Computers and Electronics in Agriculture 227 (2024) 109494

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Enhancing soil organic carbon prediction of LUCAS soil database using deep learning and deep feature selection

Mohammadmehdi Saberioon [a,b,*], Asa Gholizadeh [c], Ali Ghaznavi [a], Sabine Chabrillat [a,d], Vahid Khosravi [c,a]

[a] Section 1.4 Remote Sensing and Geoinformatics, Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences, Telegrafenberg, Potsdam 14473, Germany
[b] ILVO, Flanders Research Institute for Agriculture, Fisheries and Food, Technology and Food Science-Agricultural Engineering, 9820 Merelbeke, Belgium
[c] Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, Kamycka 129, Suchdol, Prague 16500, Czech Republic
[d] Institute of Soil Science, Leibniz University Hannover, Herrenhauser Straße 2, 30419 Hannover, Germany

ARTICLE INFO

Keywords:
Soil organic carbon
Soil spectral library
Convolutional neural network
Fully connected neural network
Deep feature extraction
Large-scale
Stacked autoencoder

ABSTRACT

The main terrestrial carbon (C) fraction is soil organic carbon (SOC), which has a considerable effect on climate change and greenhouse gas emissions through the absorption and sequestration of carbon dioxide (CO_2). This has made SOC assessment very important from both economic and environmental viewpoints. The growing count of soil spectral libraries (SSLs) from regional to global scales has brought a tremendous opportunity for the quantification of SOC through developing spectral-based prediction models. Hence, there is a need to take advantage of big data analytics for spectral data processing. The unique ability of deep learning (DL) techniques to leverage important features of high-dimensional large-scale SSLs has made them top-demanding for more sophisticated modeling. The core objective of the present study was to assess the ability of two different DL algorithms, i.e., one-dimensional convolutional neural network (1DCNN) and fully connected neural network (FCNN) coupled with stacked autoencoder (SAE) feature extraction for SOC prediction based on the data from the land use/cover area frame statistical survey (LUCAS) database. SAE extracted the high-level deep features from the visible–near-infrared–shortwave infrared (Vis–NIR–SWIR) spectra of 11441 soil samples, which were then considered as inputs to the 1DCNN and FCNN models for predicting the SOC content. Both SAE-DL feature-selected models yielded higher accuracy than those the DL developed on the entire spectra and a random forest (RF) model was constructed for comparison. The best prediction was achieved by SAE-1DCNN (R^2 = 0.78, RMSE = 3.94%, RPD = 4.88, RPIQ = 3.91) followed by 1DCNN (R^2 = 0.73, RMSE = 5.43%, RPD = 3.67, RPIQ = 2.84) proving the superiority of 1DCNN over FCNN in this study. These results supported the applicability of combined deep features extraction and regression methods for predicting SOC using high dimensional large-scale SSLs.

1. Introduction

Soil degradation is a serious concern worldwide and may be evident in increasing carbon dioxide (CO_2) emissions following deforestation, the reduction in above- and below-ground carbon (C) storage, and its impact on the ecosystems' ability to control soil-vegetation-atmosphere transfer (SVAT) processes (Vågen et al., 2016). This, in turn, affects the delivery of essential ecosystem services and has significant implications for climate change and food security (Lal, 2004). Soil organic carbon (SOC) serves as a crucial indicator of soil quality, with the European Union (EU) identifying its decline as one of the primary factors contributing to soil degradation. Consequently, there is a substantial demand for regular SOC quantification and monitoring. However, wet chemical analyses are costly and time-consuming which complicates the continuous SOC monitoring, especially at large scales. Such assessment would greatly benefit from rapid and cost-effective SOC prediction techniques.

One recognized approach to accurately, rapidly, and inexpensively quantify and monitor soil attributes is soil spectroscopy across the visible–near-infrared–shortwave infrared (Vis–NIR–SWIR; 350–2500 nm) part of the electromagnetic spectrum (Ben-Dor and Banin, 1995).

* Corresponding author at: Section 1.4 Remote Sensing and Geoinformatics, Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences, Telegrafenberg, Potsdam 14473, Germany.
E-mail address: saberioon@gfz-potsdam.de (M. Saberioon).

https://doi.org/10.1016/j.compag.2024.109494
Received 20 June 2024; Received in revised form 5 September 2024; Accepted 23 September 2024
Available online 28 September 2024

Various multivariate statistical methods and machine learning (ML) algorithms, such as partial least squares regression (PLSR), artificial neural networks (ANN), random forest (RF), support vector machine regression (SVMR) and cubist have been employed to develop proxy models for SOC assessment using spectral data (Angelopoulou et al., 2020). The majority of published research has focused on small-scale calibration of SOC prediction models using local spectral datasets and hence, applying and transferring them to areas with different soil conditions poses a significant challenge (Castaldi et al., 2018). An increasing number of large-scale soil spectral libraries (SSLs) are currently developing on regional, continental, and global scales to address the abovementioned issue. Some of these SSLs are the global SSL, the European land use/cover area frame statistical survey (LUCAS), the Brazilian SSL (BSSL), The International Centre for research in Agroforestry–international soil reference and Information Centre (ICRAF-ISRIC) SSL, the open SSL (OSSL), and the Balkan, Middle East, and North African (GEO-CRADLE) SSL. These SSLs represent a tremendous opportunity for developing adapted approaches for SOC prediction models. They offer access to large-scale databases with high data volume and significant spectral variability.

