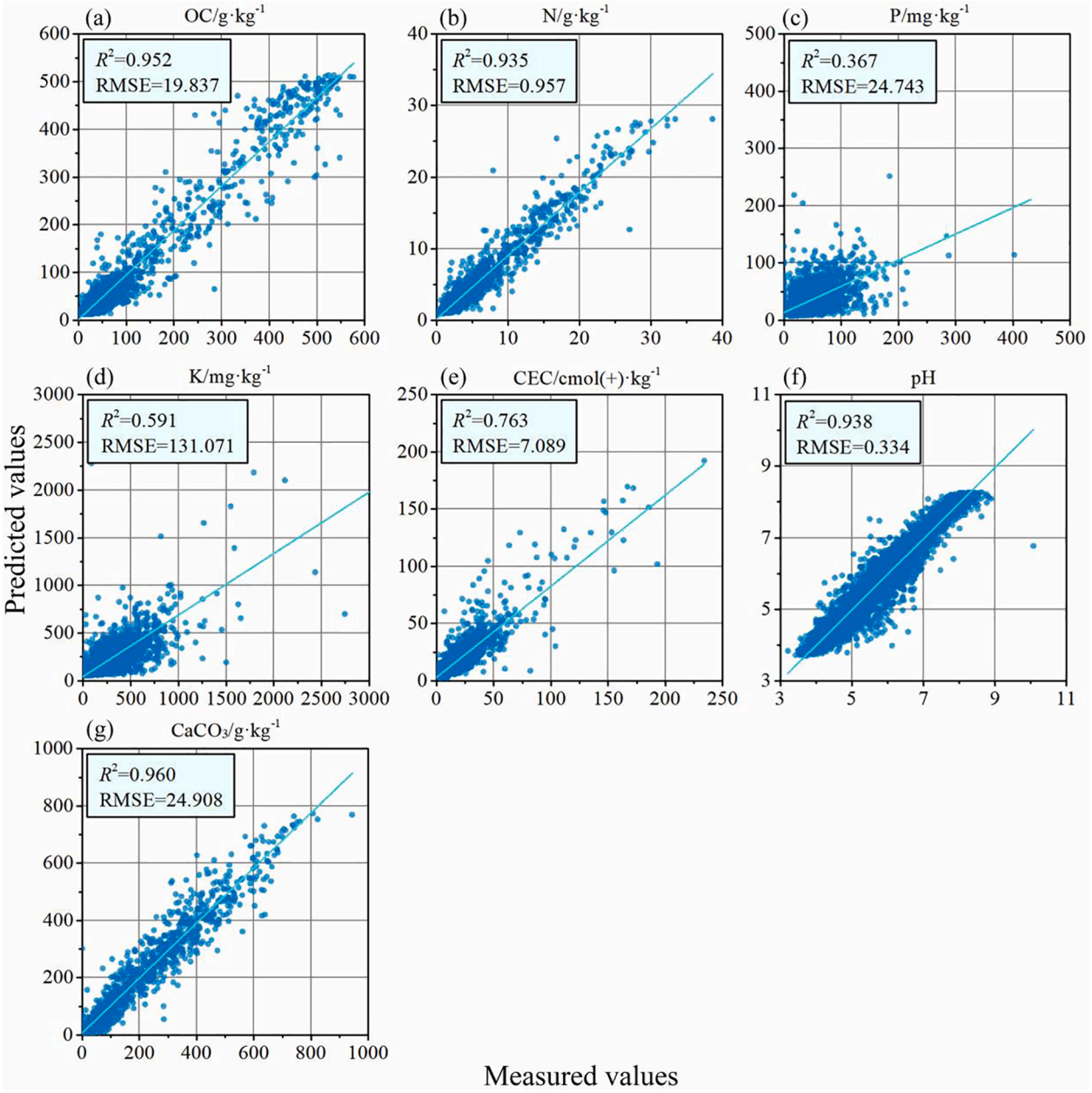
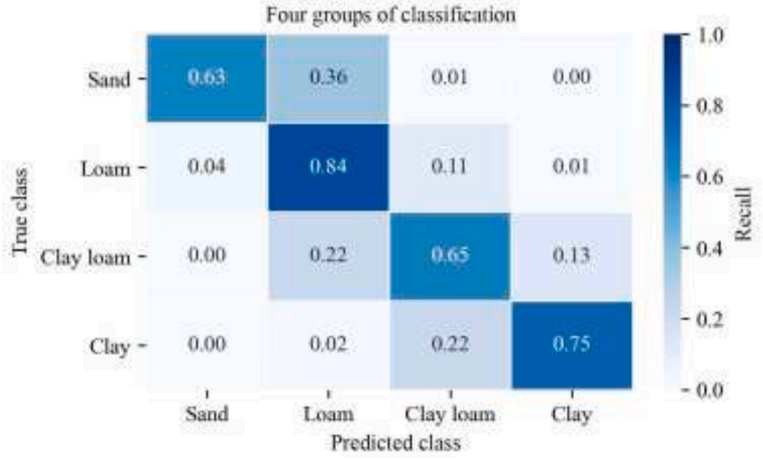


Fig. 3. Scatter plots of measured and predicted values for seven soil properties in the LucasResNet-16 model. (a) organic carbon (OC) content; (b) nitrogen (N) content; (c) phosphorus (P) content; (d) potassium (K) content; (e) cation exchange capacity (CEC); (f) pH; (g) calcium carbonate ($CaCO_3$) content.

