Rapid detection of soil carbonates by means of NIR spectroscopy, deep learning methods and phase quantification by powder X-ray diffraction.

Lykourgos Chiniadis[a*], Petros Tamvakis[b]

[a] Athena Research and Innovation Centre, ILSP Xanthi's Division, Xanthi 67100, Greece,

b Hellenic Open University, School of Sciences and Technology, 26335 Patras, Greece

e-mail: Lykourgos Chiniadis: lykchiniadis@gmail.com and Petros Tamvakis: tamvakis.petros@ac.eap.gr

*Corresponding author

ORCID(s): LC 0000-0002-7118-9942, PT 0000-0001-9514-8283

Authorship contribution statement

Lykourgos Chiniadis: Conceptualization, Data acquisition, Writing – review and editing, Validation, Formal analysis, Supervision.

Petros Tamvakis: Conceptualization, Data curation, Writing – review and editing, Validation, Programming, Formal Analysis, Visualization.

ABSTRACT

In this study we propose a novel rapid and efficient way to predict carbonates content in soil by means of Fourier Transform Near-Infrared (FT-NIR) reflectance spectroscopy and by use of deep learning methods. In addition to using traditional machine learning algorithms, we exploited multiple deep learning methods, such as: 1) a Multi-Layered Perceptron Regressor (MLP) and 2) a Convolutional Neural Network (CNN) in an attempt to compare their performance with other classical machine learning algorithms, which up until now were considered the field's standards, such as Partial Least Squares Regression (PLSR), Cubist and Support Vector Machines (SVM) on the combined dataset of two NIR spectral libraries: Kellogg Soil Survey Laboratory (KSSL) of the United States Department of Agriculture (USDA), a dataset of soil samples reflectance spectra collected nationwide, and Land Use and Coverage Area Frame Survey (LUCAS) TopSoil-2015 (European Soil Library) which contains soil sample absorbance spectra from all over the European Union, and use them to predict carbonate content on never-before-seen soil samples. In this study, absorbances in the NIR spectral region (1150-2500 nm) were utilized in two different ways: a) as one-dimensional data and b) as two-dimensional spectrograms which, to our knowledge, is a completely novel approach that has rarely been researched. Quantification of carbonates by means of X-ray-Diffraction is in good agreement with the volumetric method and the MLP prediction. Our work contributes to rapid carbonates content prediction in soil samples in cases where: 1) no volumetric method is available and 2) only NIR spectra absorbance data are available.

1. Introduction

NIR spectroscopy is a rapid, non-destructive method of low cost that provides excellent correlation of observed and predicted values when regression algorithms are applied in spectral data and especially in large datasets (1–3). Early research studies in soil properties are using minimal input in terms of soil samples collected and considering minimal geographical areas and locations (4,5). These studies typically modeled total carbon (tC) and soil organic carbon (SOC) (2,5–8), soil organic matter (SOM) (9,10), soil inorganic carbon and carbonates (SIC) (11), total nitrogen (tN) (12–18), phosphorous content (12), potassium content (12), clay (18,19), pH (15,16), moisture content (6,20,21) and cationic exchange capacity (CEC) (18,22). Some studies had also modeled other properties specific to their objectives, such as total elemental content e.g., in (15). Computational progress and especially new algorithms applied in the spectral data are nowadays efficient to predict soil properties for larger datasets and fields, acquiring hundreds and thousands soil samples over locations of interest to generate large databases of spectral and physicochemical properties. Recently, hyperspectral NIR spectral data acquisition is even possible using satellites with high resolution data acquired (23).

Graphical Abstract

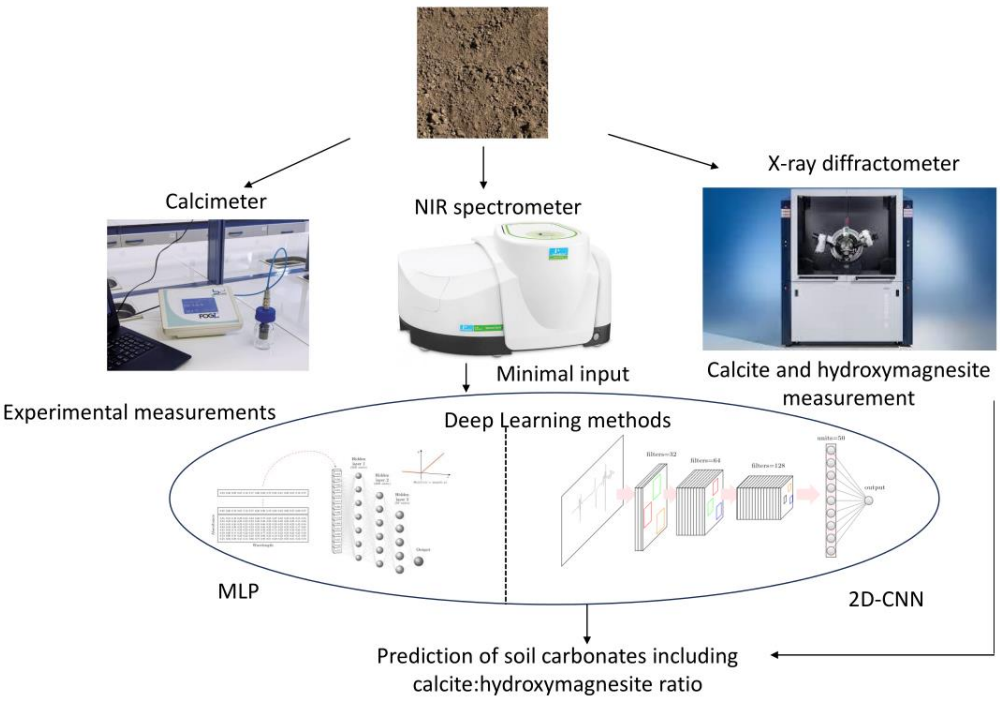


Figure 1*: The pipeline of collection of soil samples, experimental and theoretical assessment, and prediction of carbonates. Soil samples were collected from two distinct areas in Xanthi and Alexandroupolis fields and were sieved to >250μm. Furthermore, NIR spectra were acquired and calcimetry was applied in addition to calibration dataset generated by large spectral libraries (KSSL, USDA and LUCAS-2015-TopSoil). This calibration dataset, with spectra in the NIR region, was utilized for applying different regression models (namely PLSR, SVM, Cubist, MLP and 2D-CNN). MLP method outperformed other methods in prediction of carbonates in soil samples. Saliency map from 2D-CNN deep learning method, also indicated hydromagnesite peaks in the samples and XRD verified the presence of hydromagnesite quantitatively. It should be noted that apparatus is utilized exactly as appeared in this figure.*

2. Related Work

Machine Learning (ML) techniques were introduced in chemometrics and soil spectroscopy as a means for information extraction and soil properties prediction. Both traditional ML and Deep Learning (DL) approaches have been extensively adopted (1,6,28) because of their intrinsic property to handle high dimensional data such as spectra. Deep neural networks, in particular, are capable of handling data in almost every form e.g. tabular data, text, images and audio, coupled with recent increases in computational power and the expansion of system memory capabilities, now make possible the use of spectral data to perform spectrogram-based analysis for soil properties predictions. As there are usually thousands of predictor variables (reflectance values), neural networks make an excellent choice for such a task because of their intrinsic trait to handle well data that suffer from high dimensionality. Convolutional Neural Networks (CNN) (29) have been used in image recognition with remarkable success and constitute one of the most powerful DL techniques for modeling complex processes such as pattern recognition in image based applications. Recently, Tsakiridis et al., (2020) utilized CNNs to predict multiple soil properties (30). The successful use of CNNs for soil property prediction without the use of pre-processed spectra was demonstrated in (31). They proposed the

representation of raw spectral data as a two-dimensional (2D) spectrogram and showed its superior performance over traditional ML techniques such as Partial least squares regression (PLSR) (32) and Cubist regression trees (33). Ng et al., (2019) used the combined spectra of soil samples to train 1D and 2D CNNs that both outperformed traditional ML techniques (34).

In our study, we compare different methodologies and techniques based on both computational regression techniques on NIR spectral data and experimental methods that usually apply on soil samples. Neural Networks algorithms outperformed the linear models for the soil carbonates content prediction as one term, and secondly, in the prediction of important NIR peaks. Classic mineralogical and analytical methods were also applied in soil samples establishing the feasibility of introducing neural networks regression methods in soil NIR spectra.

Much effort is given towards elucidating soil properties in all continents and all recorded data is valuable knowledge for today and future research. Moreover, data science applied to spectral and physicochemical libraries for the construction of calibration libraries is a novel approach redefining universal scientific effort. Non-destructive and scientific effective methods are a new path in modern soil research and is adding value from a spectroscopic point of view.

