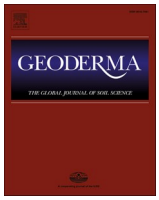
Geoderma 437 (2023) 116584

Contents lists available at ScienceDirect

Geoderma

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



Spectral fusion modeling for soil organic carbon by a parallel input-convolutional neural network

Yongsheng Hong [a], Songchao Chen [b], Bifeng Hu [c], Nan Wang [a], Jie Xue [a], Zhiqing Zhuo [d], Yuanyuan Yang [e], Yiyun Chen [f], Jie Peng [g], Yaolin Liu [f], Abdul Mounem Mouazen [h,i], Zhou Shi [a,*]

[a] Institute of Agricultural Remote Sensing and Information Technology Application, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China
[b] ZJU-Hangzhou Global Scientific and Technological Innovation Center, Hangzhou 311200, China
[c] School of Tourism and Urban Management, Jiangxi University of Finance and Economics, Nanchang 330013, China
[d] Institute of Digital Agriculture, Zhejiang Academy of Agricultural Sciences, Hangzhou 310021, China
[e] School of Spatial Planning and Design, Zhejiang University City College, Hangzhou 310015, China
[f] School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China
[g] College of Plant Sciences, Tarim University, Alar 843300, China
[h] Department of Environment, Ghent University, Coupure Links 653, 9000 Gent, Belgium
[i] Department of Agricultural Engineering and Safety, Faculty of Engineering, VYTAUTAS MAGNUS UNIVERSITY, Lithuania

ARTICLE INFO

Handling Editor: Budiman Minasny

Keywords:
Soil analysis
Visible-to-near-infrared spectroscopy
Mid-infrared spectroscopy
Data fusion
Deep learning

ABSTRACT

Visible-to-near-infrared (vis–NIR) and mid-infrared (MIR) spectroscopy have been widely utilized for the quantitative estimation of soil organic carbon (SOC). The fusion of vis–NIR and MIR data can be hypothesized to provide accurate and reliable prediction for SOC because spectral data within a specific range of each individual sensor may lack important absorptive features associated with SOC. In this study, six data fusion strategies, principally direct concatenation-partial least squares regression (DC-PLSR), outer product analysis-PLSR (OPA-PLSR), OPA-competitive adaptive reweighted sampling-PLSR (OPA-CARS-PLSR), sequentially orthogonalized-PLSR (SO-PLSR), DC-convolutional neural network (DC-CNN), and parallel input-CNN (PI-CNN), were compared for the spectral estimations of SOC. The data fusion and individual sensor models were developed using soil samples collected from Zhejiang Province, East China, and scanned under laboratory conditions with both vis–NIR and MIR spectrophotometers. The validation results of vis–NIR (validation coefficient of determination [R^2] = 0.63–0.73) were generally better than those of MIR (validation R^2 = 0.45–0.59). For data fusion, the best validation accuracy was achieved by the PI-CNN (validation R^2 = 0.84), followed in descending order by DC-CNN (validation R^2 = 0.78), SO-PLSR (validation R^2 = 0.73), OPA-CARS-PLSR (validation R^2 = 0.69), OPA-PLSR (validation R^2 = 0.66), and DC-PLSR (validation R^2 = 0.64). The better performance of PI-CNN over DC-CNN demonstrates the necessity of using different sizes of convolutional kernels before feeding into the fully connected layers in the CNN network for fusing vis–NIR and MIR spectral data. The deep-learning fusion method based on PI-CNN can be considered an efficient tool for integrating data from multiple sensors for estimating soil properties in the field of soil spectral modeling.

1. Introduction

As a key component of soil, soil organic carbon (SOC) affects various physical and chemical properties of the soil (Lal, 2016). It plays a central role in soil fertility, affecting crop production and food security (Zhao et al., 2018). Slight changes in SOC content can cause severe disturbances to the terrestrial carbon cycle and global climate change (Minasny et al., 2017; Ramesh et al., 2019; Wiesmeier et al., 2019). Constrained by the high temporal and spatial variability of soils resulting from multiple soil-forming factors, such as climate, topography, parent materials, and human management practices throughout time, there are still some challenges in accurately quantifying SOC concentrations at different geographical scales (Wang et al., 2022). Therefore, the development of accurate, inexpensive, fast and efficient SOC

* Corresponding author.
E-mail address: shizhou@zju.edu.cn (Z. Shi).

https://doi.org/10.1016/j.geoderma.2023.116584
Received 24 August 2022; Received in revised form 25 May 2023; Accepted 27 June 2023
Available online 29 June 2023

estimation approaches is crucial for monitoring soil quality and function and for timely feedback to global warming and improvement in ecosystem services.

Proximal soil sensors, such as visible-to-near-infrared (vis–NIR) and mid-infrared (MIR) spectrometers, have been widely utilized as a cost-effective and efficient alternative in the quantitative estimation of SOC (Hong et al., 2022; Hutengs et al., 2019; Ng et al., 2019b; Poppiel et al., 2022). Spectral estimations of SOC are dependent on the chromophres, and the overtones and combination absorptions of O–H, C–H, and N–H that exist within both vis–NIR and MIR spectral ranges (Nocita et al., 2015; Stenberg et al., 2010; Terra et al., 2019; Viscarra Rossel and Behrens, 2010). Due to the inherent differences in detector type, spectral range and resolutions, each sensor has its own advantages and disadvantages affecting the accuracy of predicting SOC. However, the spectral range may be the most important factor, mainly because the spectral data in a specific range may contain insufficient spectral information associated with SOC (Ng et al., 2019b). Despite the fact that fundamental vibrations of molecules associated with SOC take place in the MIR range, some literature reported better modeling performances of MIR over vis–NIR (Hong et al., 2022; Hutengs et al., 2019; Terra et al., 2019), while other studies reported contradictory results (Allo et al., 2020; McDowell et al., 2012). It is thus worth exploring modeling approaches that make full use of the complementary information from different sensors to maximize the predicted stability of SOC. Data fusion is an effective strategy to integrate multiple data sources from different sensors (Borràs et al., 2015; Gholizadeh et al., 2021a; Grunwald et al., 2015). The main purpose of data fusion modeling is to achieve accurate prediction by integrating information from multiple sources (Javadi and Mouazen, 2021; Munnaf and Mouazen, 2022).

Various data fusion strategies, such as direct concatenation (DC), feature fusion, and outer product analysis (OPA), have been introduced to fuse vis–NIR and MIR spectral data to estimate soil properties (Borràs et al., 2015). Simple DC of the vis–NIR and MIR spectra is the simplest and most commonly used method (Knox et al., 2015; Nawar et al., 2022; Wan et al., 2020). Another very promising fusion approach is the extension of the latent variable (LV) space combined with partial least squares regression (PLSR), such as the sequentially orthogonalized-PLSR (SO-PLSR) algorithm, which is the further development of the least squares-PLSR method (Biancolillo et al., 2015; Biancolillo and Næs, 2019). It is based on sequentially extracting significant features from different sensors and further allows incorporating the interactions between multiple sensors to improve the prediction accuracy and interpretations. OPA aims to generate spectral coevolutions by calculating the mutual weight of two different signals and then bringing them to the same domain (Cécillon et al., 2012; Ng et al., 2019b; Terra et al., 2019). However, OPA-fused data with large sets of predictors may lead to the amplification of spectral noise and strong collinearity among variables (Terra et al., 2019; Vohland et al., 2014), and variable selection after OPA fusion is thus indispensable. Competitive adaptive reweighted sampling (CARS) is a very rigorous procedure to select key wavebands associated with target dependent variables (Hong et al., 2020a; Li et al., 2009). However, the performance of the CARS method when integrated with OPA fusion of vis–NIR and MIR has not been explored for spectral prediction of SOC.

Recently, in the field of soil spectroscopy, deep learning algorithms, such as long short-term memory, autoencoders, and convolutional neural networks (CNNs), have been successfully employed in the spectral estimation of soil properties, usually with excellent predictive performance (Javadi et al., 2021; Ng et al., 2019b; Tsakiridis et al., 2020; Wang et al., 2022). However, the majority of existing CNN models are limited to using data from only one individual-sensor (Pullanagari et al., 2021; Shen et al., 2022); hence, there is the need to explore data fusion techniques from multiple sensors. A potential approach is to modify the built-in functions of the CNN architecture to simultaneously allow data from two sensors to be received (i.e., parallel input-CNN; PI-CNN) (Mishra and Passos, 2021a). This methodology can be implemented by first performing multiple parallel convolutional layers, followed by extracting features from each layer separately and finally passing them to the fully connected layer. The input layer can be from different data sources, such as vis–NIR, MIR, and X-ray fluorescence spectra. However, due to the unique spectral features from every data source (Hong et al., 2022; Vohland et al., 2022), using the same size of convolutional kernels for extraction of the spectral information may not be the optimal solution. It is thus necessary to individually optimize the size of the convolutional kernels for multiple PIs prior to CNN modeling. The proof-of-concept of PI-CNN was originally proposed in chemometrics (Mishra and Passos, 2021a), but its extended application in the domain of soil spectroscopy in fusing vis–NIR and MIR for SOC prediction has not been reported.

The main objective of this research was to compare spectral estimations of SOC using vis–NIR, MIR, and their fused data. Our hypothesis was that fusing vis–NIR and MIR spectra by PI-CNN could improve the modeling performance, compared to predictions made by single-sensor and commonly used data fusion models.

2. Materials and methods

2.1. Soil sampling and laboratory analysis

The study area was located in Zhejiang Province, East China (Fig. 1). Its topography across the whole province is complex and diverse, and highlands account for approximately 75% of the province's area. The terrain in the northern part is relatively low and flat, with fertile soils, high population density, and a highly developed economy. According to the Köppen-Geiger climate classification, the defining type in Zhejiang Province belongs to Cfa (i.e., a temperate zone with hot summers and no dry season) (Beck et al., 2018). The annual average temperature and precipitation are approximately 15–18℃ and 980–2000 mm, respectively. This province has more than eight different soil types (Fig. 1), with paddy soils, purplish soils, and red soils being the dominant soil orders, according to Chinese Soil Taxonomy (Shi et al., 2006).

Field sampling was designed based on the Second National Soil Survey of China and knowledge from professional pedologists. In 2009–2012, a total of 146 soil profiles were collected within the province, thus resulting in 591 soil samples depending on the genetic soil horizon classifications (Shi et al., 2006). These samples were brought back to the laboratory for air-drying, grinding, and sieving to less than 2 mm. Stones, roots, and voids were carefully excluded. Every sample was then divided into two parts, with one part used for spectral measurement and the other for soil chemical analysis. Soil texture (i.e., % clay, silt, and sand) and SOC content were determined by the pipette method and H_2SO_4–$K_2Cr_2O_7$ oxidation method, respectively.

2.2. Spectral scanning and preprocessing

The vis–NIR spectra were obtained using a Fieldspec® ProFR spectrometer equipped with a contact probe and built-in light source, covering a spectral range from 350 to 2500 nm, with a final resampling resolution of 1 nm (Analytical Spectral Devices, Boulder, CO, USA). A Spectralon panel was utilized for calibrating and optimizing the spectrometer before and during the spectral measurement. The pretreated samples were placed in plastic containers that were 5 cm in diameter and 1 cm in height, and the sample surfaces were carefully leveled before scanning. For each sample, three positions uniformly distributed over the sampling surface were selected for scanning to minimize the spectral randomness, and subsequently, 10 internal scans were collected at each position. Finally, the 30 spectral curves were averaged in one final spectrum, representing a soil sample. The vis–NIR reflectance ($\boldsymbol{R}$) was converted to absorbance using the equation $\log10(1/\boldsymbol{R})$.