The development of large-scale SSLs has caused a growing need for big data analytics to analyze vast and diverse datasets efficiently, uncover hidden patterns, and identify previously unknown correlations in soil science research (Padarian et al., 2019). Due to their inherent capacity for processing large-scale data, hierarchical learning, and support for extensive computational resources, deep learning (DL) techniques have gained increasing popularity for modeling soil properties (Tsakiridis et al., 2020a). Among DL algorithms, convolutional neural networks (CNN) and fully connected neural networks (FCNN) are common for various soil science applications. However, the application of DL in predicting SOC has predominantly focused on CNN, which excels at extracting deep features from datasets using kernel learning. For instance, Shen and Viscarra Rossel (2021) introduced an automated approach for tuning a one-dimensional convolutional neural network (1DCNN) and demonstrated its effectiveness on the LUCAS SSL for SOC quantification. In another study by Tsakiridis et al. (2020b), a localized multi-channel 1DCNN was applied to the LUCAS dataset, aiming to estimate ten different soil physicochemical properties, including SOC, simultaneously. While some studies have combined CNNs with other algorithms for SOC prediction, there is a noticeable lack of research using FCNN for SOC quantification (Zhao et al., 2021).

Soil properties are intricately encoded in the measured spectra, therefore there is a crucial need to streamline spectral complexity and enhance the prediction performance (Ma et al., 2023). This can be achieved by eliminating non-informative, redundant variables and extracting significant features that exhibit high correlation with the target variables. The utilization of effective and DL-compatible feature selection and extraction methods holds significant importance for harnessing the full potential of SSLs and modeling SOC. Some DL methods can incorporate autoencoders to simplify data complexity and reduce dimensionality by eliminating irrelevant variables (Wang et al., 2016). This process ensures that models are fed with the optimal number of selected or extracted spectral features, improving their accuracy and efficiency.

Stacked autoencoder (SAE) neural network is an unsupervised learning network made of multiple layers of sparse autoencoders. Recent advancement in autoencoder-based feature extraction has led to more applications of SAE as an alternative feature extraction technique. This has aided in solving the curse of dimensionality as well as providing more discriminating features as compared to the conventional feature selection approaches (Roy et al., 2018). Far to date, SAE had been applied to a wide variety of applications due to its feature extraction; thus, it was allowed to deal with data as complex as this. However, none of these uses has demonstrated how far the combination of SAE with deep learning techniques applies to soil reflectance spectra for SOC assessment.

Within this framework, this study intends to examine the ability of two different DL algorithms (i.e., CNN and FCNN) coupled with SAE model feature selection/extraction, to quantify and model SOC using LUCAS SSL. To our knowledge, these algorithms have not been tested for SOC prediction, neither individually nor combined. As variables of greater importance are selected/extracted using a feature selection/extraction technique, it is therefore expected that the models' accuracy will be improved compared to models developed using the DL on full-range spectra. In addition, the results of the proposed DL architectures will be compared with those obtained by using the more popular conventional ML (i.e., random forest) algorithm.

2. Materials and methods

2.1. LUCAS soil dataset and pre-processing

The LUCAS database used contained 19967 surface soil samples (0–20 cm) taken from various land use types of 25 European Union states (Fig. 1), in 2009 (Toth et al., 2013). The database consists of SOC and 11 other soil sample properties, as well as their Vis–NIR–SWIR (400–2499.5 nm) spectra with 0.5 nm spectral resolution. The SOC content of all samples was determined by dry combustion using a Vario Max CN Analyzer (Elementar Analyse Systeme GmbH, Germany). Spectral measurements were also performed using FOSS XDS Rapid Content Analyzer apparatus (FOSS Analytical, Hilleroed, Denmark) in the laboratory. For this study, 11441 samples were selected from croplands as per the purpose of this study to monitor the SOC contents of areas related to agricultural activities.

For pre-processing of the spectra, the noisy parts from 350 and 500 nm were initially removed, leaving spectra in the range of 500 nm to 2499.5 nm. Savitzky-Golay (SG) smoothing was then performed on the remaining spectra to reduce the artificial noise caused by random measurement errors. The first derivative with a second-order polynomial fit and 41 wavelength window size was the next preprocessing step to correct the background signals and offset and improve spectral features (Gholizadeh et al., 2015). The spectral outliers were then identified using principal component analysis and Mahalanobis distance methods and removed from the dataset. The final pre-processing step was normalization of the SOC distribution by the Robust-Scaler transformation which subtracts the median and scales the data according to the quantile range. The pre-processed dataset was then divided into the train (75%) and test (25%) portions using Kennard-stone (KS) algorithm. The KS algorithm selects n samples uniformly distributed over the predictor space, thus not only ensures the randomization in data selection for the training and testing sets but also optimizes the coverage of the spectral variability. The training portion was utilized to build the prediction models and the testing portion was used to verify their accuracy and generalization ability.

The proposed algorithms were implemented based on Python 3.6. The deep learning framework was Keras with the TensorFlow backend. All proposed architectures were developed and completed on a personal computer at the early step and then transferred to the Google Colab Pro premium cluster account to train the most stable models. The Google Colab Pro cluster is equipped with an NVIDIA Tesla T4 or the NVIDIA Tesla P100 GPU with 16 GB of GPU VRAM, 52 GB of RAM, and two vCPUs.