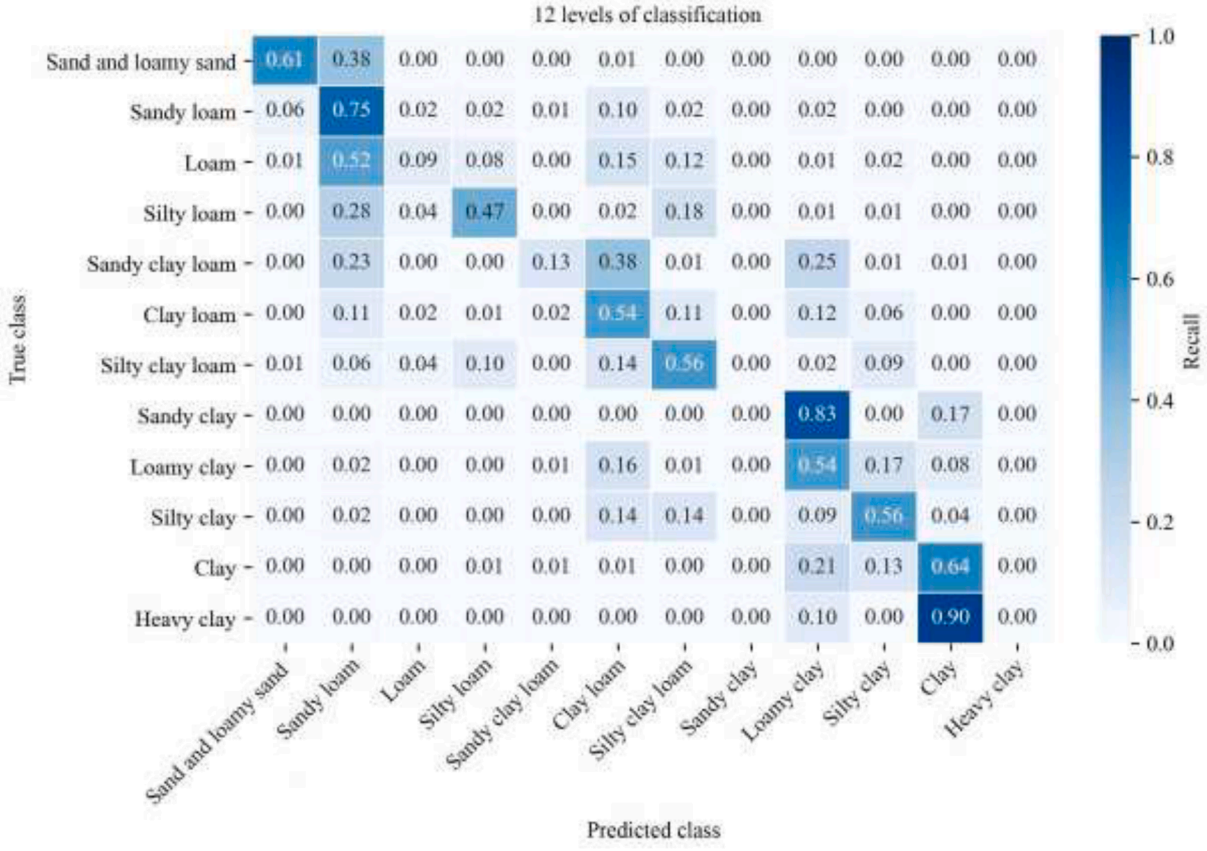
Table 4
The overall classification accuracy of four different models for soil texture.

Soil texture	Assessment indicators	LucasResNet-16	LucasVGGNet-16	SVM	RF
Four groups of classification	Calibration set accuracy	0.803	0.853	0.696	0.596
	Testing set accuracy	0.749	0.760	0.678	0.553
12 levels of classification	Calibration set accuracy	0.628	0.709	0.507	0.414
	Testing set accuracy	0.566	0.589	0.489	0.382

RF, random forest.



(a)



(b)

Fig. 4. Recall from the confusion matrix for (a) four groups and (b) 12 levels of soil texture classification by the LucasResNet-16 model (testing set).

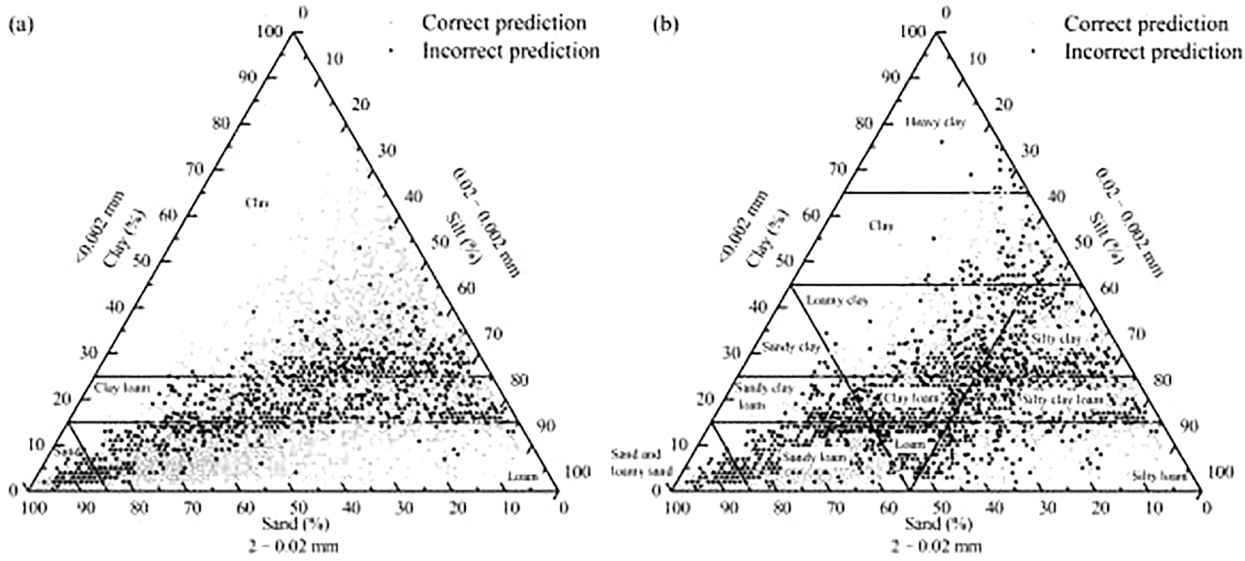


Fig. 5. Distribution of soil texture predicted by the LucasResNet-16 model for the four groups (a) and the 12 levels (b) of classification.

processing source is given (Table 5, Fig. 6). With regard to spectral transformation, single-task modeling achieved the highest accuracy for some soil properties (e.g., OC, N, and $CaCO_3$) based on the original absorbance spectra. SG1 was the best source of spectral transformation for some soil properties when modeling with 4200 data points. However, the best source of spectral transformation varied with soil properties when modeling with 195 data points. After spectral downsampling, the modeling accuracy of all soil properties decreased to different degrees.