The MIR spectral data were measured in diffuse reflectance using a handheld Agilent 4300 FTIR spectrometer with a spectral range and resolution of 4000–650 cm^{-1} and 4 cm^{-1}, respectively (Agilent

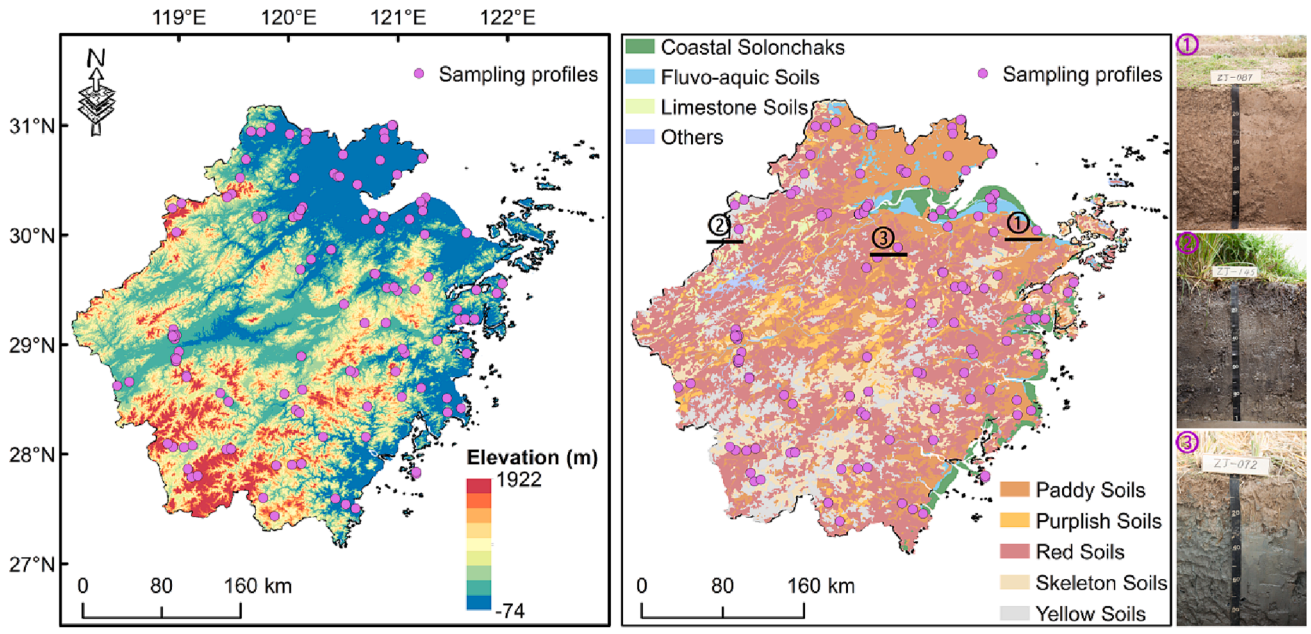


Fig. 1. Study area (Zhejiang Province, China) and location of the soil sampling profiles.

Technologies, Santa Clara, CA, USA). The soil sample was placed in a circular container 2 cm in diameter (containing approximately 2 g of soil). The surface of every sample was flattened gently without smearing before taking the measurement. After every 10 min of measurement, the spectrometer was calibrated using a gold-plated reference cap. Three replicates with 64 co-added scans were collected and then averaged into one single MIR absorbance spectrum for each sample.

Regarding the vis–NIR, only data within the 400–2400 nm range were retained to exclude the noisy wavebands. Both vis–NIR and MIR absorbance spectra were first pre-processed using Savitzky–Golay smoothing algorithm (11 points and 2 polynomial order) (Savitzky and Golay, 1964), followed by spectral first-derivative (FD) transformation. The biggest feature of the Savitzky–Golay filter is that it can keep the shape and width of the spectral signal unchanged while filtering out the noise. We selected the FD algorithm since spectral data between vis–NIR and MIR were different in terms of the spectral magnitude and intensity. Spectral FD can effectively remove the interferences from baselines and background noises, and isolate the overlapping peaks (Hong et al., 2020b; Xu et al., 2020). Finally, we resampled both types of data to an interval of 10 nm/cm^{-1} to reduce the spectral dimensionality (Chen et al., 2021). All these prepressing steps were performed in R with *prospectr* package (R Core Team, 2021; Stevens and Ramirez-Lopez, 2021).

2.3. Calibration and validation datasets

Before further data analysis, outliers in the entire dataset were cautiously excluded using the Monte Carlo outlier detection approach (Hong et al., 2019b; Li et al., 2012), thus retaining 566 samples. We then used a concentration ranking scheme to split the remaining 566 samples into two parts: a calibration set (75%, 425 samples) and a validation set (25%, 141 samples). This division method can ensure that the SOC range of the calibration set covers that of the validation set, and accordingly, their data distributions are generally similar (Chen et al., 2021; Hong et al., 2019a). The selection process was as follows: All 566 samples were sorted in ascending order according to the SOC content and then divided into 141 strata. In each stratum, one sample was assigned to the validation set, and the rest of the samples were chosen as the calibration set. Levene's test was calculated in R software with the *car* package to verify whether the data distributions of SOC in the calibration and validation sets were similar (Fox and Weisberg, 2011; Levene, 1960).

2.4. Model development and assessment

2.4.1. Individual sensor modeling

Both preprocessed vis–NIR and MIR spectra were subjected to PLSR and CNN algorithms to develop four full-spectrum-based models for the estimation of SOC. The calibration and validation results from two PLSR models were considered as the modeling benchmarks for subsequent comparisons to test the prediction accuracy of other individual as well as fusion models. The PLSR algorithm is the most widely used calibration technique in soil spectral analysis (Ng et al., 2022). This is mainly due to its unique superiority in handling multicollinearity or ill-conditioned matrices (Wold et al., 2001). It compresses spectral information into a selected number of LVs, which are obtained by projecting the spectra in the direction of maximizing covariance with the target variable. Leave-one-out cross-validation was employed to determine the optimal number of LVs, which led to the minimum root mean square error (RMSE) value. Variable importance in the projection (VIP) value derived from PLSR was used to evaluate the importance of spectral variables, with bands greater than 1 regarded as important for SOC prediction (Chong and Jun 2005; Wold et al., 2001). PLSR analysis was implemented in MATLAB 2019a (MathWorks, Natick, MA, USA) together with a toolbox named MBA-GUI (Mishra et al., 2020).

As a commonly used deep learning-based architecture, CNNs can automatically extract relevant features from specific data (Cui and Fearn, 2018; Mishra and Passos, 2021b). The 1-dimensional CNN architecture could yield relatively accurate results on one-dimensional data (Ng et al., 2019b) and thus was used in this paper. The CNN visualization architectures for both vis–NIR and MIR are shown in Fig. 2. Taking the vis–NIR as an example (Fig. 2a), a seven-layer network was created, including an input layer, a convolutional layer (with both filter and stride of 1), a flattening layer, three fully connected layers with 36, 18, and 12 neurons, and an output layer (e.g., SOC predicted values). Only the last layer utilized a linear activation function, whereas the other layers used exponential linear units as the activation function between the layers. We initialized the weights by using 'HeNormal' initialization and trained the model with an Adam optimizer, whose initial learning rate (LR) was set as 0.01*(batch size)/256. The loss function used here was MSE plus L2 penalty regularization. When the

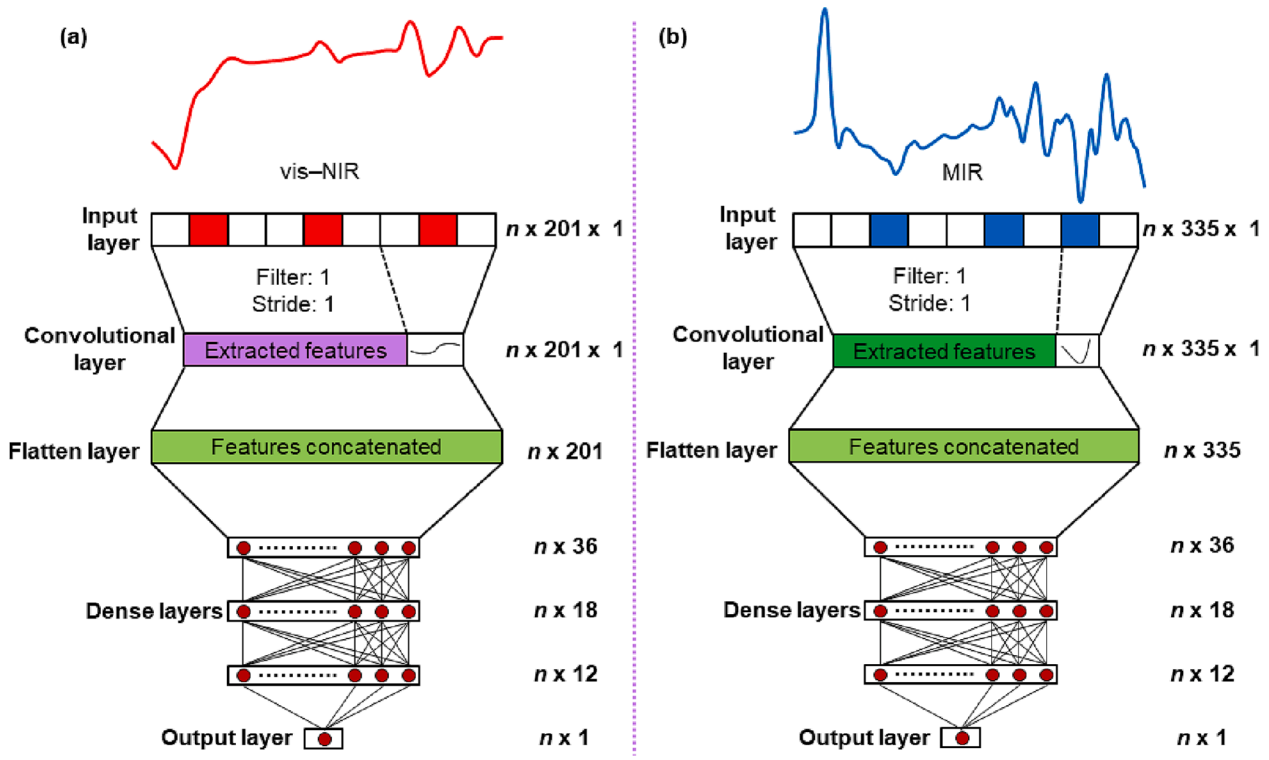


Fig. 2. Architectures of the proposed 1D-CNNs for SOC estimation: vis–NIR (a) and MIR (b).

validation loss did not improve by 10^5 after 25 epochs, the LR value was iteratively reduced by a factor of 2. The maximum number of epochs was set to 1000, and an automatic early stopping algorithm was applied to avoid overfitting. Regarding the determination of the optimal kernel size, eight different combinations with kernel sizes in the intervals of 2, 3, 4, 5, 6, 7, 8, and 9 were tested using a fivefold cross-validation method performed on the calibration set (Tsakiridis et al., 2020). For the MIR data (Fig. 2b), the same processing parameters and steps were strictly followed. The CNN algorithm was implemented in Python (Python Software Foundation, 2019) using *Keras* (Chollet, 2019) and *TensorFlow* (Abadi et al., 2019) running on a graphics processing unit cloud-computing server. Spectral variable importance (i.e., Shapley additive explanations; SHAP) from the CNN model was interpreted using the *SHAP* library (Lundberg and Lee, 2017).