2.2. Feature extraction by SAE

SAE is an unsupervised learning algorithm mainly used for feature extraction and dimensionality reduction. It is composed of multiple stacked layers of autoencoders to produce better high-level non-linear features of the input data (Meddeb et al., 2023). The autoencoder includes encoder and decoder units. During learning, the encoder transmits the input vector to a hidden feature representation while the

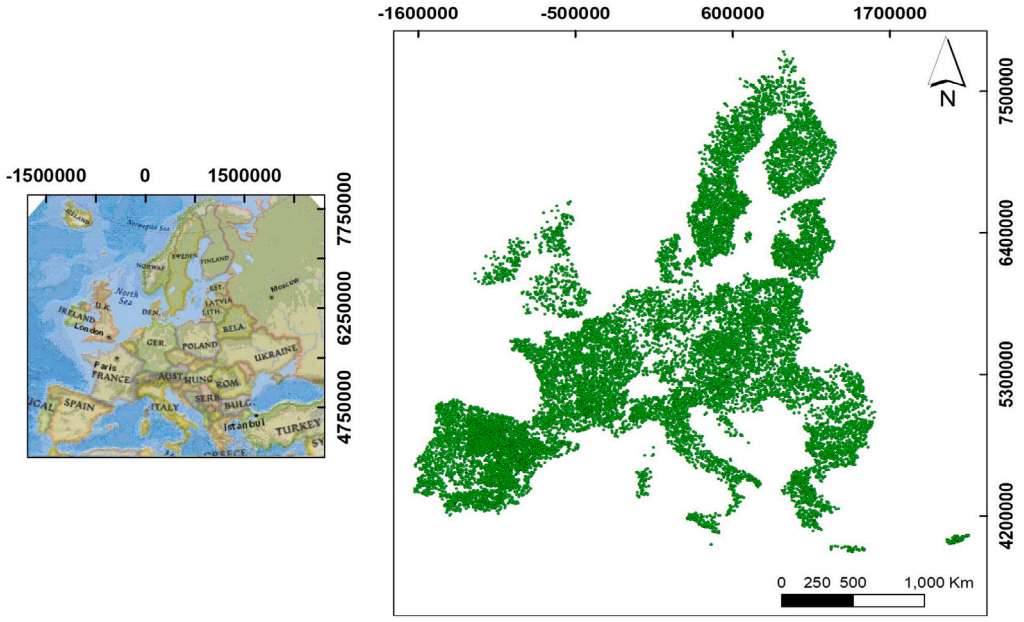

Fig. 1. LUCAS samples position in Europe.

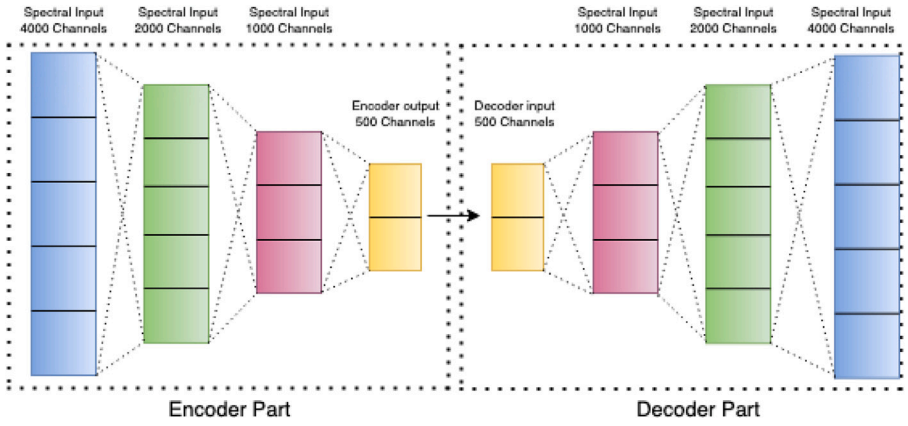

Fig. 2. Stacked autoencoder architecture.

decoder maps it back to a vector with the same size as the input space (Fig. 2). Yu et al. (2018b) have formulated these steps below:

$$y = f(w_y x + b_y)$$
$$z = f(w_z x + b_z) \quad (1)$$

where w_y and w_z are the input-to-hidden and hidden-to-output weight matrices, respectively, b_y and b_z are the hidden and output units bias, while the activation function is denoted by $f(.)$. In our study, the "leaky-ReLU" was used as the activation function $f(.)$.

The unsupervised pre-training of the SAE was initiated by introducing the original input spectra to the first layer. The numbers of channels in consecutive layers of the SAE were set to 4000, 2000, 1000, h, 1000, 2000 and 4000. The h denotes the number of channels (or deep spectral features) in the final layer of the encoding unit of SAE. After pre-training and assigning the optimal weights, the deep features extracted from the original spectra were then employed as inputs for the SAE-1DCNN and SAE-FCNN algorithms.

2.3. Spectral modeling

As the main purpose of this study, two DL models (1DCNN and FCNN) were developed on spectral features extracted by another DL methodology (SAE) to predict SOC. DL are neural network models with several "hidden" layers for gradual learning of more complex features and transforming the input data to outputs (Schmidhuber, 2015). DL models have a deeper structure with more layers than the traditional artificial neural networks (Padarian et al., 2019) which is beneficial for hierarchical learning of the deep features within the input dataset. The results of the DL approaches were compared with those obtained by the classic ML method of RF.

2.3.1. Stacked autoencoder - 1D convolutional neural network (SAE-1DCNN)

CNN is a deep feed-forward artificial neural network that mostly deals with video, image and signal processing tasks in various fields of research including soil sciences. A CNN model is usually composed of an input layer, a hidden layer (which typically contains convolution, pooling, and fully connected layers), and an output layer. For this study, a 1D-CNN architecture was designed with three convolution layers as the hidden layer. The input layer with a channel size of 500 received 500 deep spectral features extracted by SAE. The input layer output was then introduced to the convolution layers to extract features through the filter sliding (with the filter size of 8) process named the convolution operation (Yang et al., 2021). The first convolution