Multi-task modeling based on the original absorbance spectra achieved the highest modeling accuracy in both cases of spectral pre-processing. After spectral downsampling, the modeling accuracy of all soil properties also decreased to different degrees. Compared with single-task modeling, the prediction performance of multi-task modeling was slightly improved for most soil properties. The results showed that spectral transformation was effective for the prediction of specific soil properties, while spectral downsampling caused a reduction in the modeling accuracy. The multi-task DCNN model improved the modeling performance compared with the single-task DCNN model.

3.5. Contribution of feature wavelengths to soil properties

We calculated the SHAP value of each wavelength in the testing set of the LucasResNet-16 model, and obtained the averaged relative contribution from feature wavelengths to each soil property (Fig. 7). The relative contribution and position of feature wavelengths were different among the soil properties and characterized by multiple characteristic peaks. The top 10 feature wavelengths that contributed most to each soil property were extracted (Table 6). The distributions of these feature wavelengths could be roughly summarized as follows: for OC, they were concentrated around 2309 and 2180 nm; likewise, for N, around 2055, 2393, 783, and 595 nm; for P, around 667, 931, 1716, and 1960 nm; for K, around 2106, 674, 874, and 1986 nm; for CEC, around 2178, 2347, 1731, 584, and 1418 nm; for pH, around 722, 2010, and 1666 nm; for $CaCO_3$, around 1998 nm. Considering soil texture, the four groups of classification had wavelengths mainly around 1414, 2208, 1358, and 2376 nm, while wavelengths of the 12 levels predominated around 1415, 2377, and 1369 nm.

Table 5
The testing set accuracy of the LucasResNet-16 and multi-LucasResNet-16 models for prediction of soil properties based on the original spectra (4200 data points) and downsampled spectra (195 data points) in the best spectral pre-processing source.

Soil properties	Assessment indicators	LucasResNet-16 (single-task DCNN)				multi-LucasResNet-16 (multi-task DCNN)	
		4200 data points		195 data points		4200 data points (Best source: Abs)	195 data points (Best source: Abs)
OC	R^2	Abs	0.952	Abs	0.940	0.955	0.947
	RMSE		19.837		22.136	19.130	20.803
N	R^2	Abs	0.935	Abs	0.911	0.933	0.916
	RMSE		0.957		1.115	0.971	1.087
P	R^2	Abs-SG1	0.372	Abs-SG1-SNV	0.285	0.395	0.250
	RMSE		24.649		26.305	24.184	26.937
K	R^2	Abs	0.591	Abs-SG2-SNV	0.523	0.593	0.397
	RMSE		131.071		141.471	130.706	159.093
CEC	R^2	Abs-SG1	0.822	Abs-SG0	0.768	0.794	0.771
	RMSE		6.142		7.013	6.614	6.966
pH	R^2	Abs-SG1	0.941	Abs-SG2-SNV	0.914	0.942	0.897
	RMSE		0.327		0.395	0.326	0.432
$CaCO_3$	R^2	Abs	0.960	Abs	0.931	0.961	0.912
	RMSE		24.908		32.555	24.526	36.795

Abs, original absorbance spectra; SG0, zero-order Savitzky-Golay filter with a window width of 50; SG1, first-order Savitzky-Golay filter with a window width of 50; SG2, second-order Savitzky-Golay filter with a window width of 50; SNV, standard normal variate.

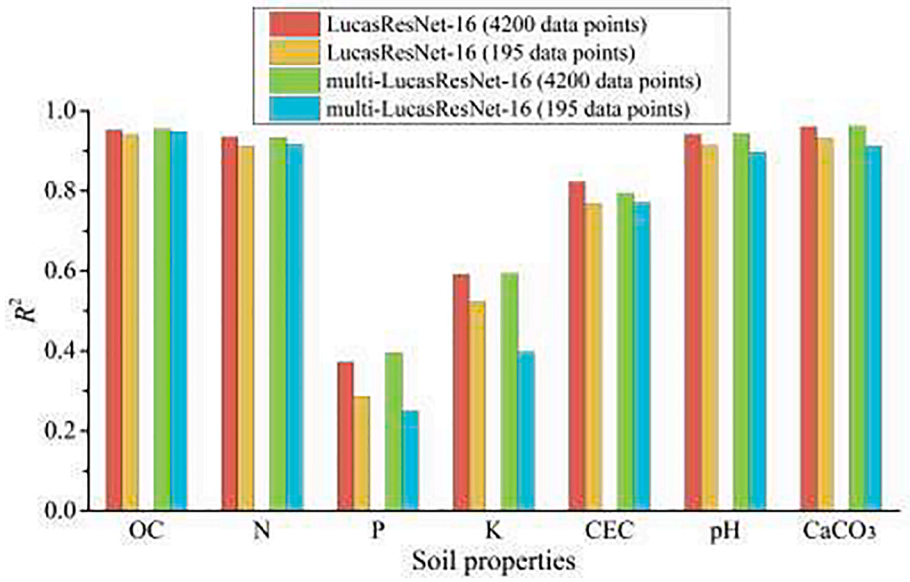


Fig. 6. Histogram of the testing set accuracy of the LucasResNet-16 and multi-LucasResNet-16 models for the prediction of soil properties based on the original spectra (4200 data points) and downsampled spectra (195 data points) in the best spectral pre-processing source. R^2, coefficient of determination.