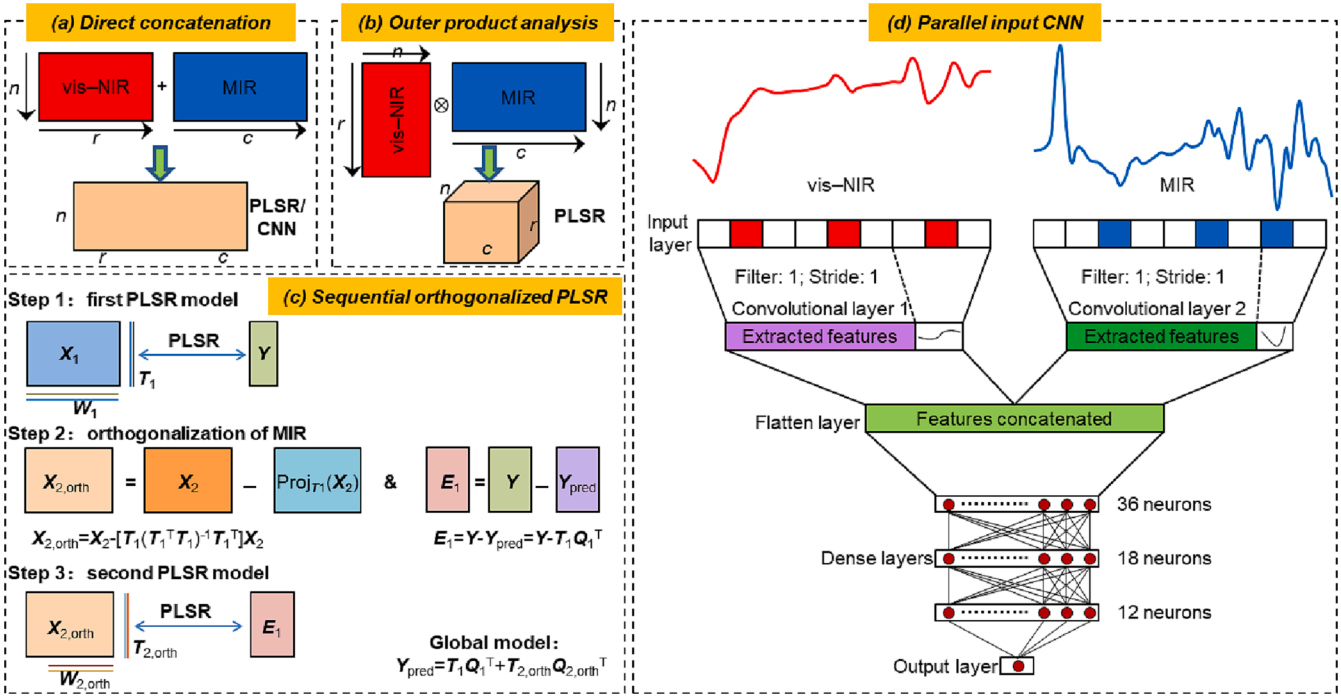


Fig. 3. Illustrations of four spectral fusion strategies. Note: X_1: vis–NIR; Y: SOC; W: weights; T: scores; X_2: MIR; E: SOC-residuals; Q: loadings; n: sample size; r: vis–NIR spectral dimension; c: MIR spectral dimension.

2.4.2. Spectral fusion modeling

Six data fusion strategies for combining vis–NIR and MIR spectra, including DC-PLSR, OPA-PLSR, OPA-CARS-PLSR, SO-PLSR, DC-CNN, and PI-CNN, were explored in this work. Conceptual diagrams of four of them are explained and illustrated in detail in Fig. 3.

(1) DC-PLSR (Hong et al., 2022; Wan et al., 2020): Simple spectral concatenation of vis–NIR and MIR data was first carried out for each sample (Fig. 3a), and then the fused spectral data were input into the PLSR algorithm for further modeling. The DC approach is the simplest and most straightforward data fusion strategy.

(2) OPA-PLSR (Fig. 3b): The OPA algorithm emphasizes the common evolution of different spectra by merging the spectral data obtained by different sensors (Jaillais et al., 2005; Terra et al., 2019). For the same sample size n, all spectral data of vis–NIR (sensor 1: r dimensions) are multiplied by all spectral values of MIR (sensor 2: c dimensions), thus producing n outer product matrices (r rows $\times$ c columns) between spectral data in both spectral domains. The spectral intensities of the resulting matrix depend upon the intensities of the original spectral data. The $r \times c$ matrix is then unfolded to an $r \times c$ vector. For n samples, the unfolded process generates a new matrix (i.e., n rows and $r \times c$ columns), from which the regression models (e.g., PLSR in this study) between SOC and the fused spectral data can be developed. Moreover, the VIP can also be derived, and its dimension is the same as that of the $r \times c$ vector. The OPA was run in R software (R Core Team, 2021). The OPA method generates a large number of predictors, which would complicate the modeling structure and lengthen the running time.

(3) OPA-CARS-PLSR: The CARS algorithm is a widely utilized variable selection method based on the rule of "survival of the fittest" (Li et al., 2009). It first develops a complete model containing all variables and then eliminates the least important variables iteratively and continuously in a backward way. In each iteration, CARS adopts a strategy of evaluating the importance performance of a subset with multiple variables rather than a single variable. Finally, the subset with the lowest cross-validation error is selected as the optimal subset. Readers can find the detailed steps of the CARS algorithm from Li et al. (2009) and Li et al. (2018). A simplified version of CARS was used to ensure that the selected variables were reproducible and stable, and all the procedures were performed in MATLAB 2019a using the libPLS library (Li et al., 2018). The number of sampling runs was set to 50.

(4) SO-PLSR (Fig. 3c): The main principle behind this method is to perform sequential incorporation of multiple predictor matrices (Biancolillo and Næs, 2019). It evaluates the incremental contributions from different sensors and removes redundant information through an orthogonalization procedure (Mishra et al., 2021b). For only two sensors (i.e., two predictor matrices), the main calculation steps are as follows: First, a PLSR model is built to relate SOC to the $\boldsymbol{X_1}$ matrix (from vis–NIR) to obtain the scores $\boldsymbol{T_1}$ matrix. Then, the PLSR model between SOC-residuals and the orthogonalized $\boldsymbol{X_2}$ matrix (from MIR) is also established. Finally, ordinary least-squares regression between SOC and the concatenated scores from the $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ matrices is established. The sequential feature along with the orthogonalization process allows direct assessment of whether data fusion of two sensors can improve the model accuracy. SO-PLSR was conducted in MATLAB 2019a using the MBA-GUI toolbox (Mishra et al., 2020). The advantages of SO-PLSR are that it can effectively deal with multisensors with more than one underlying dimension and multicollinearity among predictors.

(5) DC-CNN: Similar to the DC-PLSR method, the CNN model was constructed after concatenation of vis–NIR and MIR spectral data. The DC-CNN model was also executed to optimize the kernel size in the convolutional layer via a grid-searching method, with the same parameter intervals previously mentioned in Section 2.4.1. Other modeling parameters were set the same as listed in Section 2.4.1 (i.e., CNN on individual spectroscopic matrices).

(6) PI-CNN (Fig. 3d): For data fusion analysis, the 1D-CNN framework is adjusted to allow parallel input of multiple data sources so that different spectral features can be extracted simultaneously. Two convolutional layers consisting of two individual receptor fields (with both filter and stride of 1) were modified in the first two layers. Other parameter settings and processing steps in PI-CNN were set to be the same as those in the single sensor-based CNN architecture defined in Section 2.4.1. The kernel sizes in the individual vis–NIR or MIR range needed to be well-tuned. To achieve this goal, we optimized them in a grid-searching manner with both kernel sizes from vis–NIR and MIR changing in the space of [2, 3, 4, 5, 6, 7, 8, and 9]. The optimal kernel sizes were identified according to the lowest RMSE value obtained in the calibration set under a fivefold cross-validation scheme. Models with the optimal hyperparameters were then tested using the validation set. One advantage of PI-CNN is that it can perform specific convolutions for each data source individually.

2.4.3. Model evaluation

Three indicators, including the coefficient of determination (R^2), RMSE, and ratio of performance to interquartile range (RPIQ) (Bellon-Maurel et al., 2010), were collectively used to evaluate the models' predictive performance during calibration and validation processes (Chen et al., 2022). R^2 denotes the amount of variance explained by the model. The RMSE is the standard deviation of the residuals between the measured and estimated SOC values. The RPIQ is calculated by dividing the interquantile distance of the measured SOC values by the RMSE. Large R^2 and RPIQ and small RMSE values represent a good model.

3. Results

3.1. SOC and spectral data

The SOC in the entire dataset varied from 0.05% to 5.14%, with a mean and standard deviation of 0.88% and 0.77%, respectively (Table 1). The SOC range for the calibration set (0.05–5.14%) was larger than that for the validation set (0.07–3.93%). Levene's test yielded a p value of 0.73 at the 5% significance level, thus indicating no difference in SOC distribution between calibration and validation sets, which can be representative of the entire dataset (Fig. 4a). Except for the silt and sandy clay soil textures, the 566 samples covered all other soil texture types (Fig. 4b). Moreover, most of the samples were of the loam texture.

After FD processing, spectral absorption features of both vis–NIR and MIR data were enhanced (Fig. 5a and b), which included in particular the valleys at 520, 1450, 1950, and 2230 nm in the vis–NIR data and peaks near 3710, 1720, and 1020 cm^{-1} in the MIR data. However, absorbance features (e.g., at 520 nm) in the vis–NIR region, had relatively broader peaks compared to the thinner peaks (e.g., at approximately 3710 cm^{-1}) in the MIR range. Therefore, in practical situations, using the same convolutional kernel size for both vis–NIR and MIR may not be a successful solution, necessitating exploration of optimal convolutional kernel sizes suitable for both spectral data. Over the full-spectrum domains, spectral correlations for vis–NIR and MIR varied within −0.56–0.34 and −0.29–0.48, respectively (Fig. 5c and d).

3.2. Data fusion analysis

With OPA fusion using 201 vis–NIR wavebands and 335 MIR wavebands, a total of 67,335 combined wavebands were generated across the common spectral domain (Fig. 6a). The shape of the absorbance spectra after OPA fusion was overall consistent with the inverse of the FD absorbance of MIR spectra shown in Fig. 5b, because the spectral intensity of MIR was generally higher than that of vis–NIR. From the matrix representation (Fig. 6b), in the vis–NIR spectral range, high spectral coevolutions were located at approximately 400–800 nm, and the fused absorbance was apparent along the MIR domain with large coevolutions occurring at 3710, 3000, and 1230 cm^{-1}. The correlation between SOC and the fused data in the fused spectral range (Fig. 6c) was higher than those in both the vis–NIR and MIR individual spectral ranges. Relatively large correlation values were located in the vis–NIR

Table 1
Descriptive statistics for measured soil organic carbon (SOC).