3. Methodology

This research introduces a novel and efficient and expeditious method for forecasting soil carbonates content utilizing non-destructive Fourier Transform Near-Infrared (FT-NIR) reflectance spectroscopy and advanced deep learning techniques. The study assesses the performance of two deep learning models—namely, a Multi-Layered Perceptron Regressor (MLP) and a Convolutional Neural Network (CNN. Model training is performed on the consolidated dataset from two Near-Infrared (NIR) spectral libraries: the Kellogg Soil Survey Laboratory (KSSL) of the United States Department of Agriculture (USDA) (https://ncsslabdatamart.sc.egov.usda.gov/), which is a service agency that has been collecting spectral and other data, measuring physicochemical soil properties and location characteristics to vis-NIR and MIR spectra across the United States of America, uploading and sharing the data, since 80 and more years until nowadays and the Land Use and Coverage Area Frame Survey (LUCAS) European Topsoil dataset (25) (https://esdac.jrc.ec.europa.eu/content/lucas2015-topsoil-data). Other similar spectral libraries include the Brazilian Soil Spectral Library (BSSL) (26) global soil datasets (1) and the Chinese vis–NIR soil spectral library (CSSL) (27).

The prediction models are trained on these extensive datasets and are then applied to predict carbonate content in previously unobserved soil samples collected from Xatzisavva wine fields in Alexandroupolis, Greece and

Vourvoukelis vineyards in Avdira, Greece. We compare predictions with the predictions of more conventional machine learning algorithms like Partial Least Squares Regression (PLSR), Cubist, and Support Vector Machines (SVM) which are considered the standards in this research field. To validate models' accuracy and performance, we also evaluate all models' results against the results of our laboratory volumetric methods (Figure 1).

3.1. Study Area

The investigated area falls within the so-called areas of wine fields of Vourvoukelis winery in Avdira in Xanthi, Greece (40º 54' 20'' N, 25º 48' 24''E) and of organic wine fields of Xatzisavva winery in Alexandroupolis, Greece (41º 00' 09'' N, 24º 55' 00'' E).

3.2. Sample pretreatment

The diffuse NIR reflectance spectra of air-dried and (<250 μm) sieved soil samples were measured in the laboratory. Before spectral measurements, the samples were placed in glass Petri dishes and spectral acquisition was performed with thickness of the sample of ~3-4 cm to avoid transmission effects, for all samples. Background spectra was subtracted from all acquired NIR spectra.

3.3. Spectral acquisition

A Perkin-Elmer Spectrum N Two FT-NIR spectrophotometer (quartz beamsplitter and $LiTaO_3$ NIR detector based on diode array) was used for spectral measurements in the spectral region of 1150–2500 nm. The spectral data was screened to ensure percentage reflectance measurements did not exceed theoretical limits, from 0.0 to 100.0. All spectra were exported to wavelength intervals of 0.5 nm. Absorbance spectra were calculated by means of equation (1).

3.4. Pretreatment of spectral libraries as calibration dataset

The soil spectra were transformed using three pretreatment methods prior to chemometric modeling, as the best treatment was not known a priori. This technique included the (pseudo) absorbance transformation, normalization of spectra between values 0-1, Savitzky-Golay smoothing filtering with a window size of 11, polyorder of 2 for the differentiating SG-1 and window size of 13, polyorder of 2 for the differentiating SG-2 (35) were applied to the spectral data to reduce noise and enhance spectral features to all calibration data of both KSSL-LUCAS 2015

TopSoil libraries combined, in order to reduce undesirable variance of the data to improve the predictive capacity of the calibration models. These procedures were applied to the original reflectance spectra (R) for KSSL (USDA) and TopSoil 2015 (Europe) and transformed to Absorbance (A) by the following equation:

$$A = \log_{10}(1/R) \quad (1)$$

No transmission (T) is observed due to the large thickness of the sample measured with the NIR spectrophotometer. Data screening resulted 28,615 samples for modeling with PLSR, Cubist, SVM, MLP and CNN models. Spectra of KSSL-USDA spectral library were transformed from percent reflectance to absorbance, by means of equation (1). The values of $CaCO_3$ were normalized to g/100g since LUCAS-TopSoil laboratory measurement unit of $CaCO_3$ is in g/kg.

In machine learning applications, it is a well-known fact that the size of the training set plays a crucial role in generalization. Generalization is the ability of the model to perform well on never-before-seen data. Typically, we aim for a large, diverse training set and to that end a decision was made to merge the two spectral libraries into a larger one. Before merging, we compare both spectral datasets based on the Wasserstein metric which is a common way to compare the probability distributions of two variables:

$$W_p(P,Q) = \left(\frac{1}{n}\sum_{i=1}^{n}\left|\left|X_{(i)} - Y_{(i)}\right|\right|^p\right)^{\frac{1}{p}} \quad (2)$$

where ***P, Q*** are probability distributions (one-dimensional) and ***p*** is the number of moments. The results show that the two libraries are similar enough to be merged (Wasserstein distance=1.78). Comparison of the new combined training dataset with the test set shows low similarity between them (Wasserstein distance=6.48), which raises a concern of how well our model will perform on the test set.

3.5. Laboratory carbonates measurement

Results throughout this study are given as CO_3^{2-} equivalents. The measurement is based on measuring the gaseous CO_2 released from carbonate reacting with 6M HCl and is expected to be equally effective at measuring soil carbonates released from the full range of carbonate soil minerals including common forms such as calcite, magnesite, and dolomite and their hydrous compounds.

Carbonates equivalents were determined at both locations by pressure calcimeter method treating the <250 μm soil fraction with 6M HCl in a closed vial. At KSSL this was volumetric method 4E1a1a1 (https://www.nrcs.usda.gov/sites/default/files/2023-01/SSIR42.pdf, accessed on 20 November 2022), and at LUCAS-Topsoil this method was similar volumetric method with ISO 10693:1995 (https://esdac.jrc.ec.europa.eu/public_path/shared_folder/dataset/66/JRC121325_lucas_2015_topsoil_survey_final_1.pdf, accessed on 21 November 2022).

3.6. Model Accuracy Evaluation

In order to evaluate the model accuracy, four statistical metrics were applied, namely the coefficient of determination (R2), the Root Mean Square Error of calibration (RMSE), the Residual Prediction Deviation (RPD), and the Ratio of Performance to Inter-Quartile distance (RPIQ) as shown in the following equations (2)-(5). The R^2 measures the percentage of variance of the dependent variable as influenced by the independent variable. The RPD is explained as the ratio of standard deviation of the measured reference values to RMSE and it is used for NIR spectra in soil science as a value of correctness of the model. Also, RPIQ is explained as the ratio of the standard deviation of the inter-quartile distance of the measured reference data to RMSE.

A combination of the R^2 and RPD statistical metrics is allowing predictions to be favored or not. In more details, when $R^2 > 0.90$ and RPD > 3.0, excellent prediction is provided by the model. When $0.82 < R^2 < 0.90$ and $2.5 < RPD < 3.0$ a good approximation is feasible by the model. In addition, moderate approximation is made when $0.66 < R^2 < 0.82$ and $2 < RPD < 2.5$. Finally, poor distinction of high and low values are performed when $R2 < 0.66$ and $RPD < 2$ as shown in (3).

$$\boldsymbol{R^2}(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

where y_i and $\hat{y}_i$ are measured values and predicted values, respectively; n is the number of samples in the training set,

$$\textbf{RMSE} = \sqrt{\frac{1}{N}\sum(y_{pred} - y_{obs})2} \quad (4)$$

where N is the sample size, y_{pred} is the predicted value, and y_{obs} is the observed value. Typically, the model with the lowest RMSE is chosen.

$$\mathbf{RPD} = \mathrm{STDEV(yobs)/RMSE} \quad (5)$$

RPIQ is the ratio of performance to inter-quartile distance of the reference data in the external validation set). RPIQ is proposed to be applied instead of RPD in soil samples sets, for which often show a skewed distribution. As a result, RPIQ is a better way to standardize the RMSE in terms of population spread compared to RPD. RPIQ is based on quartiles representing the Q1 as the value below which 25% of the samples is found, Q2 as the value below where 50% of the samples are found and Q3 as the value below where 75% of the samples are found. Such an approach as described in equation 5 is useful to determine equivalent ranges of population spread (3).

$$\mathbf{RPIQ} = \mathrm{IQ/RMSE} \quad (6)$$

where IQ = (Q3-Q1)

3.7 XRD Quantification

The samples were measured in a Bruker D8 Focus diffractometer with Bragg– Brentano configuration, operating at a voltage of 40 kV and an intensity of 40 mA. It contains a primary monochromator, working with a copper anticathode, Cu $K\alpha_1$ monochromatic radiation (λ = 1.54056 Å). Acquisition was performed in 2θ scanning mode covering the 5-90° range with steps/sec of 0.02°. Samples were grinded until the fine powder was less than 20 μm. The powder was mounted in a circular quartz holder. The phases included in the refinement were all the minerals that were identified by Profex and BGNM libraries according to their peak intensities.

The Rietveld method (Rietveld, 1969) is a method that theoretically adjusts the structural and experimental parameters to the complete powder diffractogram profile of the sample, considering it as the sum of the Bragg reflections that appear at respective angular positions form crystalline material in the sample. It is considered as a total approach to mineralogical quantification. In addition, $\chi^2 < 5$ was considered as acceptable refinement statistics based on the

complexity of soil samples examined in this study and based on previous similar work of clay material XRD analysis (36).