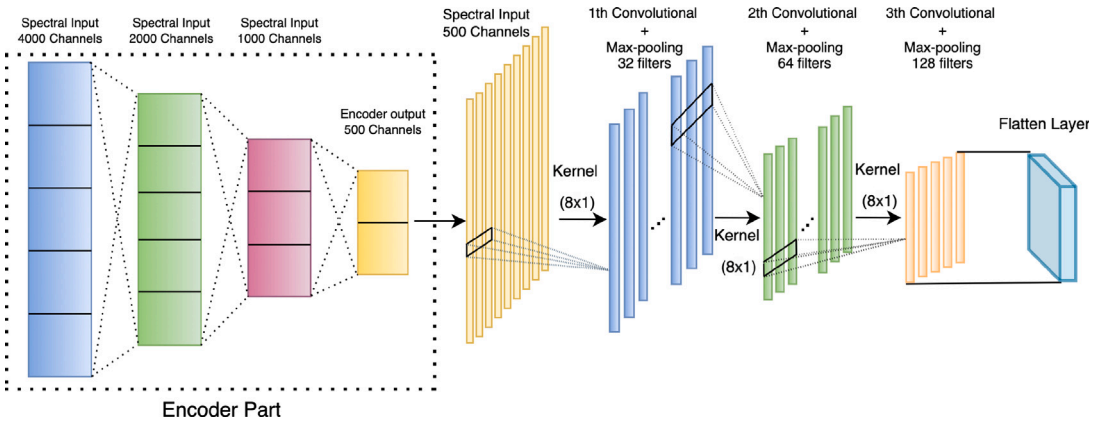

Fig. 3. Stacked autoencoder-1D Convectional Neural Network.

layer contained 32 neurons, while the neurons in the second and third convolution layers were doubled and quadrupled to 64 and 128, respectively. Each layer's output was then passed through the "leaky-ReLU" activation function and then employed as the next layer's input. To reduce the resulting output data dimensions and the over-fitting risk, a pooling layer was added to the network. The max-pooling method with a filter size of 2 by 2 was used for this step. A one-dimension vector was then created by flattening the pooling step's output and passed to a fully connected network which contained a multitude of neurons connecting to the SOC content as the output of this experiment. The model was properly trained with 700 epochs. Adam optimizer was employed with the learning rate of 1e-3, to obtain the optimal model while mean square error (MSE) was used as the loss function. Fig. 3 shows the architecture of SAE-1DCNN used in this study.

2.3.2. Stacked autoencoder - fully connected neural network (SAE-FCNN)

SAE-FCNN is a deep neural network architecture (Fig. 4) that leverages the capabilities of SAE for unsupervised extraction of deep spectral features and FCNN for supervised prediction of target variables using the extracted deep features (Yu et al., 2018b). After pre-training of SAE and extraction of the deep spectral features, the decoding part was removed and FCNN was added to the SAE encoding layer forming a SAE-FCNN regression network for SOC prediction. Dropout regularization was used which is the random selection and ignoring the neurons while the remaining can assist in predicting the ignored ones. Dropout forces the network to be less sensitive to the specific weight of neurons.

Training of the network was conducted with the batch size of twenty and maximum epoch size of 400. The early stopping with 25 epochs of patience was used for validation of the training phase. To minimize the RMSE function, the Adam optimizer was used to train the model with a starting learning rate of 0.001.

2.3.3. Random forest

RF is a tree-based learning algorithm for both classification and regression applications. Multiple decision trees are produced in the forest and fitted on various data subsets. The average result obtained by each tree was used to improve the accuracy of prediction and control over-fitting. Three important RF parameters needed to be tuned before modeling: (i) a number of regression trees (n_{tree}) in the forest, (ii) minimum data per node (*nodesize*), and (iii) a number of predictors (m_{try}) selected at each node.

2.3.4. Assessment metrics

The accuracy of each developed model was determined using four commonly used evaluation techniques including root mean squared error (RMSE), coefficient of determination (R^2), ratio of prediction to deviation (RPD) and ratio of performance to inter-quartile distance (RPIQ).

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

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

$$RPD = \frac{StD}{RMSE} \quad (4)$$

$$RPIQ = \frac{Q_3(obs) - Q_1(obs)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(obs_i - pred_i)^2}} \quad (5)$$

Table 1
Statistical summary of SOC concentration (%) of dataset after removing outliers.

n	Minimum	Maximum	Mean	Std.	CV	Skewness
9894	0.00	19.92	2.26	19.70	8.72	3.42

n: Number of samples, Std: Standard deviation, CV: Coefficient of variation.

3. Results

3.1. Samples statistical summary and spectra

The SOC content statistics of the remaining LUCAS dataset after outliers removal, are summarized in the Table 1. It includes minimum, maximum, mean, standard deviation (StD), coefficient of variation (CV), and skewness. Accordingly, the average SOC concentration was 2.26%±19.7% (StD), with a very high variability of CV = 8.72 which was expected due to the diverse soil and land use types that the samples were gathered from several European countries and geographical regions (Toth et al., 2013). In addition, the dataset was positively skewed from the normal distribution and hence was subjected to the Robust-Scaler transformation.

The average reflectance and related standard deviation of the LUCAS samples used in this study are presented in Fig. 5. As a brief overview, the overall shape of the spectra and reflectance patterns are roughly typical of soil spectra, with a continuous ascending trend and two 1400 nm and 1900 nm located absorption minima stemmed from the soil hygroscopic moisture content (Gholizadeh et al., 2023). The other absorption features between 2000 nm and 2500 nm can mostly be attributed to the soil mineral and organic constituents. The relatively high variations of the spectra were not surprising as the samples have been collected from various soil types and land uses across Europe.