Dataset	Sample size	Min (%)	1st quartile	3rd quartile	Max (%)	Mean (%)	Median (%)	Standard deviation (%)	Coefficient variation (%)
Entire	566	0.05	0.33	1.21	5.14	0.88	0.60	0.77	87.76
Calibration	425	0.05	0.33	1.21	5.14	0.89	0.60	0.79	88.77
Validation	141	0.07	0.33	1.20	3.93	0.87	0.59	0.73	84.74

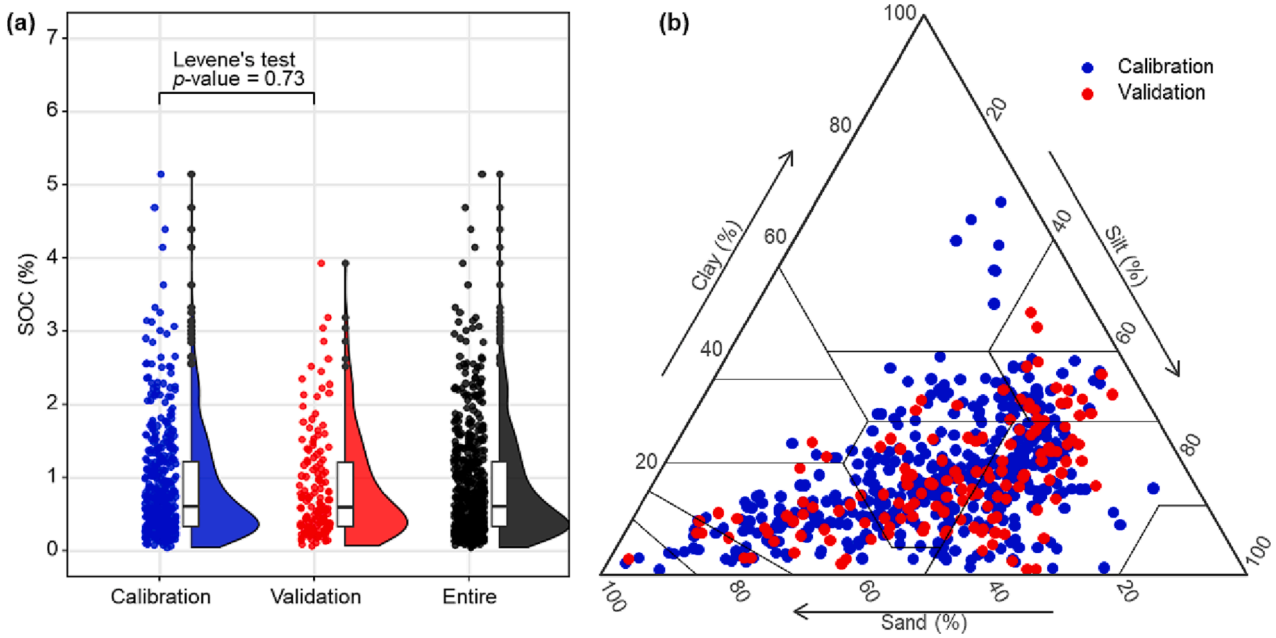


Fig. 4. Raincloud plots of SOC for three different datasets (a) and soil texture triangles for the calibration and validation datasets (b).

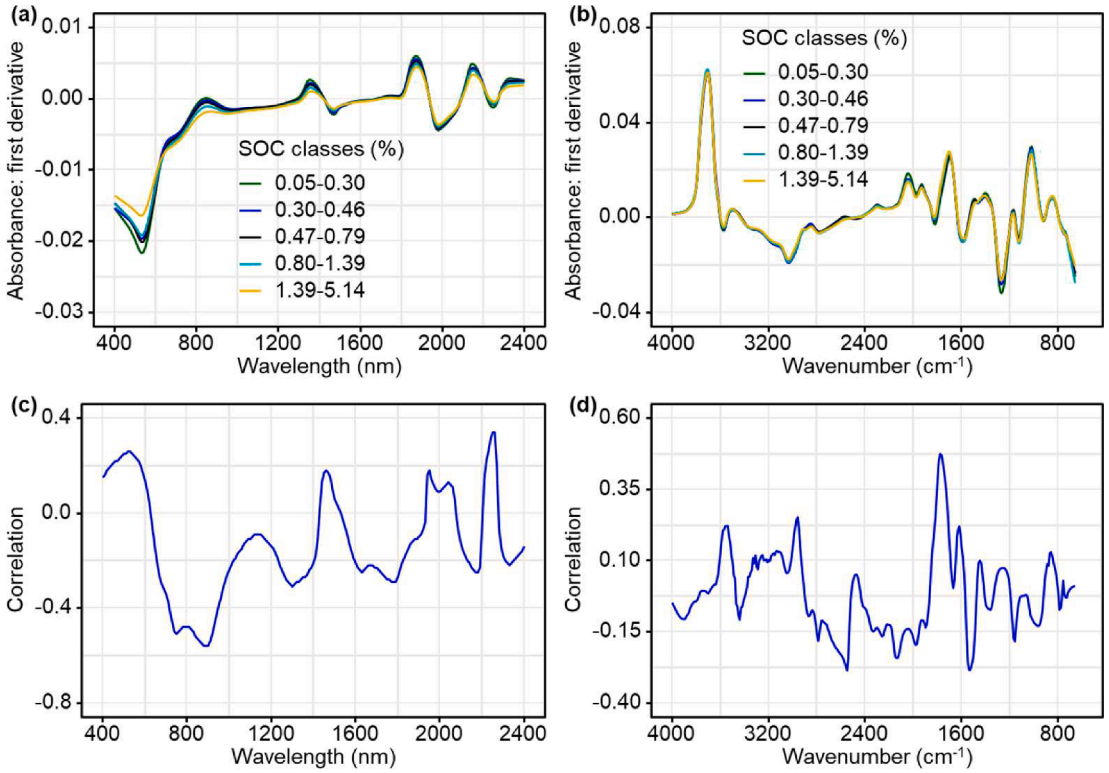


Fig. 5. First-derivatives of vis–NIR (a) and MIR absorbance (b) under five SOC classes, and spectral correlations for vis–NIR (c) and MIR (d).

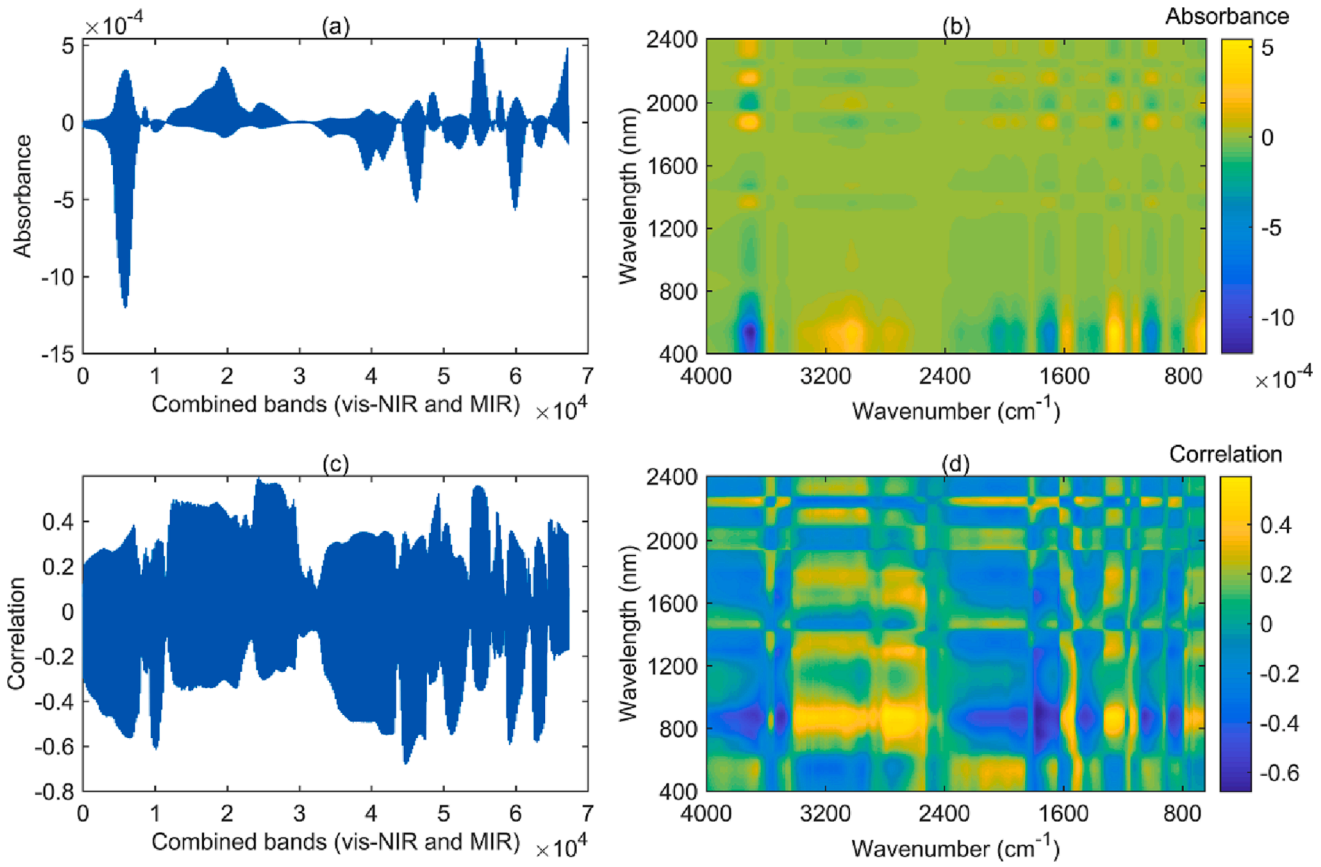


Fig. 6. Fused results by OPA: unfold absorbance (a) and folded absorbance matrix (b), and unfold spectral correlation (c) and folded correlation matrix (d).

range of 600–1000 nm and in the MIR domain of 4000–3400, 2400–1600, and 1000–800 cm^{-1} (Fig. 6d).

The output of the CARS algorithm applied to select the most informative variables for SOC after OPA fusion (Fig. 7) showed that with an increase in the sampling run, the uninformative wavelengths were eliminated in a gradual manner (Fig. 7a). By analyzing Fig. 7b, when the number of sampling runs was equal to 31, the cross-validation RMSE (RMSECV) reached the lowest value (i.e., the position marked by the arrow). The selected variables were mainly codistributed within the vis–NIR range at 400–800 and 1900–2350 nm and the MIR range at 3800–2800 cm^{-1} (Fig. 7c).

The outputs of the SO-PLSR analysis are presented in Fig. 8. The minimal cross-validation error was achieved when 15 LVs were used in the fusion model (i.e., the black point in Fig. 8a), of which 10 LVs were determined for the vis–NIR range and 5 LVs for the MIR range. The regression vectors fluctuated between positive and negative values (Fig. 8b and c), indicating that the relations between SOC and vis–NIR and MIR spectra were not stable at each waveband.