4. Algorithms and implementation

4.1 Partial Least Squares Regression

PLSR is a linear chemometric technique used for analysis of spectroscopic data for different applications. In our study, PLSR is used for the determination of soil carbonate content and is presented as a common modelling technique for quantitative spectroscopic analysis in soil mapping and classification as found in the literature (7,32,37). Decomposition of the spectral data into features (namely collective variables) is performed. The collective variables include most of the variance that exists in the reflectance NIR spectral data and thus linear models of the most correlated features are created.

$$X = T\,.\,P\hat{}T + Residuals(E) \qquad (7)$$

$$Y = T\,.\,C + Error(f) \qquad (8)$$

In PLSR, a decomposition of the ***X*** and ***Y*** variables with finding new latent variables and a selection of orthogonal factors that is maximizing the relation between prediction variables (X-soil reflectance) and response variables (Y-laboratory measured data) is performed. Components T that allows the decomposition of the predictors are searched by the PLSR (eq.1) and prediction of the response variables is also performed (equation 2). ***P*** and ***C*** are the factor loadings, and ***E*** and ***f*** are the residuals and errors matrices, respectively (38).

In our study, the PLSR was performed with the optimum 22 collective variables (CV's) in the existing NIR reflectance spectra acquired by the KSSL-USDA combined to LUCAS-TopSoil libraries (Figure 2).

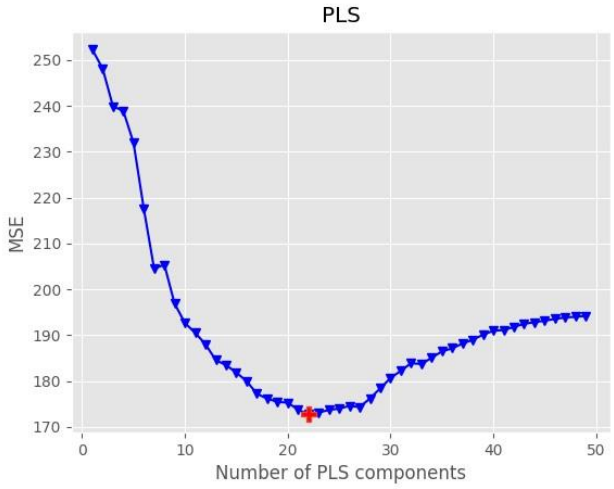


Figure 2: *Evolution of Mean Square Error (MSE) in progressing number of PLS components (collective variables, CV's) is showing the optimum selection of 22 collective variables (CV's) to best contribute to our PLSR model.*

4.2 Cubist

The Cubist chemometric technique is based on the M5 algorithm of Quinlan (33) and is widely and successfully applied in vis-NIR spectroscopy analysis. Thus, the Cubist method is considered as a competitive method to other methods of multivariate regression in terms of prediction accuracy.

The Cubist model is based on a regression tree construction with intermediate linear models extracted at each step of the procedure. It splits the original dataset that has similar attributes into subsets of sample and then generates multi-linear regression rules by optimal predictor variables selected from all spectral variables.

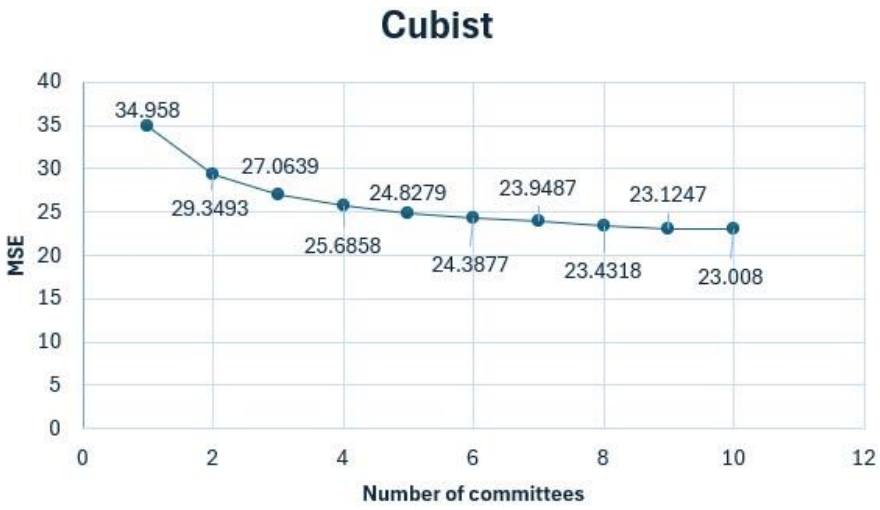


Figure 3: *Evolution of Mean Square Error (MSE) in progressing number of committees is showing the optimum selection of 10 committees to best contribute to our Cubist model.*

To optimize our Cubist model, we experimented on the number of committees (boosting). The results of our tests showed that after 10 committees, model performance saturates and we get diminishing returns (see Figure 3). Regarding the number of neighbors, it was kept to zero, based on our decision to not develop a composite model because the main focus of this research was on the novel deep learning methods. Furthermore, the number of trained models was ample enough for our research interest.

4.3 Support Vector Machines

Support vector machines (SVM) is a method that incorporates in its algorithm, linear equations for the regression analysis of multivariate cases (39,40).

The input parameters used for training the SVM are the NIR features that will be derived from the CV's calculated from the PLS regression model (CV's = 22). A linear kernel was used, and all spectra were normalized using a standard scaler (all values truncated between 0-1).

4.4 Multi-Layer Perceptrons

Multi-layer perceptrons (MLPs) or deep feedforward networks are machine learning models that use intermediate computations to transform their input $\boldsymbol{x}$ to the output $\boldsymbol{y}$ and in doing so evaluate a function f:

$$\boldsymbol{y} = f(\boldsymbol{x}; \boldsymbol{\theta}) \tag{9}$$

The transformations (linear and non-linear) that each model's layer implements to the data is parameterized by its weights $\boldsymbol{\theta}$. In this context, the model's goal is to find the set of values for the weights $\boldsymbol{\theta}$ of all layers in the network that correctly map inputs instances to their corresponding targets. Put in another way, the weights $\boldsymbol{\theta}$ that minimize the difference between the distribution of the output $\boldsymbol{y}$ and the true underlying distribution of the targets.

In contrast to linear models and to extend them to represent nonlinear functions of $\boldsymbol{x}$, MLPs apply the linear model to a transformed input $\boldsymbol{\phi}(\boldsymbol{x})$ and not to $\boldsymbol{x}$. The ultimate goal is to learn $\boldsymbol{\phi}$:

$$y = f(\boldsymbol{x}; \boldsymbol{\theta}; \boldsymbol{w}) = \boldsymbol{\phi}(\boldsymbol{x}; \boldsymbol{\theta})\text{^}(\top \boldsymbol{w}) \tag{10}$$

where $\boldsymbol{\theta}$ are parameters used to learn $\boldsymbol{\phi}$ from a broad family of functions and parameters $\boldsymbol{w}$ that map from $\boldsymbol{\phi}(\boldsymbol{x})$ to the desired output. The choice of how to represent the output determines the form of the cost function which usually expresses the difference between the predicted and the true distribution. In most cases the parametric model defines a distribution $p(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta})$ and the principle of maximumm likelihood is used:

$$J(\theta) = -E_{x,y\sim\widehat{p_{data}}} log p_{model}(y|x) \tag{11}$$

In such cases, the cost function is the negative log-likelihood between targets and model's predictions. A simpler approach is to merely predict a statistic of y conditioned on x. Note that the cost function will often combine a regularization term such as weight decay or dropout layers in order to avoid overfitting.

Our MLP model consists of three layers of 500, 200 and 50 units respectively (Figures 4, 5). Since we are predicting only one soil attribute ($CaCO_3$) the model has a single output. To introduce non-linearity we apply the Rectified Linear Unit (ReLU) activation function and as a measure to mitigate overfitting L_1 and L_2 regularizers are applied to the third layer. We use the Adam optimizer (41) with a decaying learning rate. Savitzky-Golay second derivative smoothing filter is applied to the spectral dataset before being fed to the network.

To train a neural network, usually one has to split their dataset in three smaller datasets: a) a training set that will be used to train the network and fix its weights (training sets are usually large in size and preferably diverse because training set diversity contributes to better generalization) b) a validation set which size is typically ~20% of training's set size, and is used to test the model's performance c) the unknown set which contains data samples that the model has never "seen" before. The performance of the model on the unknown dataset is the key objective.

In our study, we performed an 80/20 training/validation dataset split and used 10-fold validation. K-fold validation is a technique typically used when the training set is relatively small, however, it has been proved that it yields slightly better predictive results even for larger datasets. For our experiments we combined both KSSL and LUCAS datasets. Before merging the two datasets, the Wasserstein test statistic was performed to gain insight on the "resemblance" of the two datasets. The result showed that the two datasets are very similar to each other (Wasserstein distance=1.78). High dataset similarity is often considered a good indicator for dataset merging. On the other hand, a training set should be diverse and present a relatively high degree of variability to avoid overfitting.