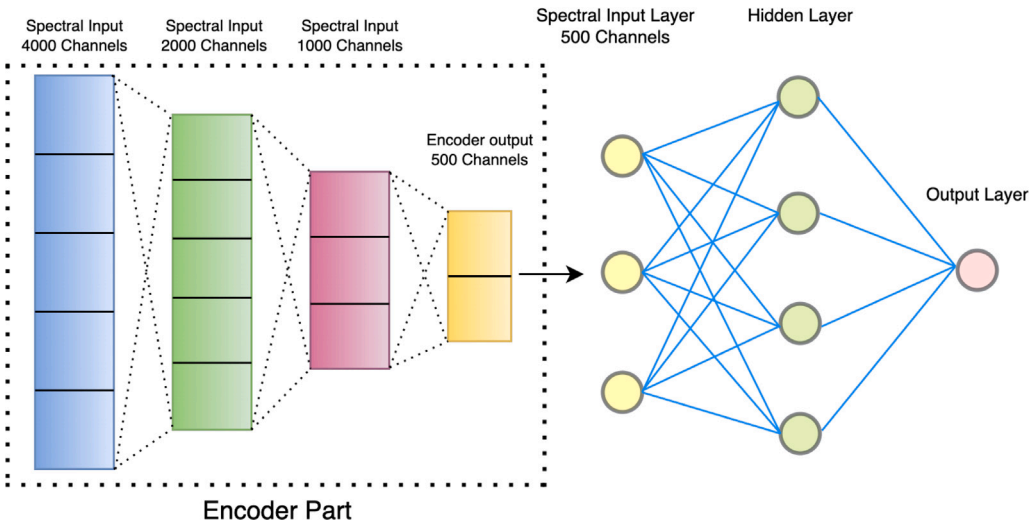

Fig. 4. Stacked autoencoder fully connected neural Network.

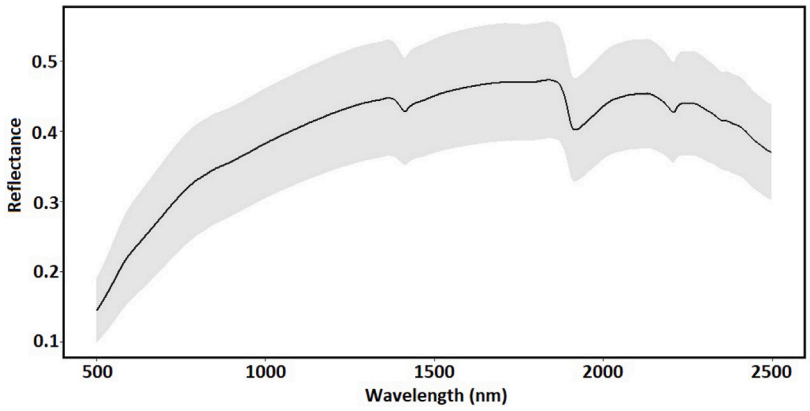

Fig. 5. Average reflectance of all samples along with variance.

3.2. SAE extracted features

The SAE feature extraction network was well-trained and optimized after 700 epochs as shown in Fig. 6b. The calculated average deviation of the output spectra from the original input spectra was acceptably low (8.5×10^{-3}) (Fig. 6a). This low deviation indicates that the SAE effectively reconstructed the original input spectra. The mean spectra calculation, which is useful for reducing the effect of measurement variations and testing repeatability, also supports the robustness of this reconstruction. After pre-training, the feature extraction began in the first hidden layer of the SAE, where the original 4000-band input spectra were encoded into 2000 variables. These 2000 variables were then passed to the second layer, where they were further reduced to 1000 features. In the final layer, the input variables were reduced from 1000 to 500, resulting in the extraction of the most informative deep spectral features. This consecutive dimension reduction from 4000 to 500 variables ensured that the final SAE output retained the most representative features of the original input spectral data.

3.3. Models performance

All DL models were well trained and converged after running 700 epochs based on the training/validation loss (Fig. 7). The best training performance and stability were achieved by assuming that all models were trained well according to the best-optimized hyperparameter values. Several models were trained based on different values of hyperparameters to achieve the best model performance and training stability. Table 2 presents the SOC prediction results obtained by different DL-developed architectures. SOC prediction performance of the RF model developed on the whole spectra is also included for comparison.

As a general comparison between different algorithms used in this study, 1DCNN outperformed both FCNN and RF, regardless of the input data (whole spectra or extracted features). According to the results, 1DCNN developed on the extracted features (SAE-1DCNN) had the highest prediction accuracy (R^2 = 0.78, RMSE = 3.93%, RPD = 4.88, RPIQ = 3.91) followed by 1DCNN developed on the whole spectra (R^2 = 0.73, RMSE = 5.43%, RPD = 3.67, RPIQ = 2.84). The worst performance was exhibited by the RF model constructed on the whole spectra (R^2 = 0.61, RMSE = 8.87%, RPD = 2.17, RPIQ = 1.74). The FCNN and SAE-FCNN methods showed medium performances with prediction results between 1DCNN and RF (R^2 = 0.63, RMSE = 6.72%, RPD = 2.97, RPIQ = 2.29 for FCNN and R^2 = 0.64, RMSE = 6.12%, RPD = 3.25, RPIQ = 2.52 for SAE-FCNN). Considering the effect of SAE, developing models on the extracted features led to better predictions than those constructed on the whole spectra. This difference was much greater between results obtained by 1DCNN and SAE-1DCNN.