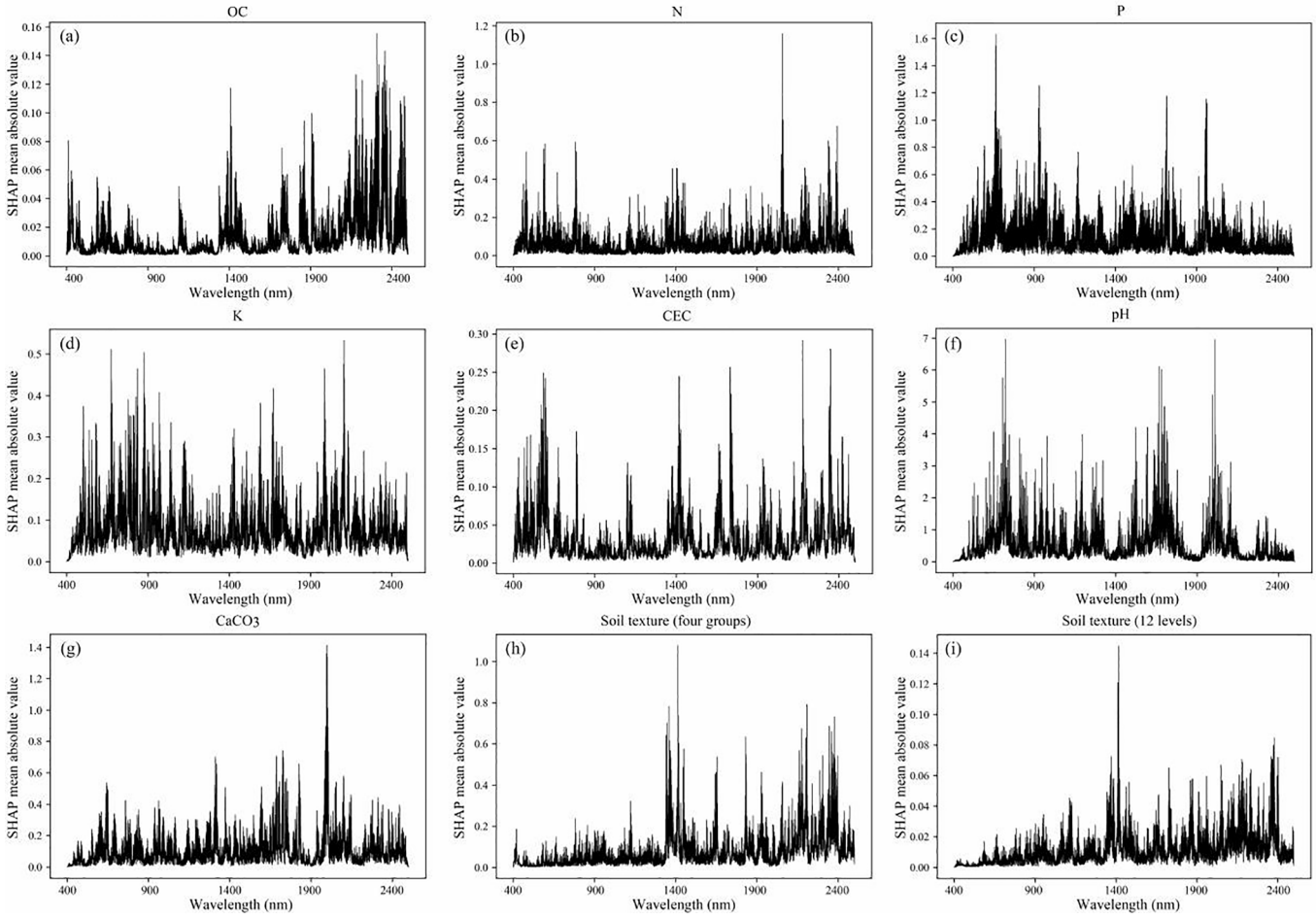


Fig. 7. Distribution of the relative contribution from feature wavelengths to each soil property in the testing set of the LucasResNet-16 model. (a) OC content; (b) N content; (c) P content; (d) K content; (e) CEC; (f) pH; (g) $CaCO_3$; (h) soil texture (four groups); (i) soil texture (12 levels).

Table 6
The top 10 feature wavelengths that contributed most to different soil properties in the testing set of the LucasResNet-16 model.

Soil properties	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6	Top 7	Top 8	Top 9	Top 10
OC	2309	2357	2319	2355	2180	2369	2219	2345	2356	2359
N	2055	2059	2054	2393	2337	783	595	2345	587	787
P	667	664	931	1716	1960	663	1963	1962	1959	1714
K	2106	674	874	2104	834	1986	2107	675	2105	676
CEC	2178	2347	1731	2177	584	1418	1419	595	2345	593
pH	722	2010	1666	1682	706	1994	1698	714	1658	1594
$CaCO_3$	1998	1994	1999	1997	1995	2002	1996	1991	2004	2003
Soil texture (four groups)	1414	2208	1358	1416	2376	1346	2344	2176	2360	1342
Soil texture (12 levels)	1415	1411	1416	1417	2377	2371	2378	2359	1369	2402

4. Discussion

4.1. Prediction accuracy of soil properties

In this study, we built three 16-layer DCNN models (LucasResNet-16, LucasVGGNet-16, and multi-LucasResNet-16) to predict different soil properties from the LUCAS spectral library. The regression predictions of soil OC content, N content, CEC, pH, and $CaCO_3$ content based on the original spectra all achieved high accuracy (Table 3). These results are consistent with those of Viscarra Rossel and Webster (2012), who used ordinary least-squares regression to predict 24 soil properties based on Vis-NIR spectral data of approximately 20,000 soil samples taken across Australia. The moderate prediction accuracy of soil K content and poor prediction accuracy of soil P content could be attributable to a narrow chemical range of these two variables. They have poor correlations with primary soil variables such as OC, clay, and $CaCO_3$ contents that are more directly assessed by Vis-NIR spectroscopy (Chang et al., 2001; Volkan Bilgili et al., 2010).

Comparing the two single-task DCNN models built in this study, we found that the prediction accuracy of LucasResNet-16 was slightly better than those of LucasVGGNet-16 for most soil properties (Table 3). This result suggests that implementing LucasResNet-16 with residual learning can alleviate the problem of gradient disappearance in a DCNN, which, in turn, improves the model accuracy and stability at later stages of calibration (He et al., 2016). Additionally, we found that the single-task DCNN models tended to underestimate the high values of soil properties. This is because in the high value region with less data, the model learning is insufficient.

To further gain insight into the performance of different modeling approaches for soil properties, we compared our 1D single-task and multi-task DCNN models with the 1D LSTM (Singh and Kasana, 2019), 2D CNN and 2D multi-CNN (Padarian et al., 2019a), and 1D local multi-CNN (Tsakiridis et al., 2020) models reported in previous studies (Table 7). We found that compared with 1D LucasResNet-16, the prediction performance of 1D multi-LucasResNet-16 was slightly improved for most soil properties. In the case of single-task modeling, compared with 1D LSTM, our 1D LucasResNet-16 model reduced RMSE by 14.7%, 16.8%, 9.0%, and 22.1% for soil OC, N, CEC, and pH, respectively. Compared with 2D CNN, our 1D LucasResNet-16 model reduced RMSE by 38.3%, 37.9%, 28.4%, and 34.6% for soil OC, N, CEC, and pH, respectively. As for multi-task modeling, compared with 1D local multi-CNN, our 1D multi-LucasResNet-16 model reduced RMSE by 9.4% and 6.2% for soil pH and $CaCO_3$, respectively. Compared with 2D multi-CNN, our 1D multi-LucasResNet-16 model reduced RMSE by 8.4% and 38.5% for soil N and pH, respectively.