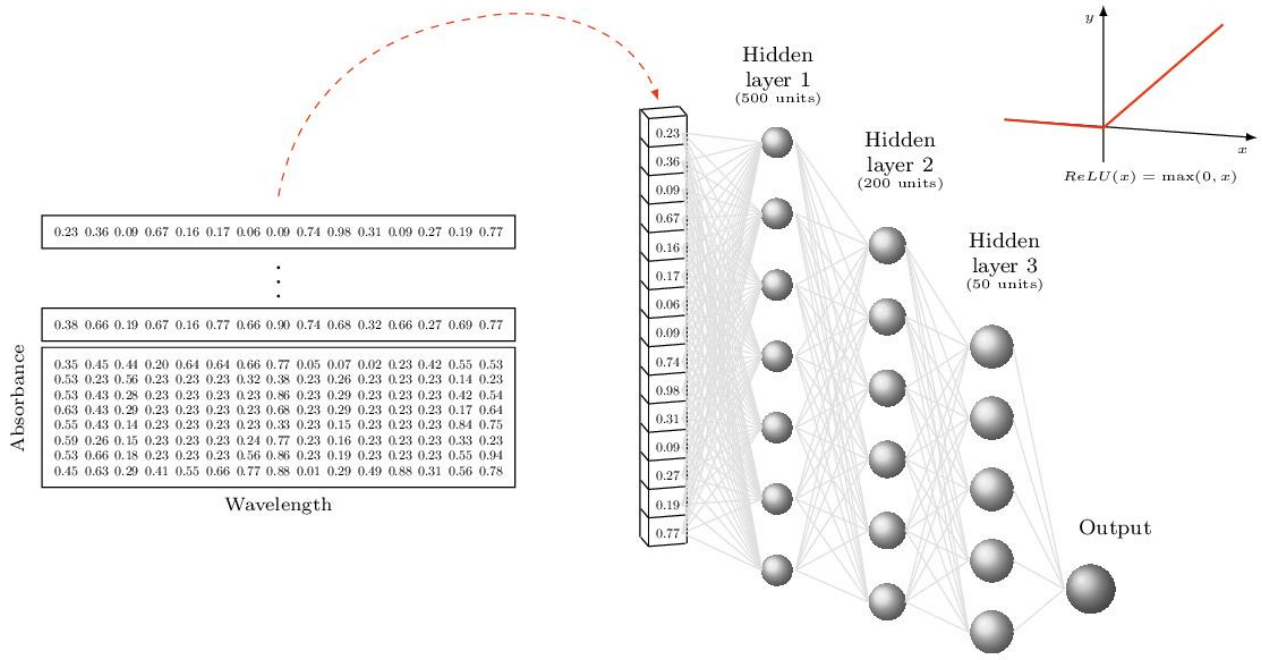


Figure 4: *MLP architecture. Three consecutive hidden layers (with 500, 200, 50 units) are filtering the combined spectral dataset in the NIR region, to a final output. An activation function ReLU is applied to facilitate training of the MLP model.*

4.5 Convolutional Neural Networks

Convolutional neural networks (CNN) are neural networks specialized in grid-like topology data i.e. images that are represented as 2D grid of pixels. Their architecture consists of stacked convolutional layers of multiple filters that convolve with their input to produce a series of feature maps. Generally, a convolution is an operation on two functions:

$$s(t) = \int (x * w)(t) \tag{12}$$

where x is the input function and w, also known as the kernel (or filter), needs to be a valid density function for the output to be a valid weighted average of x over time t.

When an instance passes through a CNN layer it usually undergoes three stages: first, the layer applies several convolutions in parallel to produce a set of linear activations. Then, each linear activation passes through a nonlinear activation function, such as a rectified linear unit. Finally, the output is further modified by a pooling function which serves both as a downsampling technique to reduce statistical and computational burden (42) as means to make the representation become approximately invariant to small translations of the input (42). Besides being translation invariant, CNNs learn spatial hierarchies of patterns: lower layers learn smaller patterns whereas layers near the output learn more complex patterns and abstract visual concepts (43).

These traits make CNNs an excellent choice for machine vision tasks such as object detection, image classification and semantic segmentation. Although designed for multidimensional data e.g. images and CT scans, CNNs perform equally good on one-dimensional inputs. In fact, CNNs have been used with relative success in tasks that involve sequences e.g. time-series. However, in our case we applied CNNs to the sample spectrograms (2D images) although we trained a CNN on spectra which yielded results on par with the MLP Regressor.

Our CNN model is made of three convolutional layers of 32, 64 and 128 filters respectively, each succeeded by a max-pooling layer (F). All convolutional and max-pooling kernels are 3X3 size. A dense layer of 50 units sits on top of the convolutional stack which in turn is connected to the network's single output. As in the MLP approach, we use ReLU activation function and Adam optimizer. In this approach, the network's input is not spectra but their respective spectrograms, each with dimensions 244X488 pixels. It is easy to convert NIR spectral absorbances into two-dimensional plots: NIR wavelengths represent the x-axis of the plot and absorbance values the y-axis. After conversion, the plots are stored as images and fed to the convolutional network. It should be considered that axis scale must be the same for all the plots.

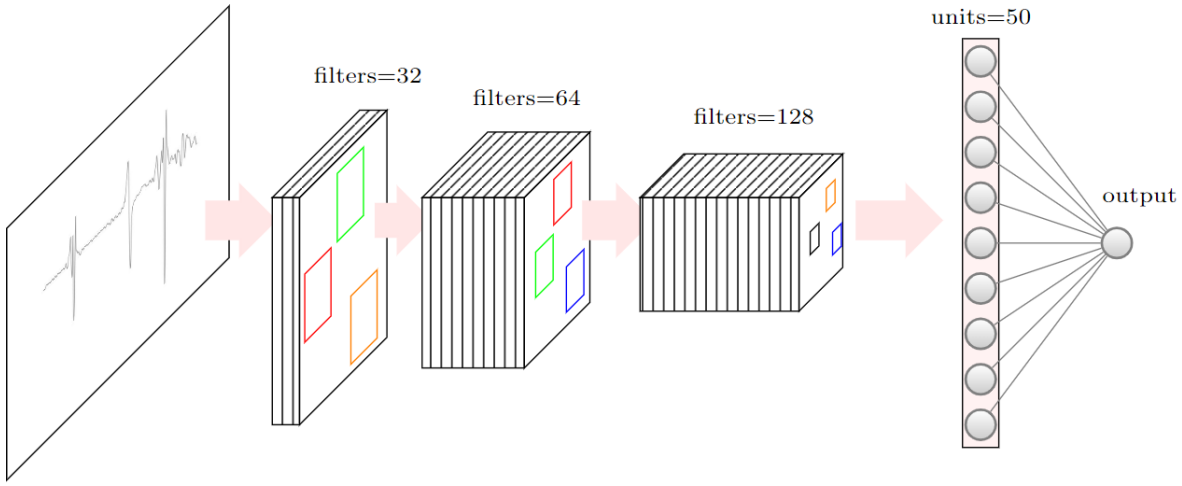
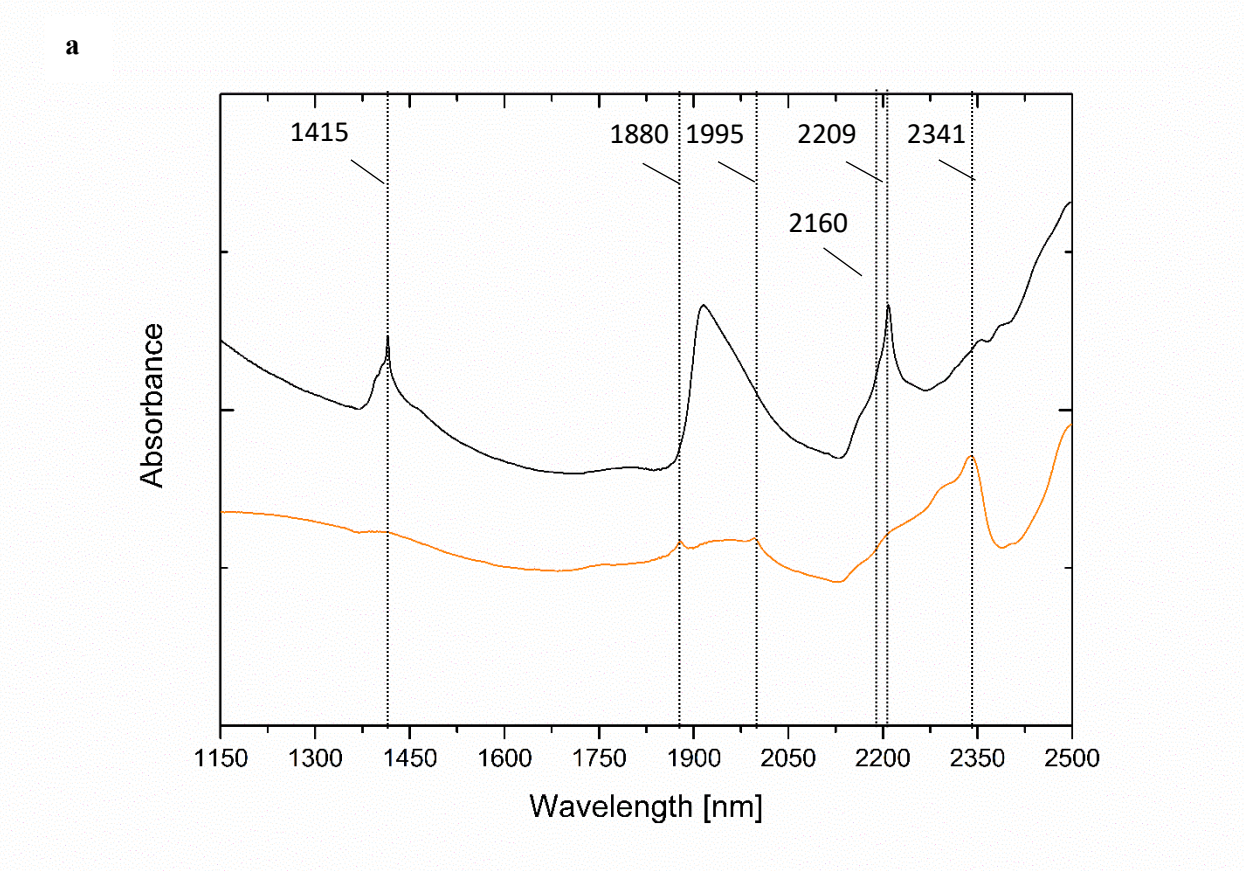