4. Discussion

4.1. Effect of the feature extraction

DL Models developed on SAE-extracted features yielded better SOC predictions than those developed on the whole spectra. As a robust

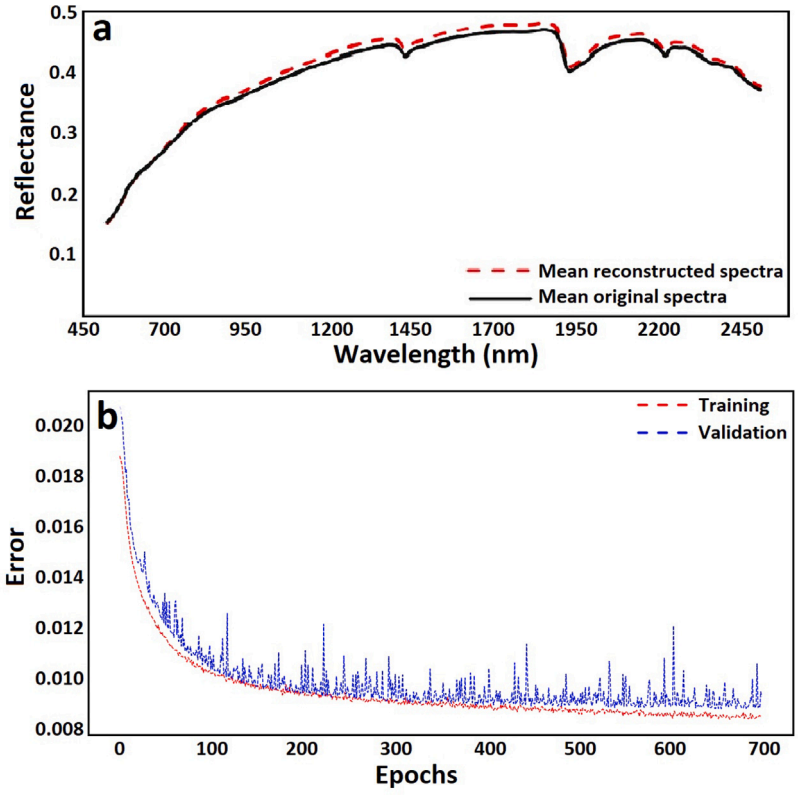

Fig. 6. The mean original and reconstructed spectra (a), and Pre-training error (b).

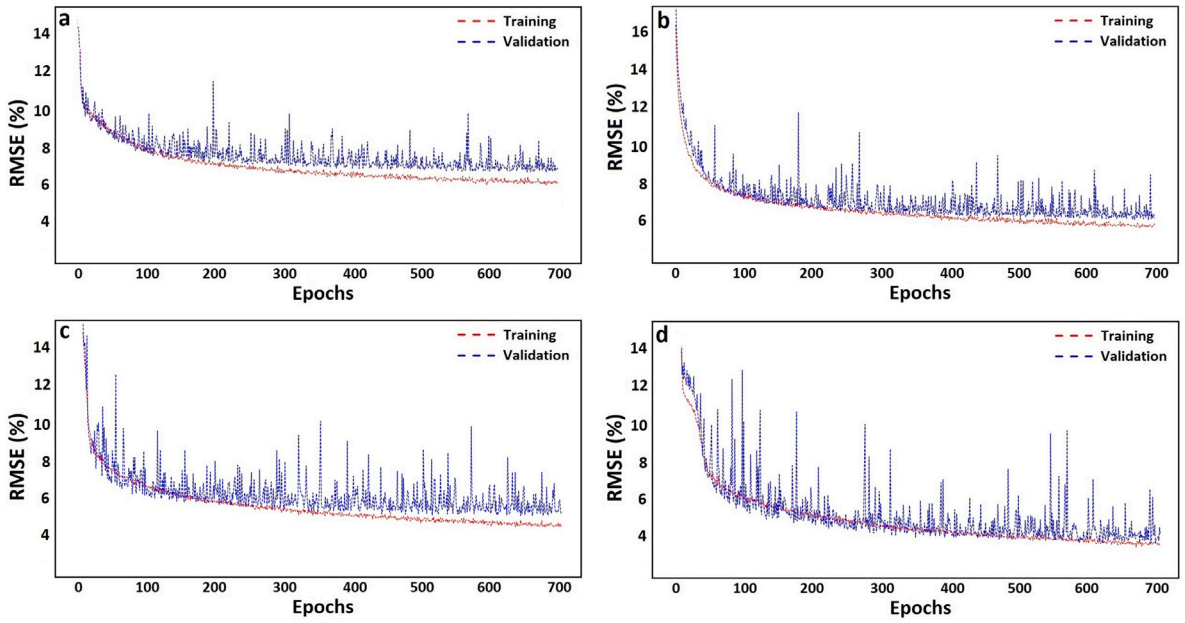

Fig. 7. Training/validation plots for FCNN (a), SAE-FCNN (b), 1DCNN (c), and SAE-1DCNN (d).

Table 2
SOC prediction models performance based on different architectures.

	EP	LR	BS	AF	R^2	RMSE	RPD	RPIQ
FCNN	700	−3	20	Leaky	0.625	6.715	2.966	2.293
SAE - FCNN	700	−3	20	Leaky	0.641	6.122	3.254	2.516
1DCNN	700	−3	16	Leaky	0.733	5.429	3.669	2.837
SAE - 1DCNN	700	−3	16	Leaky	0.784	3.935	4.884	3.914
RF	–	–	–	–	0.607	8.868	2.167	1.737

EP: Epochs, LR: learning rate, BS: Batch size, AF: Activation function.

dimension reduction algorithm, SAE extracted high-level deep spectral features from the LUCAS data and then passed them to the input layers of the DL networks. This resulted in higher informative data with lower dimensionality which substantially improved efficiency of the modeling process as well as the accuracy of the predicted SOC. Using FCNN models developed on SAE-derived spectral features, similar enhancements were reported by Yu et al. (2018a) for nitrogen concentration detection in oilseed rape with R^2 = 0.903, RMSE=0.307% and RPD=3.238, and Yu et al. (2018b) for pear's firmness and soluble solid content (SSC) prediction with R^2 = 0.890, RMSE=1.81% and RPD=3.05 for firmness, and R^2 = 0.921, RMSE=0.22% and RPD=3.68 for SSC. SAE–FCNN models constructed on the NIR hyperspectral images were also found to be potential tools for non-destructive quantification of the total viable count (TVC) in peeled shrimp, providing a new quality and safety prospecting methodology for shrimp products (Yu et al., 2019).