The results showed that DCNN was superior to the single-task shallow CNN for predicting specific soil properties, as deeper layers can learn more complex structures (Lecun et al., 2015; Zhang et al., 2016). Additionally, the multi-task DCNN model outperformed the single-task DCNN model. This is primarily because a multi-task DCNN model considers the correlation between soil properties, such as pH and $CaCO_3$ ($r = 0.52$, $P < 0.01$), in addition to P and K ($r = 0.34$, $P < 0.01$; Fig. 8), and the improvement of the modeling performance can be attributed to the high correlation between these properties.

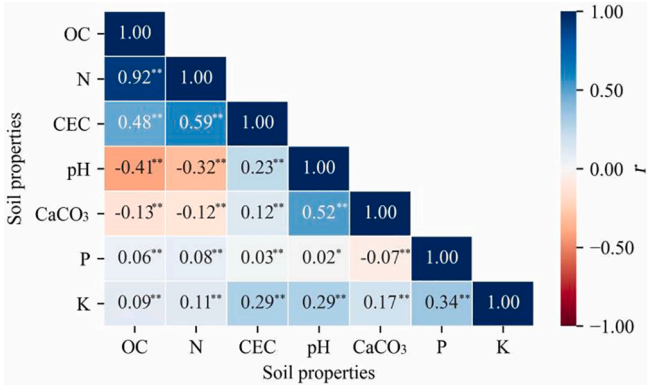


Fig. 8. The correlation between soil properties for samples of the LUCAS topsoil dataset. * $P < 0.05$; ** $P < 0.01$.

Table 7
Comparison of R^2 and RMSE on the testing set of 1D single-task and multi-task DCNN models built in this study with the results of other studies using all soil samples (mineral and organic).

Model	Assessment indicators	OC	N	CEC	pH	$CaCO_3$	P	K
1D LucasResNet-16 (this study)	R^2	0.952	0.935	0.822	0.941	0.960	0.372	0.591
	RMSE	19.837	0.957	6.142	0.327	24.908	24.649	131.071
1D multi-LucasResNet-16 (this study)	R^2	0.955	0.933	0.794	0.942	0.961	0.395	0.593
	RMSE	19.130	0.971	6.614	0.326	24.526	24.184	130.706
1D LSTM (Singh and Kasana, 2019)	R^2	0.940	0.910	0.770	0.900	NA	NA	NA
	RMSE	23.250	1.150	6.750	0.420	NA	NA	NA
2D CNN (Padarian et al., 2019a)	R^2	0.880	0.830	0.660	0.870	NA	NA	NA
	RMSE	32.140	1.540	8.580	0.500	NA	NA	NA
1D local multi-CNN (Tsakiridis et al., 2020)	R^2	0.970	0.940	0.820	0.930	0.960	NA	NA
	RMSE	15.180	0.930	5.990	0.360	26.150	NA	NA
2D multi-CNN (Padarian et al., 2019a)	R^2	0.690	0.600	0.630	0.840	NA	NA	NA
	RMSE	16.820	1.060	6.510	0.530	NA	NA	NA

CNN, convolutional neural network; LSTM, long short-term memory network; NA, not available.

Interestingly, despite the high correlation between OC and N ($r = 0.92$, $P < 0.01$), the performance of multi-task modeling for N was not improved. A plausible reason is that N also had a high correlation with CEC ($r = 0.59$, $P < 0.01$), and multi-task modeling probably ignored some of their features. In summary, a multi-task DCNN model can identify features that are important for all soil properties; it contains additional soil properties that may offer more noise than information (Tsakiridis et al., 2020), and tends to ignore some subtle features. Therefore, one should consider whether to choose the single-task or multi-task model based on the soil properties to be predicted. The modeling performance can also be improved by adding suitable auxiliary tasks for one task.

It is suitable for CNN processing to convert 1D spectral signal into 2D spectrogram by decomposing into overlapping fragments (Padarian et al., 2019a,b). However, this may affect the continuity of the spectral features related to soil properties, which, in turn, influences the prediction accuracy (Ng et al., 2019). We can also conclude that 1D CNN is superior to 2D CNN in the case of single-task or multi-task modeling (Table 7), which is in agreement with the study of Ng et al. (2019). Therefore, we did not convert 1D spectra into 2D spectrogram for modeling the soil properties. In general, the DCNN modeling applied to a large spectral library offers a certain advantage, especially in the case of only using the original spectra.

4.2. Classification accuracy of soil texture

Compared with LucasVGGNet-16, the LucasResNet-16 model performed better for classification predictions of soil texture (Table 4). Riese and Keller (2019) also proposed a LucasCoordConv model with an additional coordinate layer that achieved an overall accuracy of 0.73 to classify soil texture into four groups using the LUCAS dataset. In addition, Hu et al. (2015) constructed an improved CNN model for hyperspectral image classification with an overall accuracy of 0.74. More recently, Zhong et al. (2020) used a multilayer perceptron model to classify soil texture into four groups and 12 levels based on hyperspectral data of 245 samples, achieving a respective overall accuracy of 0.68 and 0.55. Our DCNN models showed comparable or even higher accuracy in the classification of soil texture, and most of our incorrectly predicted samples were misclassified into classes similar to the actual soil texture (Figs. 4 and 5).

For the four groups of classification, the ranked accuracy for different soil texture classes was loam > clay > clay loam > sand; samples were most easily misclassified as loam and clay loam (Figs. 4a and 5a). For the 12 levels of classification, the ensuing ranking of accuracy was sandy loam > clay > sand and loamy sand > silty clay loam and silty clay > clay loam and loamy clay > silty loam > sandy clay loam > loam > sandy clay and heavy clay; samples were most easily misclassified as sandy loam, clay loam, silty clay loam, loamy clay, and silty clay (Figs. 4b and 5b). Both sets of results are consistent with frequent misclassifications near the class boundaries as noted by Chawla et al. (2004). The misclassifications occur because samples near the boundaries and intersections are more alike in their percentage content of each particle size, with a texture class and spectral features closely resembling each other. This is why the classification accuracy of loam and sandy clay loam was low in our study; these two classes were easily misclassified as sandy loam and clay loam were similar in terms of texture. In addition, it is easy to mistakenly classify soil texture classes from large sample sizes, mainly because when the number of samples is unbalanced the model can learn more feature information for classes during calibration (Zhong et al., 2020). Conversely, sandy clay and heavy clay cannot be classified at all because of the small number of samples and the insufficient feature learning.