Figure 5: *CNN architecture. Three consecutive convolutional layers (with 32, 64, 128 units) are filtering the combined spectral dataset as images in the NIR region, with dimensions 244x488 pixels, to a final output. An activation function ReLU is also applied to facilitate training of the CNN model.*

5. Results

Our predictive machine learning models, classical (PLSR, Cubist, SVM) and deep learning ones (MLP and CNN), trained on the combined two large spectral databases, achieve high performance results (R^2 = 0.84, RPD = 2.14 for MLP) when applied to the unknown set: R^2 = 0.68 and RPD = 1.47 for CNN, for the prediction accuracy of soil carbonates, in the second derivative of NIR spectra (Figure 7).

Soil samples (no.45) of both Xatzisavva (SAM-1) wine fields in Alexandroupolis in Greece and Vourvoukelis (SAM-2) in Avdira in Greece were collected and samples (no.19) from SAM-1 and SAM-2 groups exhibited non-zero values in carbonate content, using the volumetric method as described in the methods section and were recorded (g/100g). As shown in Table 1, at least one soil sample from SAM-1 group is rich in carbonates (highest content = 18.14%) and the poorest carbonates content is in SAM-2 group (lowest content = 0.04%).


a
1415
1880
1995
2209
2341
2160
Absorbance
1150
1300
1450
1600
1750
1900
2050
2200
2350
2500
Wavelength [nm]

b

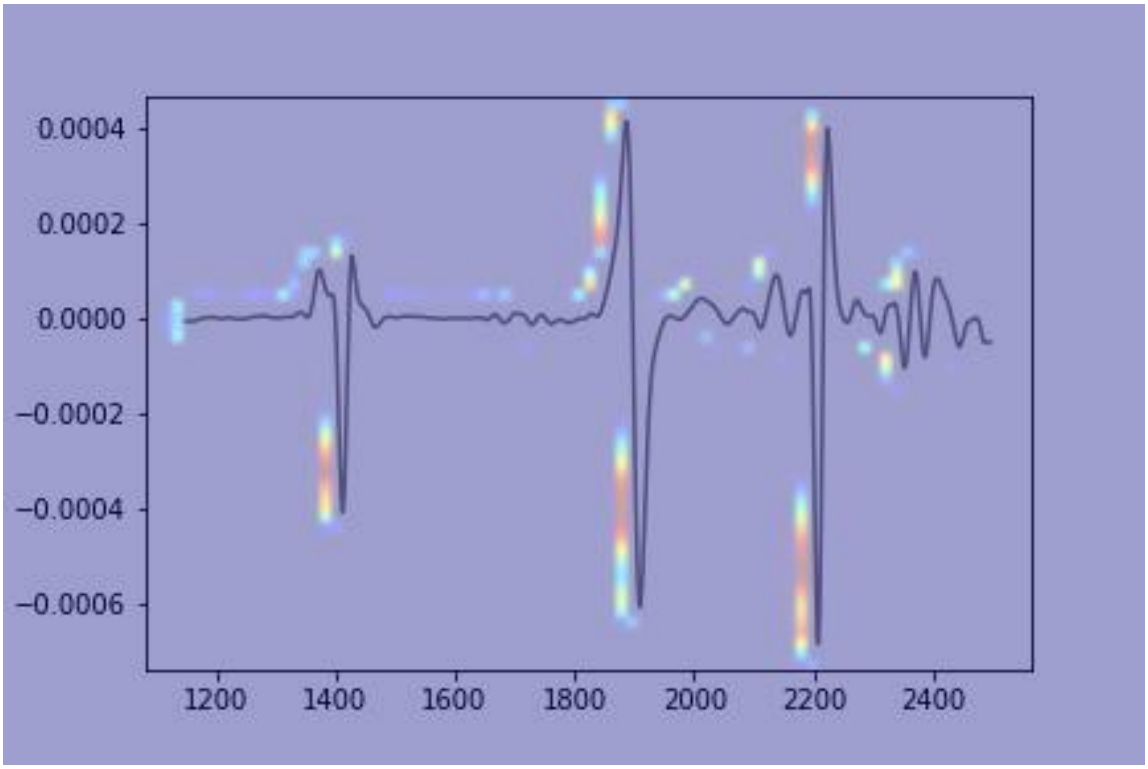


Figure 6: *On the top (a) absorbance spectra of an indicative soil sample (s03) (black) of Asyrtiko species from the Xatzisavva wine field in Alexandroupolis and calcium carbonate (Sigma, 239216) white powder spectrum (orange). Indication lines of carbonates vibration peaks are also shown as dashed lines. (b) a saliency map from CNN model in second derivative spectra, is showing the most favorable peaks that are considered to trigger the CNN model in a color gradient from red (most favored peaks) to light blue (less favored peaks).*

Table 1: *Experimental values of diverse soil samples (no.19) using the volumetric method in comparison with the predicted carbonates values from MLP model. Two groups of soil samples (SAM-1 and SAM-2) were measured and predicted.*

S/no	Exp g/100g	Pred g/100g	
S01	7.85	11.53	
S02	1.83	0.86	
S03	8.56	8.80	
S04	10.84	5.85	
S05	9.32	8.37	
S06	4.63	1.64	
S07	3.39	1.52	
S08	5.75	3.31	**SAM-1**
S09	1.18	1.07	
S10	18.14	13.44	
S11	11.28	14.18	
S12	17.32	14.38	
S13	17.00	18.57	
S14	1.12	1.80	
S15	0.07	0.60	
S16	0.13	0.59	
S17	0.13	0.53	**SAM-2**
S18	0.09	0.51	
S19	0.04	0.51	

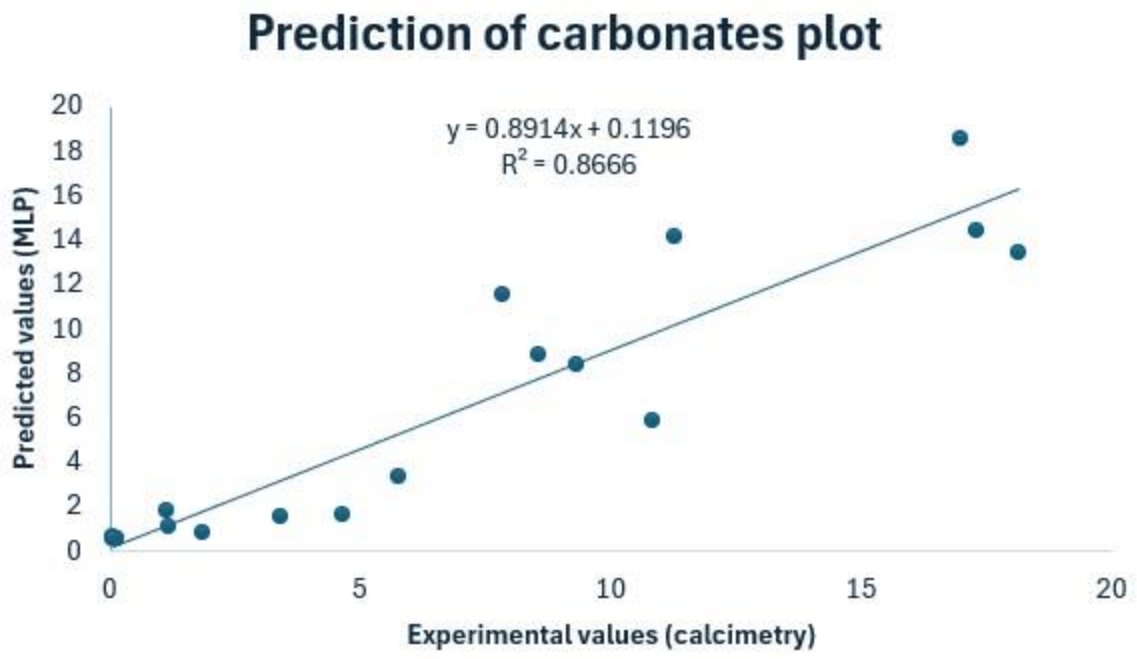
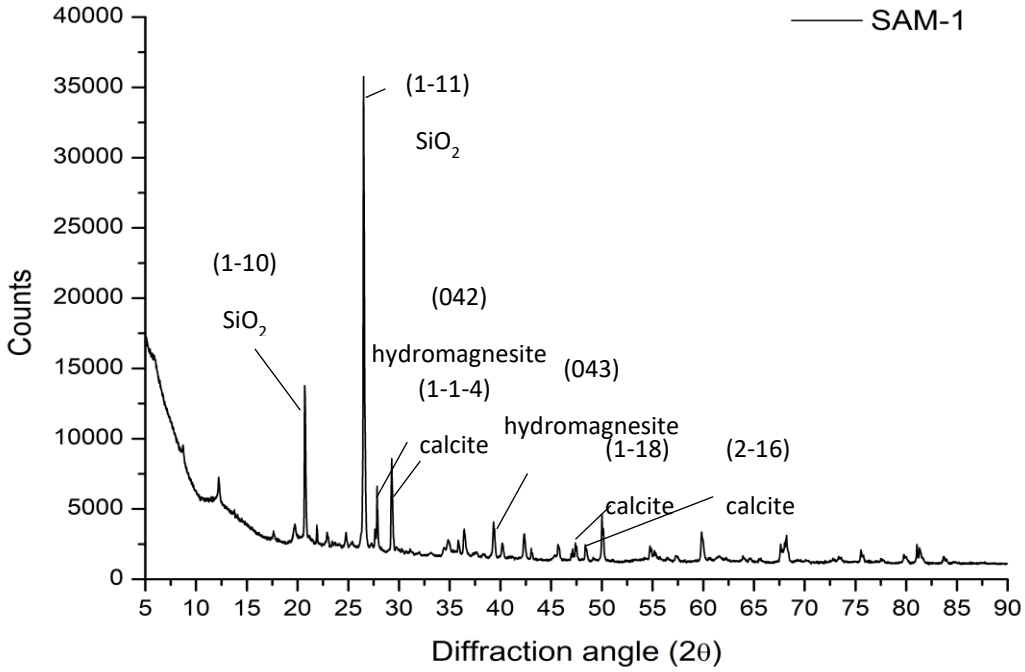