3.3. SOC modeling

The optimal kernel sizes were well tuned in the convolutional layers for the DC-CNN and PI-CNN models (Fig. 9). A kernel size with a convolutional width of 6 was identified as the best for DC-CNN, as it resulted in the minimal RMSECV value. Similarly, with respect to the PI-CNN model, the lowest RMSECV value was achieved with kernel sizes of 8

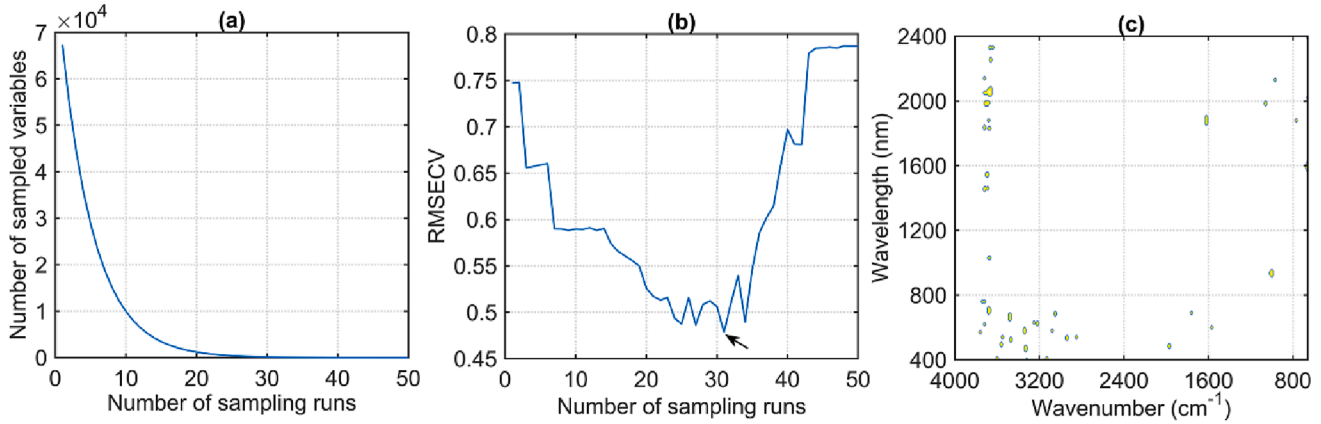


Fig. 7. Fused results by OPA-CARS: the number of sampled variables in response to each sampling run (a), change of RMSECV values in every sampling run (b), and CARS selected variables (c).

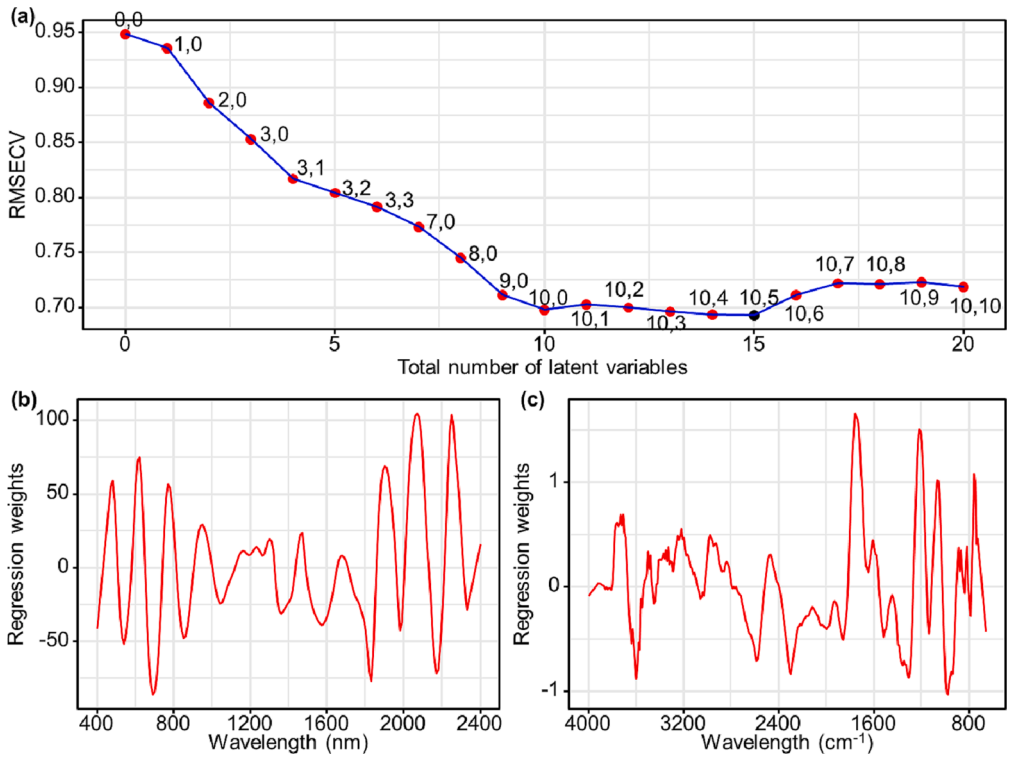


Fig. 8. Fused results by SO-PLSR: cross-validation RMSE (RMSECV) and the number of latent variables selected for vis–NIR and MIR (a), regression vector for vis–NIR (b), and regression vector for MIR (c).

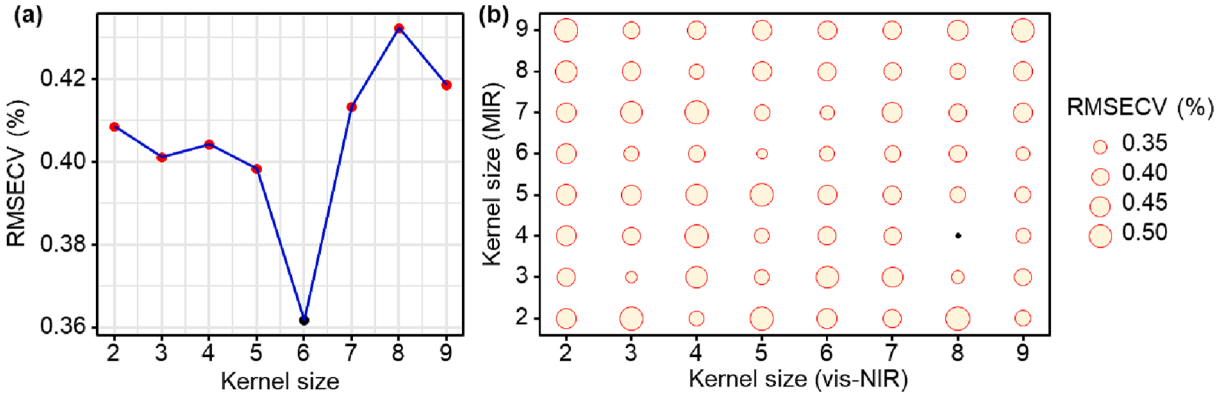


Fig. 9. Parameter optimization for kernel size in different CNN architectures: DC-CNN (a) and PI-CNN (b). The optimal values were identified as black points in both subplots.

and 4 for vis–NIR and MIR, respectively.

After optimizing the internal parameters in different CNNs, a total of ten models were developed for SOC prediction before and after data fusion (Table 2). For a single sensor, the calibration and validation results of vis–NIR (RPIQ = 2.00–2.32) were better than those of MIR (RPIQ = 1.63–1.87) for both PLSR and CNN models. In addition, the prediction accuracy of the CNN model (RPIQ = 1.87–2.32) was superior to that of the PLSR model (RPIQ = 1.63–2.00) for both vis–NIR and MIR spectral data.

Among the data fusion models (Table 2), the validation accuracy of DC-PLSR (validation R^2 = 0.64) was improved compared with that of two PLSR models built based on only vis–NIR or MIR. However, the calibration accuracy of the DC-PLSR model (calibration R^2 = 0.59) was lower than that of the vis–NIR-based PLSR model (calibration R^2 = 0.61). The OPA-PLSR model achieved a calibration R^2 of 0.67 and a validation R^2 of 0.66 for SOC estimation, which was better than the modeling accuracies from two single-sensor PLSR models (i.e., vis–NIR-PLSR and MIR-PLSR). Compared with OPA-PLSR, the calibration and validation R^2 values of the OPA-CARS-PLSR model were improved by 0.16 and 0.03, respectively. However, the accuracy gap between calibration and validation in OPA-CARS-PLSR was higher than that in OPA-PLSR, which reflects the instability in prediction of the former model. In both calibration and validation, the SO-PLSR model performed better than or comparable to the four single-sensor models (Table 2). The two deep-learning models (i.e., DC-CNN and PI-CNN) were generally superior to the remaining eight models. The best modeling result was obtained with the PI-CNN model, providing the smallest validation RMSE value (0.29%) and the largest validation R^2 (0.84) and RPIQ (3.03) values. The scatterplots of the measured versus spectrally predicted SOC values in the validation dataset (Figs. 10 and 11) confirm this finding, as the points from the PI-CNN model were more closely scattered around the 1:1 line than those from the other nine models. Despite this, the

Table 2
Model performances reported for SOC under different modeling strategies (before and after data fusion).

Spectral data[a]	Modeling algorithm	Calibration			Validation			
		R^2[b]	RMSE (%)[c]		R^2[b]	RMSE (%)[c]	RPIQ[d]	
Individual input	vis–NIR	PLSR	0.61	0.49		0.63	0.44	2.00
		CNN	0.74	0.41		0.73	0.38	2.32
	MIR	PLSR	0.49	0.56		0.45	0.54	1.63
		CNN	0.62	0.49		0.59	0.47	1.87
Data fusion	DC	PLSR	0.59	0.52		0.64	0.46	1.91
	OPA	PLSR	0.67	0.46		0.66	0.44	2.00
	OPA-CARS	PLSR	0.83	0.33		0.69	0.41	2.15
	SO	PLSR	0.74	0.40		0.73	0.38	2.32
	DC	CNN	0.81	0.34		0.78	0.35	2.51
	PI	CNN	0.85	0.30		0.84	0.29	3.03

[a] Vis–NIR: visible-to-near-infrared; MIR: mid-infrared; DC: direct concatenation; OPA: outer product analysis; OPA-CARS: OPA-competitive adaptive reweighted sampling; SO: sequential orthogonalized; PI: parallel input; PLSR: partial least squares regression; CNN: convolutional neural network.
[b] Determination coefficient.
[c] Root mean squared error.
[d] Ratio of performance to interquartile range.

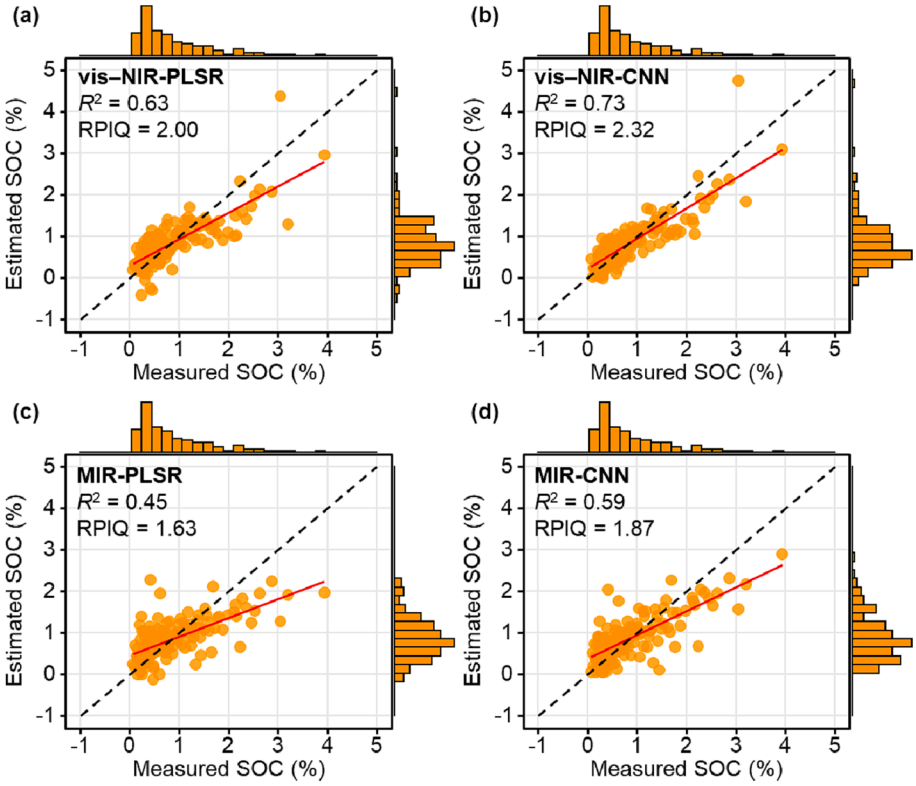


Fig. 10. Measured versus estimated SOC values for validation set with individual vis–NIR or MIR used as modeling input (before data fusion).

predicted performance of the PI-CNN model was slightly weaker at lower and higher SOC concentrations.