4.2. RF or DL techniques? a comparison

Generally speaking, RF is a powerful and efficient algorithm with relatively fast computational speed and limited computational cost (Bai et al., 2022). The SOC prediction performance of RF in this study was acceptable to some extent (Table 2) which further proves the RF efficiency for this specific application. However, all DL and SAE-DL techniques outperformed RF in the prediction of SOC. It might be due to some drawbacks that are associated with the RF method especially when facing large datasets with high dimensionality. RF may overfit to the uninformative noisy data especially when its structure is composed of a large number of trees which makes the forest structure too complex. In addition, RF is like a black box and hence, except for tuning of some parameters, there is no control on the inner processes to enhance its performance and accuracy (Bai et al., 2022).

Better performance of 1DCNN over FCNN can be attributed to, the strong learning ability of CNN which has made it capable of detecting important features within complex and high-dimensional LUCAS datasets (Ma et al., 2019). The great potential of CNN in dealing with large datasets can be considered as another reason that has made it the best option for modeling SOC using LUCAS SSL. This potential has provided a great opportunity for large-scale monitoring of the soil properties (Yang et al., 2021). The superiority of CNN-based models constructed on proximal and remote sensing data has been acknowledged in several SOC prediction-related researches. In a SOC assessment study using Vis–NIR–SWIR spectra in China, CNN outperformed both RF and PLSR techniques with higher prediction accuracies (Bai et al., 2022). Similarly, in the study of SOC prediction using USDA SSL with 37540 Vis–NIR–SWIR reflectance spectra, CNN outperformed RF and other well-known ML models such as PLSR, K-Nearest Neighbors (KNN) and Ridge (Wang et al., 2022). The performance of CNN is however affected by some important factors including hardware quality, data size (Bai et al., 2022) and computational power demand (Wang et al., 2022). Several parameters are also needed to be optimized during the modeling to prevent the overfitting (Darwish et al., 2020).

Unlike CNN, few studies have investigated the applicability of FCNN in prediction of the soil properties. In a study conducted by Gholizadeh et al. (2020) on the forest soils, FCNN developed on Vis–NIR–SWIR spectra outperformed RF, PLSR and SVMR in predicting the toxic elements. They also noticed the higher performance of FCNN on larger datasets, while in the lower-size datasets, the results obtained by RF and SVMR were comparable to those obtained by FCNN. This was justified by the strong capability of RF and SVMR to develop models using small-sized datasets (Jiang et al., 2019). However, through all the investigation, 1DCNN and FCNN are obviously performing better, and one problem remains to be critical: the interpretability. Deep learning models, though at least offering some interpretability by the user in the form of features of importance and decision trees, are still mostly treated as black boxes, given their architecture can be quite complex and data is very high dimensional. This can render the interpretability of how the concrete spectral features are working in the prediction of SOC quite impenetrable and reduces practical usability for such models. In this respect, interpretability provides further decision-making capabilities, which may be one of the limiting factors to broader acceptance in the realm of soil science, let alone beyond, despite high predictive performance these models can otherwise achieve. Emphasis should therefore be placed on model improvements for better performance and on developing methods that could possibly make deep learning models more interpretable with the use of visualization techniques, explainable AI architectures, integration of interpretable models in deep learning architectures. Notably, Grushetskaya et al. (2024) have recently explored the use of XAI frameworks with the same LUCAS SSL data used in this study, demonstrating how such methods can help explain deep learning model decisions for SOC prediction, thereby increasing transparency and usability in practical applications.

5. Conclusion

The main purpose of this research was to compare two DL-based SOC prediction models (1DCNN and FCNN) established on the whole LUCAS SSL spectra and related features extracted by the SAE algorithm. According to the results, models developed on the SAE retrieved features yielded higher accuracy compared to when they were constructed on the whole Vis–NIR–SWIR spectra. SAE-1DCNN model showed the highest accuracy with R^2 = 0.784, RMSE = 3.935%, RPD = 4.884, RPIQ = 3.914 followed by 1DCNN (R^2 = 0.733, RMSE = 5.429%, RPD = 3.669, RPIQ = 2.837) proving the superiority of 1DCNN over FCNN within the current study. Regardless of the input dataset used (whole spectra or SAE derived), both DL methods outperformed the RF model developed on the whole spectra. This could further prove the superiority of DL algorithms when facing large high-dimensional datasets, specifically the SSLs. Results obtained in this research can be encouraging for more investigation efforts on both feature extraction and regression applications of DL algorithms on SSLs to predict SOC or other soil properties. Despite the promising results, there are some limitations in the current study. Models were developed and tested on the LUCAS SSL dataset alone and may not generalize well to other soil spectral libraries or regions. Reliance on SAE for feature extraction suggests that an investigation of the application of other advanced feature extraction techniques might further improve model performance. Other issues concern the huge computational resource needs of deep learning models, especially 1DCNN, which raises questions about their practicality in any setting. In addition, areas that continue to need further research include the lack of external validation on independent data sets and the intrinsic complexity and challenges of interpretability with deep learning models. This would, therefore, also become important in addressing these limitations for future studies to enhance the robustness and applicability of DL-based SOC prediction models in subsequent works.

CRediT authorship contribution statement

Mohammadmehdi Saberioon: Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Asa Gholizadeh:** Writing – review & editing, Writing – original draft, Conceptualization. **Ali Ghaznavi:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis. **Sabine Chabrillat:** Writing – review & editing, Writing – original draft, Conceptualization. **Vahid Khosravi:** Writing – review & editing, Writing – original draft, Visualization, Formal analysis.

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

Data will be made available on request.

Acknowledgments

This work was supported by the European Union's Horizon H2020 research and innovation European Joint Programme Cofund on Agricultural Soil Management (EJP-SOIL grant number 862695) and was carried out in the framework of the STEROPES of EJP-SOIL.