Figure 7: *Plot of carbonates prediction with MLP vs Calcimetry (Volumetric method) of Table 1, exhibits $R^2 = 0.87$. MLP deep learning methods calibrated throughout the combined dataset of KSSL-USDA NIR spectral library and LUCAS-2015-TopSoil NIR spectral library outperformed other regression methods utilized in our work. MLP performs well, even below the Limit of Detection of the Volumetric method that is 2.5 g/100g. All values reported are in g/100g.*

Table 2: *Regression prediction models applied to all data for carbonates (two spectral libraries, firstly 6,833 NIR spectra from KSSL (USA) combined with 21,782 NIR spectra from Lucas TopSoil (all European fields) for carbonates). Statistics of R^2, Root Mean Square Error (RMSE), Residual Prediction Deviation (RPD) and Ratio of Performance to Inter-Quartile distance (RPIQ) are shown. From all pretreatments, the second derivative outperformed the other two methods of absorbance and first derivative spectra.*

	PLSR	**SVM**	**Cubist**	**MLP**	**CNN (images)**
Second Derivative	R^2 = -1.43 RMSE = 13.11 RPD = 0.54 RPIQ = 0.82	R^2 = -0.19 RMSE = 6.59 RPD = 0.91 RPIQ = 1.43	R^2 = 0.86 RMSE = 4.80 RPD = 2.70 RPIQ = 2.13	R^2 = 0.87 RMSE = 2.11 RPD = 2.14 RPIQ = 3.33	R^2 = 0.68 RMSE = 4.11 RPD = 1.47 RPIQ = 2.29

In order to describe in full detail, the carbonates content of the SAM-1 (sample S03), and SAM-2 (S18) we took insight in the quantification of the minerals by means of powder X-ray diffraction (XRD) method. Diffractograms were acquired in an in-house diffractometer, as mentioned in the methods section. Rietveld refinement using structural models was then performed (Figure 8).



F

S

Quantification of carbonates minerals in crystalline phase, in the representative soil sample (S03) of SAM-1 group, shows that total carbonates are contained in 6.07 % (w/w) with a crystalline index = 0.72. The total carbonates content is then calculated to 8.43 % (w/w). The minerals found were firstly calcite 3.88 % (w/w), and also hydromagnesite 2.19 %, as shown in Table 3. A total of 8.43 % (w/w) in ratio of total minerals and in specific of both their crystalline phases of SAM-1 group and amorphous content based on crystalline index, is in good agreement with both volumetric method and MLP prediction of NIR spectra based on the calibration libraries of KSSL (USA) and LUCAS-Topsoil (EU).

Table 3: *Comparison of total carbonates content of S03 from SAM-1 group by volumetric method, MLP prediction and XRD phase quantification of both calcite and hydromagnesite minerals shows a very good values agreement.*

SAM-1	**Molecular formula**	**Phase quantification from XRD (wt-%)**	**Volumetric method (wt-%)**	**MLP (wt-%)**
Calcite	$CaCO_3$	3.88	-	-
Hydro-magnesite	$Mg_5(CO_3)_4(OH)_2 \cdot 4H_2O$	2.19	-	-
Total carbonates	CO_3^{2-}	6.07 (Crystalline Index = 0.72) **Total = 8.43**	**8.56**	**8.80**

In SAM-2 group, the carbonates content is negligible, as shown in Figure 9 and the XRD diffractogram where all peaks of calcite are absent, at 29.4º for (1-1-4) calcite plane and around 47.4º and 48.5º for (2-2-2) and (1-18) planes, respectively. Also, hydromagnesite is absent at 27.7º for (042) and at 39.2º for (080).

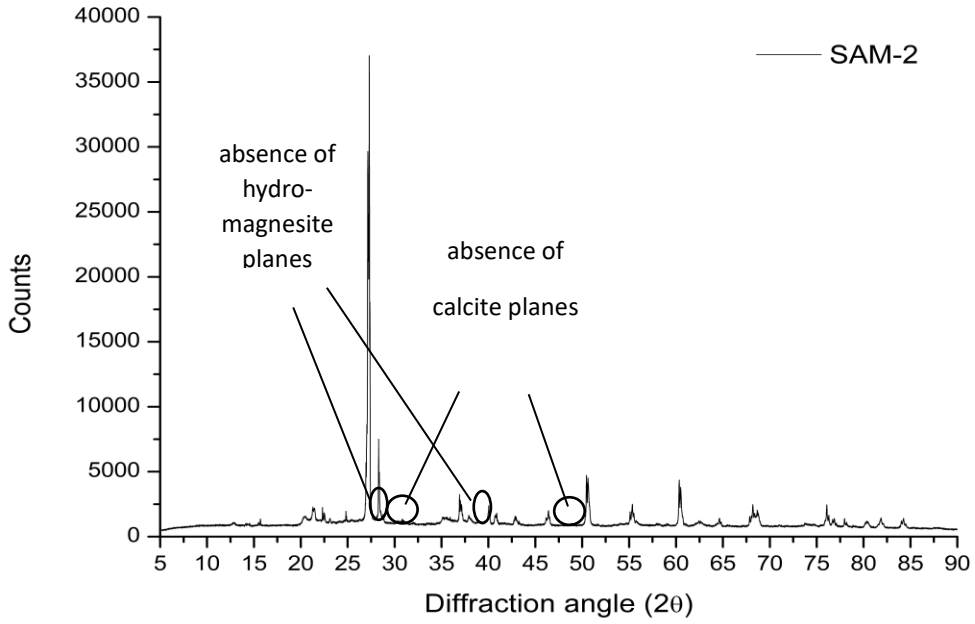


Figure 9: *Absence of calcite and hydromagnesite crystalline phases in diffractogram for SAM-2 group representative sample (S18).*

6. Discussion

Our computational machine learning models trained on two extensive spectral databases: the USDA's KSSL spectral library and the European Union's LUCAS 2015 TopSoil spectral library demonstrated impressive performance metrics (R^2 = 0.87, RPD = 2.14 for MLP) in the prediction set, and (R^2 = 0.68, RPD = 1.47 for CNN) specifically in predicting soil carbonates.

Notably, the MLP exhibits remarkable accuracy, even in predicting extreme experimental values of soil carbonates, below the limit of detection (LOD) of the volumetric method. Its predictions align well with the quantification achieved through X-ray Diffraction and the volumetric method.

Soil carbonates, primarily composed of calcium and magnesium carbonates, manifest distinctive absorption peaks in the NIR region. These peaks stem from the overtone and combination bands of the fundamental bands of CO_3^{2-}, prevalent in the mid-IR region. In the NIR, carbonates exhibit five characteristic bands. Notably, a conspicuous peak around 2340 nm represents an overtone of the asymmetrical stretching v_3 observed at 1415 cm^{-1} in the mid-IR region. Additionally, the peak near 2500-2550 nm corresponds to a combination of symmetrical stretching (v_1) and the first

overtone of asymmetrical stretching (v_3), denoted as ($v_1 + 2v_3$). Weaker absorptions are evident near 1415 nm due to crystallized water, around 1900 nm ($v_1 + 3v_3$), near 2000 nm ($2v_1 + 2v_3$), and around 2160 nm ($3v_1 + 2v_4$), where v_4 represents the in-plane bending (v_4, 680 cm-1). It's noteworthy that the positions of these absorption bands vary based on the composition of the carbonates. While few studies have quantified carbonate composition in soil utilizing vis-NIR spectroscopy for estimating carbonate content, a recent study examined soil carbonates content using mid-IR spectroscopy, employing the KSSL library for calibration (44).