3.4. Interpretations of important predictors

The variable importance plots are shown in Fig. 12 for the vis–NIR-PLSR and MIR-PLSR models before data fusion and for the DC-CNN and PI-CNN models after data fusion. Overall, the PLSR models tended to select more parts of the important variables compared to the CNN models, which seemed to identify some of the most informative and important predictors with the sharpest values. For the vis–NIR-PLSR model (Fig. 12a), important spectral bands with VIP values larger than 1 were mainly located within 400–820, 1800–1880, 2090–2150, and 2380–2400 nm. For the MIR-PLSR model (Fig. 12b), wavebands with VIPs larger than 1 were observed in the ranges of 3670–3600, 3080–3010, 2150–2050, 1890–1550, and 1380–980 cm^{-1}. Comparing the SHAP values derived from both DC-CNN and PI-CNN models demonstrated that the important wavebands from the latter were almost higher than those from the former across both vis–NIR and MIR regions (Fig. 12c). The most prominent SHAP peaks occurred in the spectral ranges of 800–900 nm, 1920–1950 nm, 2200–2280 nm, 2250–2130 cm^{-1}, and 1800–1730 cm^{-1}, with some smaller peaks observed at approximately 740 nm, 1610 nm, 2180 nm, 1630 cm^{-1}, and 950 cm^{-1}.

4. Discussion

As shown in Fig. 5, MIR provided stronger and more intense absorption bands than vis–NIR, mainly because fundamental molecular vibrations occur in the MIR region (Mendes et al., 2022a; Nocita et al., 2015). The substantial difference between vis–NIR and MIR is

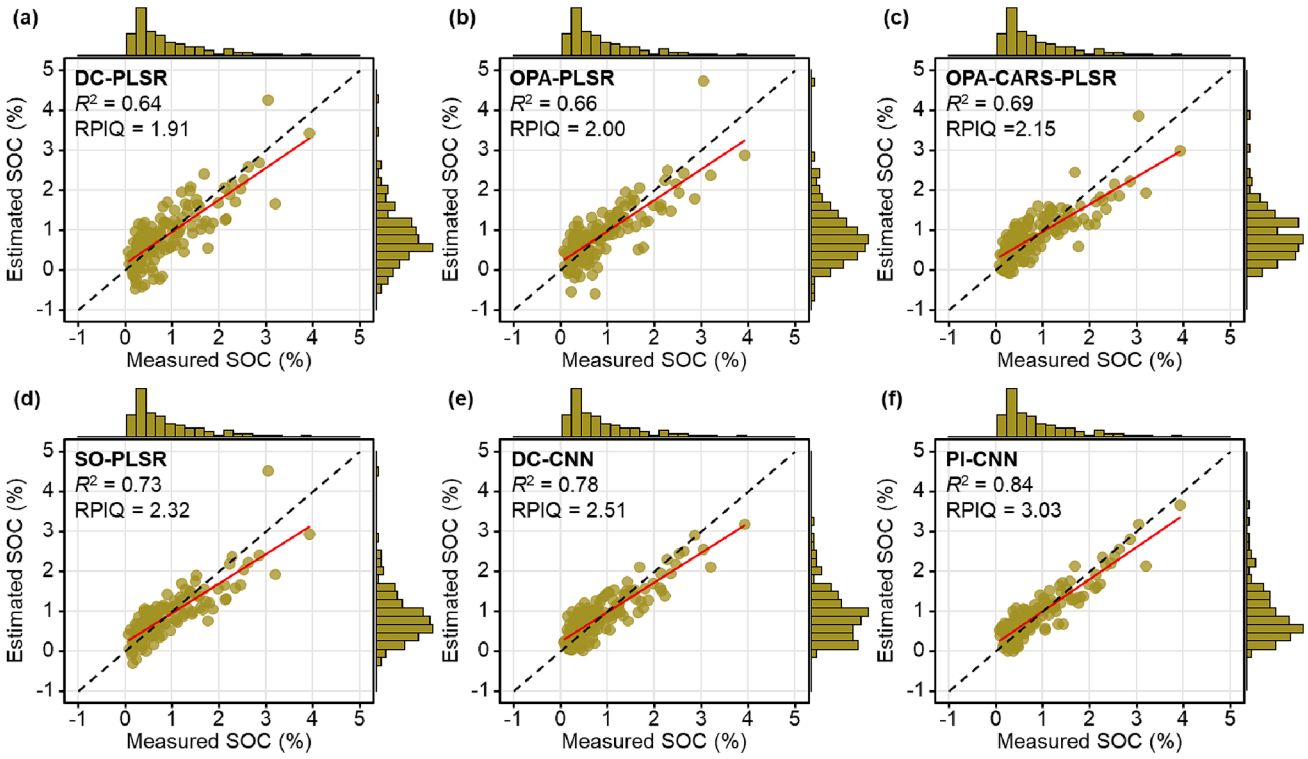


Fig. 11. Measured versus estimated SOC values for validation dataset after different spectral fusions.

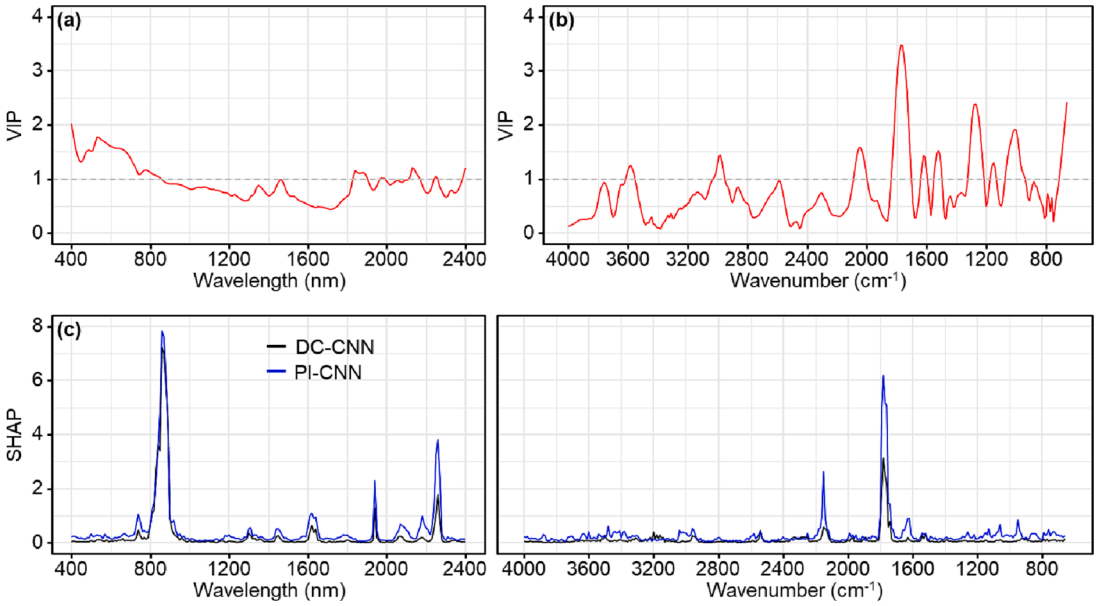


Fig. 12. Comparisons of the variable importance: variable importance in projection (VIP) extracted from vis–NIR-PLSR (a), VIP extracted from MIR-PLSR (b), and the averaged Shapley additive explanations (SHAP) extracted from DC-CNN and PI-CNN (c). In subplots (a) and (b), the important threshold of 1 is indicated by the dashed lines.

dominantly sourced from the interactive mode between soil and electromagnetic energy (Mendes et al., 2022b; Nocita et al., 2015; Stenberg et al., 2010). The vibrational absorptions in the MIR region are generally stronger and more defined; in contrast, the absorptions in the vis–NIR range are weaker, more overlapping, and less distinguishable (Clairotte et al., 2016; Nocita et al., 2015; Soriano-Disla et al., 2014). However, for the modeling results of the individual sensors (Table 2), vis–NIR (validation R^2 = 0.63–0.73) outperformed the MIR model (validation R^2 = 0.45–0.59), irrespective of the PLSR or CNN algorithms used. This trend is supported by the stronger spectral correlation of the vis–NIR data compared to that of the MIR data across the full-spectrum range (Fig. 5). Our conclusions were not in line with the findings of some previous studies (Hong et al., 2022; Hutengs et al., 2021; Vohland et al., 2014; Vohland et al., 2022). This can be attributed to several factors affecting the modeling accuracy, including different measurement conditions during vis–NIR and MIR scanning (e.g., light exposures and measurement positions) and instrument noise (Gholizadeh et al., 2021b; Nocita et al., 2015). These results demonstrate that a single spectral sensor may

suffer from the instability of the modeling structure and the prediction uncertainty in estimating SOC, which necessitates the use of data fusion to achieve more accurate and stable SOC prediction.

The DC-PLSR, OPA-PLSR, OPA-CARS-PLSR, and SO-PLSR models improved the validation accuracies for SOC prediction relative to the vis–NIR-PLSR and MIR-PLSR models (Table 2). With OPA fusion, spectral data from different sensors were mutually processed to the same domain, where spectral coevolutions between absorbance intensities were obtained, which is a straightforward and very effective approach (Terra et al., 2019; Veselá et al., 2007). In comparison with simple concatenation of spectral data from different sensors, OPA enhances the possibility of capturing spectral coevolutions more relevant to SOC. This point is exemplified by the improved spectral correlation (e.g., with the maximal absolute value of 0.68) after OPA fusion of vis–NIR and MIR in Fig. 6c and d compared to the spectral correlations reported for individual sensors in Fig. 5c and d. However, the number of wavebands after OPA fusion was close to 70,000 (Fig. 6), which will increase the running time. Moreover, the fused data constrained by relatively narrow wavebands contained noisy and redundant information, possibly causing a reduction in spectral responses. Therefore, extracting the most informative spectral predictors rather than using full-spectrum data is crucial in developing accurate prediction models after OPA fusion (Ng et al., 2019a). In Table 2, the calibration and validation R^2 values of the OPA-CARS-PLSR model were improved by 0.16 and 0.03, respectively, relative to OPA-PLSR. Hong et al. (2020a) and Vohland et al. (2014) demonstrated a positive effect of the CARS algorithm on choosing the optimal subset of predictors for high-resolution spectral data in estimating soil properties. However, the difference in the modeling accuracy between the calibration and validation datasets from the OPA-CARS-PLSR (i.e., R^2 gap = 0.14) was larger than that from the OPA-PLSR (i.e., R^2 gap = 0.01) (Table 2). This is primarily because the modeling structure of OPA-CARS-PLSR is less stable (with fewer input variables) than that of OPA-PLSR.