4.3. Effects of data pre-processing

DCNN is able to achieve high prediction accuracy based on the original spectra, which saves the time spent in data pre-processing and thereby improves the efficiency of real-time monitoring (Padarian et al., 2019a; Xu et al., 2019; Zhang et al., 2019a). However, we still found that spectral transformation (especially SG1) was effective for the prediction of some soil properties by DCNN modeling (Table 5, Fig. 6), which is consistent with the results of Tsakiridis et al. (2020) and Haghi et al. (2021). In contrast, spectral downsampling led to a reduction in the modeling accuracy (Table 5, Fig. 6), mainly because the reduction in data dimension could change the original pattern of spectral signals and cause the loss of useful information (Zhang et al., 2019a). Moreover, deep learning approaches can discover intricate structures in high-dimensional data, while reducing the need for prior knowledge and human effort for feature engineering (Lecun et al., 2015). Therefore, the DCNN models built in this study do not need spectral dimension reduction for the prediction of soil properties based on a large soil spectral library.

4.4. Spectral feature wavelengths contributing to soil properties

The results of this study showed that the position of feature wavelengths related to different soil properties were not alike, with multiple characteristic peaks generally present (Fig. 7, Table 6). First, OC has distinct spectral characteristics in the near-infrared region (Viscarra Rossel and Behrens, 2010), which are mainly attributed to the C–H of hydrocarbyl groups (1340–1380 nm), as well as the N–H of amide groups and the O–H of hydroxyl groups (1860–1900 nm) (Castaldi et al., 2018). The feature wavelengths of OC can also be related to lignin (1600–1800 nm) and cellulose (2100 nm), in addition to phenol, amide, and aliphatic groups (2300 nm) (Ben-Dor et al., 1997). Second, N is the main component of soil organic matter, so a strong correlation between the two, at 595 and 783 nm, is expected. These two wavelengths are consistent with the 550–700 nm reported by Galvao and Vitorello (1998) and the 600–800 nm reported by Ji et al. (2012) as sensitive wavelengths of organic matter. The remaining feature wavelengths of N are mainly related to the N–H of amino and amide groups (Zhang et al., 2019b).

The main feature wavelengths of P and K were relatively close (Fig. 7), which could explain the poor prediction accuracy of soil P and K contents. The wavelength positions of P and K are mainly related to iron oxides, water, and carboxyl groups (Viscarra Rossel and Webster, 2012; Volkan Bilgili et al., 2010). The feature wavelengths of CEC are similar to previous findings of Viscarra Rossel and Webster (2012), which are related primarily to iron oxides and clay minerals. The feature wavelengths of pH are mainly related to key chemical bonds, such as the O–H of carboxyl groups, the C–H of aromatic compounds, and the N–H of amino groups. The feature wavelength of $CaCO_3$ is related to the C=O near 1998 nm (Viscarra Rossel et al., 2016).

Considering soil texture, the feature wavelengths related to the four groups and 12 levels of classification were very similar (Fig. 7, Table 6). Soil texture influences soil spectral reflectance mainly through soil moisture content and particle size distribution (Bedidi et al., 1992). The soil with higher clay content absorbs more water, resulting in stronger absorption bands at 1400 and 1900 nm. Conversely, the soil with finer particles has smaller inter-particle spaces and smoother surfaces, which contributes to higher spectral reflectance. The feature wavelength of 1400 nm is mainly linked to water, and the wavelengths around 2208 and 2376 nm are principally related to clay minerals, such as kaolinite, montmorillonite, and illite (Peng et al., 2014; Castaldi et al., 2019).

Generally, CNN modeling requires a huge amount of data and computing power, and many hyperparameters need to be adjusted (LeCun et al., 2015; Zhu et al., 2017; Ng et al., 2019; Yuan et al., 2020). However, the prediction accuracy of DCNNs is usually better than that of traditional machine learning methods. Therefore, DCNNs provide useful tools for soil spectral modeling.

5. Conclusions

This study explored the modeling potential of DCNNs when applied to a large soil spectral library. Two single-task 16-layer DCNN models were successfully used to make regression predictions of seven soil properties and classification predictions of soil texture. We also assessed the effects of data pre-processing and the performance of multi-task DCNN modeling. Based on the original spectra, the DCNN models built in this study can predict most soil properties with high accuracy. Our models outperform the single-task shallow CNN architecture proposed in previous studies and other traditional machine learning methods. DCNNs do not need spectral dimension reduction for modeling soil spectral data. The application of deep learning tools has changed multiple fields of knowledge, including computer vision, natural language processing, and medical image analysis. Our study soundly demonstrates the modeling potential of deep learning for soil properties based on soil spectral data, enabling the timely adjustment of agricultural management measures and sustainable land use. This study also provides basic data for real-time quantitative monitoring of changes in soil properties, soil quality assessments, and crop yield estimations, which are also useful for attaining the goal of precision agriculture.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The LUCAS topsoil dataset used in this work was made available by the European Commission through the European Soil Data Centre managed by the Joint Research Centre (http://esdac.jrc.ec.europa.eu/). This work was supported by the National Natural Science Foundation of China (grant number: 42071068).

References

Araújo, S.R., Wetterlind, J., Demattê, J.A.M., Stenberg, B., 2014. Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. Eur. J. Soil Sci. 65 (5), 718–729. https://doi.org/10.1111/ejss.12165.

BEDIDI, A., Cervelle, B., Madeira, J., Pouget, M., 1992. Moisture effects on visible spectral characteristics of lateritic soils. Soil Sci. 153 (2), 129–141. https://doi.org/10.1097/00010694-199202000-00007.

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 Sens. Environ. 61, 1–15. https://doi.org/10.1016/S0034-4257(96)00120-4.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Brodský, L., Klement, A., Penížek, V., Kodešová, R., Borůvka, L., 2011. Building soil spectral library of the Czech soils for quantitative digital soil mapping. Soil Water Res. 6 (No. 4), 165–172. https://doi.org/10.17221/24/2011-SWR.