In an attempt to determine which wavelength absorbance peaks, possess more weight on the model's performance and due to the fact that it is easier to visualize this through saliency mappings (peaks on the second derivative absorbance spectra) we develop a CNN model as discussed previously. The most favored peaks are in the 1415 nm, in the 1908 nm, in the 2209 nm and in the 2335 nm. The sharp band at 1415 nm that is shown in soil NIR spectra in Figure 6, is due to calcite and hydromagnesite ($Mg_5(CO_3)_4(OH)_2 \cdot 4(H_2O)$) that is a mineral also present in calcareous soils especially in SAM-1 group and absent in SAM-2 group. Hydrous carbonates, like other hydrous minerals, typically exhibit strong bands around 1400 nm due to O–H stretching or a combination of the symmetric H–O–H stretch and H–O–H bend (45). The shape of the absorption peaks can indicate whether this feature is produced by a hydroxyl or water group. The characteristic band near 1415 nm is also due to kaolinite that shows absorption wavelengths near 1400 nm (1395 and 1415 nm) that are overtone vibrations of the O–H stretch near 2778 nm (3600 cm^{-1}), and can to one part be attributed to the first overtone of structural O–H stretching mode in its octahedral layer (46). The 1880 nm band is due to $v_1 + 2v_3$, where v_1 is the totally symmetric C-O stretch and v_3 is the doubly degenerate antisymmetric C-O stretching mode occurring near 7000 nm (47).

Also, at 1908 nm occurs a combination of H–O–H bending and the asymmetrical stretching fundamentals. This band is possible overlapping with the main band of other minerals and in particular O-H stretch of kaolinite. Also, the peak near 2209 nm is triggered by the presence of calcite. Lastly, around 2235 nm a prominent peak of calcite triggers the saliency map derived from CNN model. Other peaks that are triggered by the content of carbonates by visual inspection of the saliency map, are difficult to assign because of various peak overlaps with hydroxyl and/or water molecules bound to mineral crystallites. As a result, NIR spectral assignment towards elucidation of carbonates peaks is a multiparametric task, especially in complex samples such as soil samples.

In this study, soil samples were collected from two wine fields in Greece: Xatzisavva (SAM-1) in Alexandroupolis and Vourvoukelis (SAM-2) in Avdira, Xanthi. Carbonate content was determined using the volumetric method, recorded as g/100g. Table 1 reveals that at least one soil sample from the SAM-1 group has notably high carbonate content (highest content = 18.14%), while the SAM-2 group exhibits the lowest carbonate content (lowest content = 0.04%). To identify influential wavelength absorbance peaks in the model's performance, saliency mappings were developed through a CNN model. The peaks with the highest significance were observed at 1415 nm, 1880 nm, 1995 nm, and 2209 nm. The sharp band at 1415 nm corresponds to minerals like calcite, monohydrocalcite ($Ca(CO_3) \cdot (H_2O)$), and hydromagnesite ($Mg_5(CO_3)_4(OH)_2 \cdot 4(H_2O)$), prevalent in calcareous soils, particularly in the SAM-1 group and absent in the SAM-2 group.

In addition, the ratio of calcite/hydromagnesite in our soil samples is also calculated by means of X-ray diffraction (XRD). The two farms from which the soil samples were collected, have different farming conditions of operations and handling of soil, with Vourvoukelis farm introducing additional fertilizer in soil while Xatzisavva's operational practice is pure organic. The ratio of calcite/hydromagnesite of soil fertility due to hydrous magnesite complexes that are present in addition to calcite, are important and totally detectable by XRD and saliency maps of 2D-CNN method (Figures 6, 9). The elegant method of 2D-CNN in the analysis of the combined dataset of KSSL and LUCAS-2015 spectral libraries is reported for the first time to our knowledge, exhibiting the major peaks of influence in the regression and in the prediction. The prediction of secondary peaks (hydromagnesite in our prediction set) is in great accordance with the XRD quantification in two samples with different origin (namely SAM-1 and SAM-2).

This research significantly contributes to the field by offering a swift prediction approach for soil carbonates content, particularly in scenarios where a volumetric method is unavailable, and only visible-NIR spectra absorbance data are at hand.

To the best of our knowledge, this study is pioneering, presenting the first prediction model trained on such a comprehensive dataset, showcasing promising outcomes on previously unseen data. These findings robustly affirm the efficacy of deep learning models as potent tools for predicting soil carbonates content and the potency to identify NIR peaks of hydrous minerals by saliency map of 2D-CNN method. The ratio and in particular the quantification of calcite/hydrous compounds (hydromagnesite) is verified by XRD method.

7. Conclusion

Carbonates content of soil is an essential soil chemical property, with a significant impact on growth and production of crops. Soil genesis records and soil classifications as shown in various soil profiles are also mostly affected by the distribution of carbonates content of soils. Here we present a modified way of soil carbonates prediction other than classical volumetric method. A minimal input of a NIR spectra provides the total carbonates content in soil samples, as calibrated with two spectral libraries, namely USDA-NRCS-KSSL and EU-LUCAS-2015-TopSoil with 28,615 NIR spectra, where MLP method is trained in this dataset and applied for prediction. MLP method performed good in a diverse test set ranging from diverse carbonates content as low as 0.07%, to high content (18.14%) in soil samples. CNN method produced less favorable results in terms of prediction with experimentally derived carbonates values and a saliency map showed us the peaks triggered in training and prediction spectral sets. Peak assignment of various carbonates and hydrous carbonates minerals in the NIR region was also established. The presence of hydromagnesite in soil samples with highest carbonates content (SAM-1) was verified by means of XRD Rietveld analysis.

Quantification of the total carbonates content in SAM-1 group with XRD, and in representative sample S03, is in good agreement with MLP prediction and volumetric method applied, while the absence of carbonates content is obvious in SAM-2 group with XRD analysis.

8. Acknowledgments

This study was supported by the project "AGRO4+" - Holistic approach to Agriculture 4.0 for new farmers" (MIS 5046239) which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

Authors would like to thank Dr. Nikolaos Kazakis, Dr. Nestor Tsirliganis and Dr. Chairi Kiourt for access to instrumentation and the staff at the USDA-NRCS-KSSL and EU-LUCAS-TopSoil for the data provided in this study.

Code availability section

Carbonates_prediction

Contact: tamvakis.petros@ac.eap.gr, +1-262-960-2449 (USA)

Hardware requirements: CPU/GPU

Program language: Python 3

Software required: https://www.tensorflow.org/ and dependencies therein.

Program size: 110 KB

The source codes are available for downloading at the link: https://github.com/petamva/carbonates_prediction

References

1. Viscarra Rossel RA, Behrens T, Ben-Dor E, Brown DJ, Demattê JAM, Shepherd KD, et al. A global spectral library to characterize the world's soil. Earth-Sci Rev. 2016 Apr;155:198–230.

2. Guerrero C, Wetterlind J, Stenberg B, Mouazen AM, Gabarrón-Galeote MA, Ruiz-Sinoga JD, et al. Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? Soil Tillage Res. 2016 Jan;155:501–9.

3. Bellon-Maurel V, Fernandez-Ahumada E, Palagos B, Roger JM, McBratney A. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. TrAC Trends Anal Chem. 2010 Oct;29(9):1073–81.

4. Nocita M, Stevens A, van Wesemael B, Aitkenhead M, Bachmann M, Barthès B, et al. Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring. In: Advances in Agronomy [Internet]. Elsevier; 2015 [cited 2022 Dec 19]. p. 139–59. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0065211315000425

5. Bai Z, Xie M, Hu B, Luo D, Wan C, Peng J, et al. Estimation of Soil Organic Carbon Using Vis-NIR Spectral Data and Spectral Feature Bands Selection in Southern Xinjiang, China. Sensors. 2022 Aug 16;22(16):6124.

6. Morellos A, Pantazi XE, Moshou D, Alexandridis T, Whetton R, Tziotzios G, et al. Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. Biosyst Eng. 2016 Dec;152:104–16.

7. Kooistra L, Wehrens R, Leuven RSEW, Buydens LMC. Possibilities of visible–near-infrared spectroscopy for the assessment of soil contamination in river floodplains. Anal Chim Acta. 2001 Nov;446(1–2):97–105.

8. Askari MS, O'Rourke SM, Holden NM. A comparison of point and imaging visible-near infrared spectroscopy for determining soil organic carbon. J Infrared Spectrosc. 2018 Apr;26(2):133–46.

9. Pribyl DW. A critical review of the conventional SOC to SOM conversion factor. Geoderma. 2010 May;156(3–4):75–83.

10. Conforti M, Matteucci G, Buttafuoco G. Using laboratory Vis-NIR spectroscopy for monitoring some forest soil properties. J Soils Sediments. 2018 Mar;18(3):1009–19.