The SO-PLSR retains the original natures of PLSR. That is, it can effectively handle the problem of ill-conditioned matrices among independent variables and can also capture the complementary LV information from each sensor (Biancolillo and Næs, 2019). In addition, the SO-PLSR model is independent of different variances originating from each sensor because the vis–NIR and MIR data involved in PLSR are mutually orthogonal (Biancolillo et al., 2015; Mishra et al., 2021a). Another benefit of using SO-PLSR is that the number of LVs in each sensor is optimized individually (Biancolillo and Næs, 2019). The results in Fig. 8 show that 10 and 5 LVs (a total of 15 LVs) for vis–NIR and MIR, respectively, were optimal for the SO-PLSR model, meaning that vis–NIR contributes more spectral information than MIR in estimating SOC. This conclusion is in line with our finding that the individual vis–NIR model outperformed the MIR model in Table 2.

Overall, both the DC-CNN and PI-CNN models outperformed the DC-PLSR, OPA-PLSR, OPA-CARS-PLSR, and SO-PLSR models in predicting SOC (Table 2). Compared with the linear features of the PLSR model, the CNN models can capture the nonlinearity (Cui and Fearn, 2018; Ng et al., 2019b), thus performing better, as SOC was reported to have nonlinear spectral features (Tekin et al., 2012). The nonlinearity may be sourced from both scattering and absorption features in the soil spectra (Nocita et al., 2015). The main difference between DC-CNN and PI-CNN was that the former adopted the same kernel size for both vis–NIR and MIR, while the latter used two different convolutional kernel sizes. From Fig. 5, it can be observed that the spectral absorption peaks of MIR were narrower than those of vis–NIR; hence, different kernel sizes for both vis–NIR and MIR are justified for use. Using multiple kernels in the convolutional layers also fails to ensure that each convolutional kernel is effectively optimized to the specific wavebands of all sensors. For the PI-CNN model, the optimal kernel sizes were 8 and 4 for vis–NIR and MIR, respectively, while for the DC-CNN model, its optimal kernel width was 6 (Fig. 9). The development of PI-CNN can guarantee that the kernel size of each convolutional layer is specifically optimized to vis–NIR or MIR.

The effects of the same and different kernel sizes in the convolutional layers of the CNN models on variable importance are shown in Fig. 12. Compared with the DC-CNN model, the PI-CNN model received higher importance values at the peaks at approximately 740, 1610, and 2250 nm in the vis–NIR region and at 2200 and 1700 cm^{-1} in the MIR region and thus more easily detected the local detailed features of the narrow peaks. Such enhanced features captured by PI convolutions may explain why the calibration and validation accuracies of the PI-CNN model were better than those of the DC-CNN model, as indicated in Table 2 and Fig. 11. The important spectral variables jointly determined by VIP and SHAP metrics indicated that their shared intervals in the vis–NIR region were mainly distributed within 400–900, 1800–1950, and 2100–2300 nm. These spectral ranges were associated with the presence of goethite, ferrihydrite, hematite, organic matter, N–H, C–H, carbonates, humic acid, C = O, AL–OH, kaolin, and O–H (Knadel et al., 2013; Viscarra Rossel and Behrens, 2010). In the MIR domain, they were mainly located within 3670–3600, 3080–3010, 2150–2050, 1890–1550, and 1380–980 cm^{-1}, which resulted from clay minerals, O–H, aliphatic C–H stretching, quartz, carbohydrate, silicates, C = O, aromatic and aliphatic compounds, and C–O–C (Matamala et al., 2019; Nocita et al., 2015; Vohland et al., 2014).

5. Conclusions

We presented a new data fusion strategy called PI-CNN to fuse vis–NIR and MIR data for predicting SOC. Among all the models tested for SOC prediction, the best validation accuracy was achieved by the PI-CNN, followed by the DC-CNN. One important shortcoming of the DC-CNN model is that it enforces the same kernel size for both vis–NIR and MIR spectra in the convolutional layers, which is not a suitable solution, as the absorption peaks of each spectral region are of different sizes, necessitating different kernel sizes to be adopted. The main advantage of the PI-CNN algorithm is the ability to perform specific convolutions separately to extract the relevant features of vis–NIR and MIR, thereby improving the modeling performance. Due to the flexibility of the CNN internal structure, future works can extend the PI-CNN approach to test data combinations using other instruments (e.g., laser-induced breakdown spectroscopy and X-ray fluorescence) to achieve spectral fusion for specific soil properties.

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.

Data availability

The authors do not have permission to share data.

Acknowledgments

The research is supported by National Natural Science Foundation of China (grant number: 42201058; U1901601; 41771440; 42101061) and National Key Research and Development Program of China (2018YFE0107000). Authors acknowledged the financial support received from the Research Foundation - Flanders (FWO) for Odysseus I SiTeMan Project (Nr. G0F9216N). Yongsheng Hong also gratefully acknowledges the Project funded by China Postdoctoral Science Foundation (2021 M702840).

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X.,

2019. TensorFlow: large-scale machine learning on heterogeneous systems. Software available from. https://www.tensorflow.org/.
Allo, M., Todoroff, P., Jameux, M., Stern, M., Paulin, L., Albrecht, A., 2020. Prediction of tropical volcanic soil organic carbon stocks by visible-near- and mid-infrared spectroscopy. CATENA 189, 104452.
Beck, H.E., Zimmermann, N.E., McVicar, T.R., Vergopolan, N., Berg, A., Wood, E.F., 2018. Present and future Köppen-Geiger climate classification maps at 1-km resolution. Scientific Data 5 (1), 180214.
Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.-M., McBratney, A., 2010. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. TrAC–Trend. Anal. Chem. 29 (9), 1073–1081.
Biancolillo, A., Måge, I., Næs, T., 2015. Combining SO-PLS and linear discriminant analysis for multi-block classification. Chemometr. Intell. Lab. 141, 58–67.
Biancolillo, A., Næs, T., 2019. Chapter 6 - The Sequential and Orthogonalized PLS Regression for Multiblock Regression: Theory, Examples, and Extensions. In: Cocchi, M. (Ed.), Data Handl. Elsevier, Sci. Techn, pp. 157–177.
Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., Busto, O., 2015. Data fusion methodologies for food and beverage authentication and quality assessment – A review. Anal. Chim. Acta. 891, 1–14.
Cécillon, L., Certini, G., Lange, H., Forte, C., Strand, L.T., 2012. Spectral fingerprinting of soil organic matter composition. Org. Geochem. 46, 127–136.
Chen, S., Xu, H., Xu, D., Ji, W., Li, S., Yang, M., Hu, B., Zhou, Y., Wang, N., Arrouays, D., Shi, Z., 2021. Evaluating validation strategies on the performance of soil property prediction from regional to continental spectral data. Geoderma 400, 115159.
Chen, S., Arrouays, D., Leatitia Mulder, V., Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., Hannam, J., Meersmans, J., Richer-de-Forges, A. C., Walter, C., 2022. Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. Geoderma 409, 115567.
Chollet, F., 2019. Keras. https://keras.io/.
Chong, I.-G., Jun, C.-H., 2005. Performance of some variable selection methods when multicollinearity is present. Chemometr. Intell. Lab. 78 (1), 103–112.
Clairotte, M., Grinand, C., Kouakoua, E., Thébault, A., Saby, N.P.A., Bernoux, M., Barthès, B.G., 2016. National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy. Geoderma 276, 41–52.
Cui, C., Fearn, T., 2018. Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration. Chemometr. Intell. Lab. 182, 9–20.
Fox, J., Weisberg, S., 2011. An R Companion to Applied Regression, Second ed. SAGE Publications, Thousand Oaks, CA.
Gholizadeh, A., Coblinski, J.A., Saberioon, M., Ben-Dor, E., Drábek, O., Dematte, J.A.M., Borůvka, L., Němeček, K., Chabrillat, S., Dajčl, J., 2021a. vis–NIR and XRF Data Fusion and Feature Selection to Estimate Potentially Toxic Elements in Soil. Sensors 21 (7), 2386.
Gholizadeh, A., Neumann, C., Chabrillat, S., van Wesemael, B., Castaldi, F., Borůvka, L., Sanderman, J., Klement, A., Hohmann, C., 2021b. Soil organic carbon estimation using VNIR–SWIR spectroscopy: The effect of multiple sensors and scanning conditions. Soil Till. Res. 211, 105017.
Grunwald, S., Vasques, G.M., Rivero, R.G., 2015. Chapter One - Fusion of Soil and Remote Sensing Data to Model Soil Properties. In: Sparks, D.L. (Ed.), Adv. Academic Press, Agron, pp. 1–109.
Hong, Y., Chen, S., Liu, Y., Zhang, Y., Yu, L., Chen, Y., Liu, Y., Cheng, H., Liu, Y., 2019a. Combination of fractional order derivative and memory-based learning algorithm to improve the estimation accuracy of soil organic matter by visible and near-infrared spectroscopy. CATENA 174, 104–116.
Hong, Y., Liu, Y., Chen, Y., Liu, Y., Yu, L., Liu, Y., Cheng, H., 2019b. Application of fractional-order derivative in the quantitative estimation of soil organic matter content through visible and near-infrared spectroscopy. Geoderma 337, 758–769.
Hong, Y., Chen, S., Chen, Y., Linderman, M., Mouazen, A.M., Liu, Y., Guo, L., Yu, L., Liu, Y., Cheng, H., Liu, Y., 2020a. Comparing laboratory and airborne hyperspectral data for the estimation and mapping of topsoil organic carbon: Feature selection coupled with random forest. Soil Till. Res. 199, 104589.
Hong, Y., Guo, L., Chen, S., Linderman, M., Mouazen, A.M., Yu, L., Chen, Y., Liu, Y., Liu, Y., Cheng, H., Liu, Y., 2020b. Exploring the potential of airborne hyperspectral image for estimating topsoil organic carbon: Effects of fractional-order derivative and optimal band combination algorithm. Geoderma 365, 114228.
Hong, Y., Munnaf, M.A., Guerrero, A., Chen, S., Liu, Y., Shi, Z., Mouazen, A.M., 2022. Fusion of visible-to-near-infrared and mid-infrared spectroscopy to estimate soil organic carbon. Soil Till. Res. 217, 105284.
Hutengs, C., Seidel, M., Oertel, F., Ludwig, B., Vohland, M., 2019. In situ and laboratory soil spectroscopy with portable visible-to-near-infrared and mid-infrared instruments for the assessment of organic carbon in soils. Geoderma 355, 113900.
Hutengs, C., Eisenhauer, N., Schädler, M., Lochner, A., Seidel, M., Vohland, M., 2021. VNIR and MIR spectroscopy of PLFA-derived soil microbial properties and associated soil physicochemical characteristics in an experimental plant diversity gradient. Soil Biol. Biochem. 160, 108319.
Jaillais, B., Pinto, R., Barros, A.S., Rutledge, D.N., 2005. Outer-product analysis (OPA) using PCA to study the influence of temperature on NIR spectra of water. Vib. Spectrosc. 39 (1), 50–58.
Javadi, S.H., Mouazen, A.M., 2021. Data Fusion of XRF and Vis-NIR Using Outer Product Analysis, Granger-Ramanathan, and Least Squares for Prediction of Key Soil Attributes. Remote Sensing 13 (11), 2023.
Javadi, S.H., Munnaf, M.A., Mouazen, A.M., 2021. Fusion of Vis-NIR and XRF spectra for estimation of key soil attributes. Geoderma 385, 114851.
Knadel, M., Viscarra Rossel, R.A., Deng, F., Thomsen, A., Greve, M.H., 2013. Visible-Near Infrared Spectra as a Proxy for Topsoil Texture and Glacial Boundaries. Soil Sci. Soc. Am. J. 77 (2), 568–579.
Knox, N.M., Grunwald, S., McDowell, M.L., Bruland, G.L., Myers, D.B., Harris, W.G., 2015. Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy. Geoderma 239–240, 229–239.
Lal, R., 2016. Soil health and carbon management. Food and Energy Security 5 (4), 212–222.
Levene, H., 1960. Robust tests for equality of variances. Robust Tests Equal. Var. 278–292.
Li, H., Liang, Y., Xu, Q., Cao, D., 2009. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. Anal. Chim. Acta. 648 (1), 77–84.
Li, H.-D., Liang, Y.-Z., Cao, D.-S., Xu, Q.-S., 2012. Model-population analysis and its applications in chemical and biological modeling. TrAC–Trend. Anal. Chem. 38, 154–162.
Li, H.-D., Xu, Q.-S., Liang, Y.-Z., 2018. libPLS: An integrated library for partial least squares regression and linear discriminant analysis. Chemometr. Intell. Lab. 176, 34–43.
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neur. In. 30.
Matamala, R., Jastrow, J.D., Calderón, F.J., Liang, C., Fan, Z., Michaelson, G.J., Ping, C.-L., 2019. Predicting the decomposability of arctic tundra soil organic matter with mid infrared spectroscopy. Soil Biol. Biochem. 129, 1–12.
McDowell, M.L., Bruland, G.L., Deenik, J.L., Grunwald, S., Knox, N.M., 2012. Soil total carbon analysis in Hawaiian soils with visible, near-infrared and mid-infrared diffuse reflectance spectroscopy. Geoderma 189–190, 312–320.
Mendes, W.d.S., Demattê, J.A.M., Rosin, N.A., Terra, F.d.S., Poppiel, R.R., Urbina-Salazar, D.F., Boechat, C.L., Silva, E.B., Curi, N., Silva, S.H.G., José dos Santos, U., Souza Valladares, G., 2022a. The Brazilian soil Mid-infrared Spectral Library: The Power of the Fundamental Range. Geoderma 415, 115776.
Mendes, W.d.S., Sommer, M., Koszinski, S., Wehrhan, M., 2022b. Peatlands spectral data influence in global spectral modelling of soil organic carbon and total nitrogen using visible-near-infrared spectroscopy. J. Environ. Manage. 317, 115383.
Minasny, B., Malone, B.P., McBratney, A.B., Angers, D.A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z.-S., Cheng, K., Das, B.S., Field, D.J., Gimona, A., Hedley, C.B., Hong, S.Y., Mandal, B., Marchant, B.P., Martin, M., McConkey, B.G., Mulder, V.L., O'Rourke, S., Richer-de-Forges, A.C., Odeh, I., Padarian, J., Paustian, K., Pan, G., Poggio, L., Savin, I., Stolbovoy, V., Stockmann, U., Sulaeman, Y., Tsui, C.-C., Vågen, T.-G., van Wesemael, B., Winowiecki, L., 2017. Soil carbon 4 per mille. Geoderma 292, 59–86.
Mishra, P., Marini, F., Brouwer, B., Roger, J.M., Biancolillo, A., Woltering, E., Echtelt, E. H.-v., 2021a. Sequential fusion of information from two portable spectrometers for improved prediction of moisture and soluble solids content in pear fruit. Talanta 223, 121733.
Mishra, P., Passos, D., 2021a. Deep multiblock predictive modelling using parallel input convolutional neural networks. Anal. Chim. Acta. 1163, 338520.
Mishra, P., Passos, D., 2021b. A synergistic use of chemometrics and deep learning improved the predictive performance of near-infrared spectroscopy models for dry matter prediction in mango fruit. Chemometr. Intell. Lab. 212, 104287.
Mishra, P., Roger, J.M., Rutledge, D.N., Biancolillo, A., Marini, F., Nordon, A., Jouan-Rimbaud-Bouveresse, D., 2020. MBA-GUI: A chemometric graphical user interface for multi-block data visualisation, regression, classification, variable selection and automated pre-processing. Chemometr. Intell. Lab. 205, 104139.
Mishra, P., Roger, J.-M., Jouan-Rimbaud-Bouveresse, D., Biancolillo, A., Marini, F., Nordon, A., Rutledge, D.N., 2021b. Recent trends in multi-block data analysis in chemometrics for multi-source data integration. TrAC–Trend. Anal. Chem. 137, 116206.
Munnaf, M.A., Mouazen, A.M., 2022. Removal of external influences from on-line vis-NIR spectra for predicting soil organic carbon using machine learning. CATENA 211, 106015.
Nawar, S., Richard, F., Kassim, A.M., Tekin, Y., Mouazen, A.M., 2022. Fusion of Gamma-rays and portable X-ray fluorescence spectral data to measure extractable potassium in soils. Soil Till. Res. 223, 105472.
Ng, W., Minasny, B., Malone, B.P., Sarathjith, M.C., Das, B.S., 2019a. Optimizing wavelength selection by using informative vectors for parsimonious infrared spectra modelling. Comput. Electron. Agr. 158, 201–210.
Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., McBratney, A.B., 2019b. Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. Geoderma 352, 251–267.
Ng, W., Minasny, B., Jones, E., McBratney, A., 2022. To spike or to localize? Strategies to improve the prediction of local soil properties using regional spectral library. Geoderma 406, 115501.
Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Ben Dor, E., Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A.M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Sakai, H., Soriano-Disla, J.M., Shepherd, K.D., Stenberg, B., Towett, E. K., Vargas, R., Wetterlind, J., 2015. Chapter Four - Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring. In: Sparks, D.L. (Ed.), Adv. Academic Press, Agron, pp. 139–159.
Poppiel, R.R., Paiva, A.F.d.S., Demattê, J.A.M., 2022. Bridging the gap between soil spectroscopy and traditional laboratory: Insights for routine implementation. Geoderma 425, 116029.
Pullanagari, R.R., Dehghan-Shoar, M., Yule, I.J., Bhatia, N., 2021. Field spectroscopy of canopy nitrogen concentration in temperate grasslands using a convolutional neural network. Remote Sens. Environ. 257, 112353.
Python Software Foundation, 2019. Python language reference. Python Software Foundation. https://www.python.org.