References

Angelopoulou, T., Balafoutis, A., Zalidis, G., Bochtis, D., 2020. From laboratory to proximal sensing spectroscopy for soil organic carbon estimation—A review. Sustainability 12 (2), 443.

Bai, Z., Xie, M., Hu, B., Luo, D., Wan, C., Peng, J., Shi, Z., 2022. Estimation of soil organic carbon using vis-nir spectral data and spectral feature bands selection in southern Xinjiang, China. Sensors 22 (16), 6124.

Ben-Dor, E., Banin, A., 1995. Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. Soil Sci. Soc. Am. J. 59 (2), 364–372. http://dx.doi.org/10.2136/sssaj1995.03615995005900020014x.

Castaldi, F., Chabrillat, S., Chartin, C., Genot, V., Jones, A., 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.

Darwish, A., Ezzat, D., Hassanien, A.E., 2020. An optimized model based on convolutional neural networks and orthogonal learning particle swarm optimization algorithm for plant diseases diagnosis. Swarm Evol. Comput. 52, 100616.

Gholizadeh, A., Borˊ uvka, L., Saberioon, M., Kozák, J., Vašát, R., Němeček, K., 2015. Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features. Soil Water Res. 10 (4), 218. http://dx.doi.org/10.17221/113/2015-SWR.

Gholizadeh, A., Saberioon, M., Ben-Dor, E., Rossel, R.A.V., Borůvka, L., 2020. Modelling potentially toxic elements in forest soils with vis–nir spectra and learning algorithms. Environ. Pollut. 267, 115574. http://dx.doi.org/10.1016/j.envpol.2020.115574.

Gholizadeh, A., Saberioon, M., Pouladi, N., Ben-Dor, E., 2023. Quantification and depth distribution analysis of carbon to nitrogen ratio in forest soils using reflectance spectroscopy. Int. Soil Water Conserv. Res. 11 (1), 112–124.

Grushetskaya, Y., Sips, M., Schachtschneider, R., Saberioon, M., Mahan, A., 2024. HPExplorer: XAI method to explore the relationship between hyperparameters and model performance. In: Bifet, A., Krilavičius, T., Miliou, I., Nowaczyk, S. (Eds.), Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track. Springer Nature Switzerland, pp. 319–334. http://dx.doi.org/10.1007/978-3-031-70378-2_20.

Jiang, H., Rusuli, Y., Amuti, T., He, Q., 2019. Quantitative assessment of soil salinity using multi-source remote sensing data based on the support vector machine and artificial neural network. Int. J. Remote Sens. 40 (1), 284–306.

Lal, R., 2004. Soil carbon sequestration impacts on global climate change and food security. Science 304 (5677), 1623–1627.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. ISPRS J. Photogram. Remote Sens. 152, 166–177.

Ma, Y., Minasny, B., Demattê, J.A., McBratney, A.B., 2023. Incorporating soil knowledge into machine-learning prediction of soil properties from soil spectra. Eur. J. Soil Sci. 74 (6), e13438.

Meddeb, R., Jemili, F., Triki, B., Korbaa, O., 2023. A deep learning-based intrusion detection approach for mobile Ad-hoc network. Soft Comput. 1–15.

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

Roy, M., Bose, S.K., Kar, B., Gopalakrishnan, P.K., Basu, A., 2018. A stacked autoencoder neural network based automated feature extraction method for anomaly detection in on-line condition monitoring. In: 2018 IEEE Symposium Series on Computational Intelligence. SSCI, IEEE, pp. 1501–1507.

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural Netw. 61, 85–117.

Shen, Z., Viscarra Rossel, R., 2021. Automated spectroscopic modelling with optimised convolutional neural networks. Sci. Rep. 11 (1), 1–12.

Toth, G., Jones, A., Montanarella, L., Alewell, C., Ballabio, C., Carre, F., DE, B.D., Guicharnaud, R.A., Gardi, C., Hermann, T., et al., 2013. LUCAS Topoil Survey-Methodology, Data and Results.

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

Tsakiridis, N.L., Keramaris, K.D., Theocharis, J.B., Zalidis, G.C., 2020b. Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network. Geoderma 367, 114208.

Vågen, T.-G., Winowiecki, L.A., Tondoh, J.E., Desta, L.T., Gumbricht, T., 2016. Mapping of soil properties and land degradation risk in Africa using MODIS reflectance. Geoderma 263, 216–225.

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.

Wang, Y., Yao, H., Zhao, S., 2016. Auto-encoder based dimensionality reduction. Neurocomputing 184, 232–242.

Yang, L., Cai, Y., Zhang, L., Guo, M., Li, A., Zhou, C., 2021. A deep learning method to predict soil organic carbon content at a regional scale using satellite-based phenology variables. Int. J. Appl. Earth Obs. Geoinf. 102, 102428.

Yu, X., Lu, H., Liu, Q., 2018a. Deep-learning-based regression model and hyperspectral imaging for rapid detection of nitrogen concentration in oilseed rape (Brassica napus L.) leaf. Chemometr. Intell. Lab. Syst. 172, 188–193.

Yu, X., Lu, H., Wu, D., 2018b. Development of deep learning method for predicting firmness and soluble solid content of postharvest Korla fragrant pear using vis/NIR hyperspectral reflectance imaging. Postharvest Biol. Technol. 141, 39–49.

Yu, X., Yu, X., Wen, S., Yang, J., Wang, J., 2019. Using deep learning and hyperspectral imaging to predict total viable count (TVC) in peeled Pacific white shrimp. J. Food Meas. Charact. 13, 2082–2094.

Zhao, W., Wu, Z., Yin, Z., 2021. Estimation of soil organic carbon content based on deep learning and quantile regression. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. IEEE, pp. 3717–3720.