Burges, C.J., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Disc. 2, 121–167. https://doi.org/10.1023/A:1009715923555.

Castaldi, F., Chabrillat, S., Chartin, C., Genot, V., Jones, A.R., van Wesemael, B., 2018. Estimation of soil organic carbon in arable soil in Belgium and Luxembourg with the LUCAS topsoil database. Eur. J. Soil Sci. 69 (4), 592–603. https://doi.org/10.1111/ejss.12553.

Castaldi, F., Hueni, A., Chabrillat, S., Ward, K., Buttafuoco, G., Bomans, B., Vreys, K., Brell, M., van Wesemael, B., 2019. Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands. ISPRS. J. Photogramm. 147, 267–282. https://doi.org/10.1016/j.isprsjprs.2018.11.026.

Chang, C.W., Laird, D.A., Mausbach, M.J., Hurburgh, C.R., 2001. Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties. Soil Sci. Soc. Am. J. 65, 480–490. https://doi.org/10.2136/sssaj2001.652480x.

Chawla, N.V., Japkowicz, N., Kotcz, A., 2004. Special issue on learning from imbalanced data sets. ACM SIGKDD explorations Newslett. 6, 1–6. https://doi.org/10.1145/1007730.1007733.

Galvao, L.S., Vitorello, I., 1998. Role of organic matter in obliterating the effects of iron on spectral reflectance and colour of Brazilian tropical soils. Int. J. Remote Sens. 19 (10), 1969–1979. https://doi.org/10.1080/014311698215090.

Gogé, F., Joffre, R., Jolivet, C., Ross, I., Ranjard, L., 2012. Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. Chemometr. Intell. Lab. 110 (1), 168–176. https://doi.org/10.1016/j.chemolab.2011.11.003.

Grunwald, S., Yu, C., Xiong, X., 2018. Transferability and scalability of soil total carbon prediction models in Florida, USA. Pedosphere 28 (6), 856–872. https://doi.org/10.1016/S1002-0160(18)60048-7.

Guerrero, C., Wetterlind, J., Stenberg, B., Mouazen, A.M., Gabarrón-Galeote, M.A., Ruiz-Sinoga, J.D., Zornoza, R., Viscarra Rossel, R.A., 2015. Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? Soil Till Res. 155, 501–509. https://doi.org/10.1016/j.still.2015.07.008.

Haghi, R.K., Pérez-Fernández, E., Robertson, A.H.J., 2021. Prediction of various soil properties for a national spatial dataset of Scottish soils based on four different chemometric approaches: a comparison of near infrared and mid-infrared spectroscopy. Geoderma 396, 115071. https://doi.org/10.1016/j.geoderma.2021.115071.

Hartemink, A.E., 2015. On global soil science and regional solutions. Geoderma Reg. 5, 1–3. https://doi.org/10.1016/j.geodrs.2015.02.001.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

Hu, W., Huang, Y., Wei, L., Zhang, F., Li, H., 2015. Deep convolutional neural networks for hyperspectral image classification. J. Sens. 2015, 1–12. https://doi.org/10.1155/2015/258619.

Islam, K., Singh, B., McBratney, A., 2003. Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy. Soil Res. 41, 1101–1114. https://doi.org/10.1071/SR02137.

Ji, W., Shi, Z., Zhou, Q., Zhou, L., 2012. VIS-NIR reflectance spectroscopy of the organic matter in several types of soils. J. Infrared Millim. W. 31, 277–282. https://doi.org/10.3724/SP.J.1010.2012.00277.

Lal, R., 2004. Soil carbon sequestration impacts on global climate change and food security. Science 304 (5677), 1623–1627. https://doi.org/10.1126/science:1097396.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. https://doi.org/10.1038/nature14539.

Lipovetsky, S., Conklin, M., 2001. Analysis of regression in game theory approach. Appl. Stoch. Model. Bus. 17 (4), 319–330. https://doi.org/10.1002/asmb.446.

Liu, L., Ji, M., Buchroithner, M., 2018. Transfer learning for soil spectroscopy based on convolutional neural networks and its application in soil clay content mapping using hyperspectral imagery. Sensors 18, 3169. https://doi.org/10.3390/s18093169.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inform. Process. Syst. 4765–4774. https://arxiv.org/abs/1705.07874.

Moran, M.S., Inoue, Y., Barnes, E.M., 1997. Opportunities and limitations for image-based remote sensing in precision crop management. Remote Sens. Environ. 61 (3), 319–346. https://doi.org/10.1016/S0034-4257(97)00045-X.

Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., McBratney, A.B., 2019. Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. Geoderma 352, 251–267. https://doi.org/10.1016/j.geoderma.2019.06.016.

Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., Montanarella, L., 2014. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. Soil Biol. Biochem. 68, 337–347. https://doi.org/10.1016/j.soilbio.2013.10.022.

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., Fernández-Ugalde, O., 2018. LUCAS soil, the largest expandable soil dataset for Europe: a review. Eur. J. Soil Sci. 69 (1), 140–153. https://doi.org/10.1111/ejss.12499.

Padarian, J., Minasny, B., McBratney, A.B., 2019a. Using deep learning to predict soil properties from regional spectral data. Geoderma Reg. 16, e00198. https://doi.org/10.1016/j.geodrs.2018.e00198.

Padarian, J., Minasny, B., McBratney, A.B., 2019b. Transfer learning to localise a continental soil vis-NIR calibration model. Geoderma 340, 279–288. https://doi.org/10.1016/j.geoderma.2019.01.009.

Padarian, J., McBratney, A.B., Minasny, B., 2020. Game theory interpretation of digital soil mapping convolutional neural networks. Soil 6, 389–397. https://doi.org/10.5194/soil-2020-17.

Panagos, P., Van Liedekerke, M., Jones, A., Montanarella, L., 2012. European Soil Data Centre: response to European policy support and public data requirements. Land Use Policy 29 (2), 329–338. https://doi.org/10.1016/j.landusepol.2011.07.003.

Peng, Y., Knadel, M., Gislum, R., Schelde, K., Thomsen, A., Greve, M.H., 2014. Quantification of SOC and clay content using visible near-infrared reflectance-mid-infrared reflectance spectroscopy with Jack-Knifing partial least squares regression. Soil Sci. 179, 325–332. https://doi.org/10.1097/SS.0000000000000074.