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Lanzhou, China (URL) http://www.R-project.org/.
Ramesh, T., Bolan, N.S., Kirkham, M.B., Wijesekara, H., Kanchikerimath, M., Srinivasa Rao, C., Sandeep, S., Rinklebe, J., Ok, Y.S., Choudhury, B.U., Wang, H., Tang, C., Wang, X., Song, Z., Freeman Ii, O.W., 2019. Chapter One - Soil organic carbon dynamics: Impact of land use changes and management practices: A review. In: Sparks, D.L. (Ed.), Adv. Academic Press, Agron, pp. 1–107.
Savitzky, A., Golay, M.J.E., 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal. Chem. 36 (8), 1627–1639.
Shen, Z., Ramirez-Lopez, L., Behrens, T., Cui, L., Zhang, M., Walden, L., Wetterlind, J., Shi, Z., Sudduth, K.A., Baumann, P., Song, Y., Catambay, K., Viscarra Rossel, R.A., 2022. Deep transfer learning of global spectra for local soil carbon monitoring. ISPRS-J. Photogramm. Remote Sens. 188, 190–200.
Shi, X.-Z., Yu, D.-S., Yang, G.-X., Wang, H.-J., Sun, W.-X., Du, G.-H., Gong, Z.-T., 2006. Cross-Reference Benchmarks for Translating the Genetic Soil Classification of China into the Chinese Soil Taxonomy. Pedosphere 16 (2), 147–153.
Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., Macdonald, L.M., McLaughlin, M. J., 2014. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. Appl. Spectrosc. Rev. 49 (2), 139–186.
Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Chapter Five - Visible and Near Infrared Spectroscopy in Soil Science. In: Sparks, D.L. (Ed.), Adv. Academic Press, Agron, pp. 163–215.
Stevens, A., Ramirez-Lopez, L., 2021. An introduction to the prospectr package. R package Vignette R package version (2), 2.
Tekin, Y., Tumsavas, Z., Mouazen, A.M., 2012. Effect of Moisture Content on Prediction of Organic Carbon and pH Using Visible and Near-Infrared Spectroscopy. Soil Sci. Soc. Am. J. 76 (1), 188–198.
Terra, F.S., Viscarra Rossel, R.A., Demattê, J.A.M., 2019. Spectral fusion by Outer Product Analysis (OPA) to improve predictions of soil organic C. Geoderma 335, 35–46.
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.
Veselá, A., Barros, A.S., Synytsya, A., Delgadillo, I., Čopíková, J., Coimbra, M.A., 2007. Infrared spectroscopy and outer product analysis for quantification of fat, nitrogen, and moisture of cocoa powder. Anal. Chim. Acta. 601 (1), 77–86.
Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158 (1), 46–54.
Vohland, M., Ludwig, M., Thiele-Bruhn, S., Ludwig, B., 2014. Determination of soil properties with visible to near- and mid-infrared spectroscopy: Effects of spectral variable selection. Geoderma 223–225, 88–96.
Vohland, M., Ludwig, B., Seidel, M., Hutengs, C., 2022. Quantification of soil organic carbon at regional scale: Benefits of fusing vis-NIR and MIR diffuse reflectance data are greater for in situ than for laboratory-based modelling approaches. Geoderma 405, 115426.
Wan, M., Hu, W., Qu, M., Li, W., Zhang, C., Kang, J., Hong, Y., Chen, Y., Huang, B., 2020. Rapid estimation of soil cation exchange capacity through sensor data fusion of portable XRF spectrometry and Vis-NIR spectroscopy. Geoderma 363, 114163.
Wang, S., Guan, K., Zhang, C., Lee, D., Margenot, A.J., Ge, Y., Peng, J., Zhou, W., Zhou, Q., Huang, Y., 2022. Using soil library hyperspectral reflectance and machine learning to predict soil organic carbon: Assessing potential of airborne and spaceborne optical soil sensing. Remote Sens. Environ. 271, 112914.
Wiesmeier, M., Urbanski, L., Hobley, E., Lang, B., von Lützow, M., Marin-Spiotta, E., van Wesemael, B., Rabot, E., Ließ, M., Garcia-Franco, N., Wollschläger, U., Vogel, H.-J., Kögel-Knabner, I., 2019. Soil organic carbon storage as a key function of soils - A review of drivers and indicators at various scales. Geoderma 333, 149–162.
Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometr. Intell. Lab. 58 (2), 109–130.
Xu, L., Hong, Y., Wei, Y., Guo, L., Shi, T., Liu, Y., Jiang, Q., Fei, T., Liu, Y., Mouazen, A. M., Chen, Y., 2020. Estimation of Organic Carbon in Anthropogenic Soil by VIS-NIR Spectroscopy: Effect of Variable Selection. Remote Sensing 12 (20), 3394.
Zhao, Y., Wang, M., Hu, S., Zhang, X., Ouyang, Z., Zhang, G., Huang, B., Zhao, S., Wu, J., Xie, D., 2018. Economics-and policy-driven organic carbon input enhancement dominates soil organic carbon accumulation in Chinese croplands. P. Natl. Acad. Sci. USA 115 (16), 4045–4050.