Petersson, H., Gustafsson, D., Bergstr, D., 2016. Hyperspectral image analysis using deep learning-A review. In Proceedings of the 2016 6th International Conference on Image Processing Theory Tools and Applications (IPTA), Oulu, Finland. pp. 1-6. doi:10.1109/IPTA.2016.7820963.

Riese, F.M., Keller, S., 2019. Soil texture classification with 1D convolutional neural networks based on hyperspectral data. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. pp, 615–621. https://doi.org/10.5194/isprs-annals-IV-2-W5-615-2019.

Romero, D.J., Ben-Dor, E., Demattê, José.A.M., Souza, A.B.e., Vicente, L.E., Tavares, T. R., Martello, M., Strabeli, T.F., da Silva Barros, P.P., Fiorio, P.R., Gallo, B.C., Sato, M. V., Eitelwein, M.T., 2018. Internal soil standard method for the Brazilian soil spectral library: Performance and proximate analysis. Geoderma 312, 95–103. https://doi.org/10.1016/j.geoderma.2017.09.014.

Sanchez, P.A., Ahamed, S., Carre, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M.d.L., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vagen, T.-G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.-L., 2009. Digital Soil Map of the World. Science 325, 680–681. https://doi.org/10.1126/science.1175084.

Shi, Z., Wang, Q., Peng, J., Ji, W., Liu, H., Li, X., Viscarra Rossel, R.A., 2014. Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. Sci. China Earth Sci. 57 (7), 1671–1680. https://doi.org/10.1007/s11430-013-4808-x.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. URL: https://arxiv.org/pdf/1409.1556.pdf.

Singh, S., Kasana, S.S., 2019. Estimation of soil properties from the EU spectral library using long short-term memory networks. Geoderma Reg. 18, e00233. https://doi.org/10.1016/j.geodrs.2019.e00233.

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Chapter five-visible and near infrared spectroscopy in soil science. Adv. Agron. 107, 163–215. https://doi.org/10.1016/S0065-2113(10)07005-7.

Stevens, A., Nocita, M., Tóth, G., Montanarella, L., van Wesemael, B., Chen, H.YH., 2013. Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. PLoS ONE 8 (6), e66409. https://doi.org/10.1371/journal.pone.0066409.

Tóth, G., Jones, A., Montanarella, L., 2013. The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. Environ. Monit. Assess. 185 (9), 7409–7425. https://doi.org/10.1007/s10661-013-3109-3.

Tsakiridis, N.L., Keramaris, K.D., Theocharis, J.B., Zalidis, G.C., 2020. Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network. Geoderma 367, 114208. https://doi.org/10.1016/j.geoderma.2020.114208.

Tziolas, N., Tsakiridis, N., Ogen, Y., Kalopesa, E., Ben-Dor, E., Theocharis, J., Zalidis, G., 2020. An integrated methodology using open soil spectral libraries and Earth Observation data for soil organic carbon estimations in support of soil-related SDGs. Remote Sens. Environ. 244, 111793. https://doi.org/10.1016/j.rse.2020.111793.

Veres, M., Lacey, G., Taylor, G.W., 2015. Deep learning architectures for soil property prediction. In: 2015 12th Conference on Computer and Robot Vision. pp. 8-15. doi: 10.1109/CRV.2015.15.

Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158 (1–2), 46–54. https://doi.org/10.1016/j.geoderma.2009.12.025.

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B.G., Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L., Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C.B., Knadel, M., Morrás, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C., Winowiecki, L.A., Hong, S.Y., Ji, W., 2016. A global spectral library to characterize the world's soil. Earth-Sci. Rev. 155, 198–230. https://doi.org/10.1016/j.earscirev.2016.01.012.

Viscarra Rossel, R.A., Webster, R., 2012. Predicting soil properties from the Australian soil visible-near infrared spectroscopic database. Eur. J. Soil Sci. 63, 848–860. https://doi.org/10.1111/j.1365-2389.2012.01495.x.

Volkan Bilgili, A., van Es, H.M., Akbas, F., Durak, A., Hively, W.D., 2010. Visible-near infrared reflectance spectroscopy for assessment of soil properties in a semi-arid area of Turkey. J. Arid Environ. 74 (2), 229–238. https://doi.org/10.1016/j.jaridenv.2009.08.011.

Wetterlind, J., Stenberg, B., Söderström, M., 2010. Increased sample point density in farm soil mapping by local calibration of visible and near infrared prediction models. Geoderma 156 (3-4), 152–160. https://doi.org/10.1016/j.geoderma.2010.02.012.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58 (2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

Xu, Z., Zhao, X., Guo, X., Guo, J., 2019. Deep learning application for predicting soil organic matter content by VIS-NIR spectroscopy. Comput. Intel. Neuros. 2019 (1–11), 1. https://doi.org/10.1155/2019/3563761.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J., Zhang, L., 2020. Deep learning in environmental remote sensing: Achievements and challenges. Remote Sens. Environ. 241, 111716. https://doi.org/10.1016/j.rse.2020.111716.

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

Zhang, X., Lin, T., Xu, J., Luo, X., Ying, Y., 2019a. DeepSpectra: An end-to-end deep learning approach for quantitative spectral analysis. Anal. Chim. Acta. https://doi.org/10.1016/j.aca.2019.01.002.

Zhang, L., Zhang, L., Du, B., 2016. Deep learning for remote sensing data: a technical tutorial on the state of the art. IEEE Geosc. Rem. Sen. M. 4 (2), 22–40. https://doi.org/10.1109/MGRS.2016.2540798.

Zhong, L., Guo, X., Guo, J., Han, Y., Zhu, Q., Xiong, X., 2020. Soil texture classification of hyperspectral based on data mining technology. Sci. Agric. Sin. 53, 4449–4459. https://doi.org/10.3864/j.issn.0578-1752.2020.21.013.

Zhong, L., Guo, X., Guo, J., Xu, Z., Zhu, Q., Ding, M., 2021. Hyperspectral estimation of organic matter in red soil using different convolutional neural network models. Trans. CSAE. 37, 203–212. https://doi.org/10.11975/j.issn.1002-6819.2021.01.025.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geosc. Rem. Sen. M. 5 (4), 8–36. https://doi.org/10.1109/MGRS.2017.2762307.