11. Barthès BG, Kouakoua E, Moulin P, Hmaidi K, Gallali T, Clairotte M, et al. Studying the Physical Protection of Soil Carbon with Quantitative Infrared Spectroscopy. J Infrared Spectrosc. 2016 Jun;24(3):199–214.

12. Jia S, Yang X, Zhang J, Li G. Quantitative Analysis of Soil Nitrogen, Organic Carbon, Available Phosphorous, and Available Potassium Using Near-Infrared Spectroscopy Combined With Variable Selection. Soil Sci. 2014 Apr;179(4):211–9.

13. Reeves JB, Van Kessel JAS. Investigations into near Infrared Analysis as an Alternative to Traditional Procedures in Manure Nitrogen and Carbon Mineralisation Studies. J Infrared Spectrosc. 1999 Jun;7(3):195–212.

14. Cho RK, Lin G, Kwon YK. Nondestructive Analysis for Nitrogens of Soils by near Infrared Reflectance Spectroscopy. J Infrared Spectrosc. 1998 Jan;6(A):A87–91.

15. Reeves JB, McCarty GW, Meisinger JJ. Near Infrared Reflectance Spectroscopy for the Analysis of Agricultural Soils. J Infrared Spectrosc. 1999 Jun;7(3):179–93.

16. Morón A, Cozzolino D. Application of near Infrared Reflectance Spectroscopy for the Analysis of Organic C, Total N and pH in Soils of Uruguay. J Infrared Spectrosc. 2002 Jun;10(3):215–21.

17. Miltz J, Don A. Optimising Sample Preparation and near Infrared Spectra Measurements of Soil Samples to Calibrate Organic Carbon and Total Nitrogen Content. J Infrared Spectrosc. 2012 Dec;20(6):695–706.

18. Genot V, Colinet G, Bock L, Vanvyve D, Reusen Y, Dardenne P. Near Infrared Reflectance Spectroscopy for Estimating Soil Characteristics Valuable in the Diagnosis of Soil Fertility. J Infrared Spectrosc. 2011 Apr;19(2):117–38.

19. Knadel M, Stenberg B, Deng F, Thomsen A, Greve MH. Comparing Predictive Abilities of Three Visible-Near Infrared Spectrophotometers for Soil Organic Carbon and Clay Determination. J Infrared Spectrosc. 2013 Feb;21(1):67–80.

20. Hummel JW, Sudduth KA, Hollinger SE. Soil moisture and organic matter prediction of surface and subsurface soils using an NIR soil sensor. Comput Electron Agric. 2001 Aug;32(2):149–65.

21. Mouazen AM, De Baerdemaeker J, Ramon H. Towards development of on-line soil moisture content sensor using a fibre-type NIR spectrophotometer. Soil Tillage Res. 2005 Jan;80(1–2):171–83.

22. Chodak M, Khanna P, Horvath B, Beese F. Near Infrared Spectroscopy for Determination of Total and Exchangeable Cations in Geologically Heterogeneous Forest Soils. J Infrared Spectrosc. 2004 Oct;12(5):315–24.

23. Liu L, Ji M, Buchroithner M. Transfer Learning for Soil Spectroscopy Based on Convolutional Neural Networks and Its Application in Soil Clay Content Mapping Using Hyperspectral Imagery. Sensors. 2018 Sep 19;18(9):3169.

24. Guo Y, Amundson R, Gong P, Ahrens R. Taxonomic Structure, Distribution, and Abundance of the Soils in the USA. Soil Sci Soc Am J. 2003 Sep;67(5):1507–16.

25. Orgiazzi A, Ballabio C, Panagos P, Jones A, Fernández-Ugalde O. LUCAS Soil, the largest expandable soil dataset for Europe: a review. Eur J Soil Sci. 2018 Jan;69(1):140–53.

26. Demattê JAM, Paiva AFDS, Poppiel RR, Rosin NA, Ruiz LFC, Mello FADO, et al. The Brazilian Soil Spectral Service (BraSpecS): A User-Friendly System for Global Soil Spectra Communication. Remote Sens. 2022 Feb 5;14(3):740.

27. Shi Z, Wang Q, Peng J, Ji W, Liu H, Li X, et al. Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. Sci China Earth Sci. 2014 Jul;57(7):1671–80.

28. Stevens A, Nocita M, Tóth G, Montanarella L, van Wesemael B. Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy. Chen HYH, editor. PLoS ONE. 2013 Jun 19;8(6):e66409.

29. Lecun Y. Generalization and network design strategies. In: Connectionism in perspective. R. Pfeifer; Z. Schreter; F. Fogelman; L. Steels. Zurich, Switzerland: Elsevier; 1989.

30. Tsakiridis NL, Keramaris KD, Theocharis JB, Zalidis GC. Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network. Geoderma. 2020 May;367:114208.

31. Padarian J, Minasny B, McBratney AB. Using deep learning to predict soil properties from regional spectral data. Geoderma Reg. 2019 Mar;16:e00198.

32. Martens H, Næs T. Multivariate calibration. Chichester [England] ; New York: Wiley; 1989. 419 p.

33. Quinlan JR. Learning with Continuous Classes. In: Proceedings of Australian Joint Conference on Artificial Intelligence. Hobart; 1992. p. 343–8.

34. Ng W, Minasny B, Montazerolghaem M, Padarian J, Ferguson R, Bailey S, et al. Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. Geoderma. 2019 Oct;352:251–67.

35. Savitzky Abraham, Golay MJE. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal Chem. 1964 Jul 1;36(8):1627–39.

36. Cuevas J, Cabrera MÁ, Fernández C, Mota-Heredia C, Fernández R, Torres E, et al. Bentonite Powder XRD Quantitative Analysis Using Rietveld Refinement: Revisiting and Updating Bulk Semiquantitative Mineralogical Compositions. Minerals. 2022 Jun 17;12(6):772.

37. Viscarra Rossel RA. ParLeS: Software for chemometric analysis of spectroscopic data. Chemom Intell Lab Syst. 2008 Jan;90(1):72–83.

38. Viscarra Rossel RA, Walvoort DJJ, McBratney AB, Janik LJ, Skjemstad JO. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma. 2006 Mar;131(1–2):59–75.

39. de Santana FB, Otani SK, de Souza AM, Poppi RJ. Comparison of PLS and SVM models for soil organic matter and particle size using vis-NIR spectral libraries. Geoderma Reg. 2021 Dec;27:e00436.

40. Suykens JAK, Vandewalle J, De Moor B. Optimal control by least squares support vector machines. Neural Netw. 2001 Jan;14(1):23–35.

41. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. 2014 [cited 2023 Mar 28]; Available from: https://arxiv.org/abs/1412.6980

42. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, Massachusetts: The MIT Press; 2016. 775 p. (Adaptive computation and machine learning).

43. Chollet F. Deep learning with Python. Second edition. Shelter Island: Manning Publications; 2021. 478 p.

44. Comstock JP, Sherpa SR, Ferguson R, Bailey S, Beem-Miller JP, Lin F, et al. Carbonate determination in soils by mid-IR spectroscopy with regional and continental scale models. Minasny B, editor. PLOS ONE. 2019 Feb 21;14(2):e0210235.

45. Harner PL, Gilmore MS. Visible–near infrared spectra of hydrous carbonates, with implications for the detection of carbonates in hyperspectral data of Mars. Icarus. 2015 Apr;250:204–14.

46. Stenberg B, Viscarra Rossel RA, Mouazen AM, Wetterlind J. Visible and Near Infrared Spectroscopy in Soil Science. In: Advances in Agronomy [Internet]. Elsevier; 2010 [cited 2022 Dec 19]. p. 163–215. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0065211310070057

47. Hunt GR. SPECTRAL SIGNATURES OF PARTICULATE MINERALS IN THE VISIBLE AND NEAR INFRARED. GEOPHYSICS. 1977 Apr;42(3):501–13.

48. Ben Dor E, Ong C, Lau IC. Reflectance measurements of soils in the laboratory: Standards and protocols. Geoderma. 2015 May;245–246:112–24.

49. Pudełko A, Chodak M. Estimation of total nitrogen and organic carbon contents in mine soils with NIR reflectance spectroscopy and various chemometric methods. Geoderma. 2020 Jun;368:114306.

50. Shi Z, Ji W, Viscarra Rossel RA, Chen S, Zhou Y. Prediction of soil organic matter using a spatially constrained local partial least squares regression and the Chinese vis-NIR spectral

library: vis-NIR predictions of soil carbon with scL-PLSR. Eur J Soil Sci. 2015 Jul;66(4):679–87.

51. Kühnel A, Bogner C. *In-situ* prediction of soil organic carbon by vis-NIR spectroscopy: an efficient use of limited field data: In-situ prediction of SOC. Eur J Soil Sci. 2017 Sep;68(5):689–702.

52. Gao Y, Cui L, Lei B, Zhai Y, Shi T, Wang J, et al. Estimating Soil Organic Carbon Content with Visible–Near-Infrared (Vis-NIR) Spectroscopy. Appl Spectrosc. 2014 Jul;68(7):712–22.

53. Xu L, Hong Y, Wei Y, Guo L, Shi T, Liu Y, et al. Estimation of Organic Carbon in Anthropogenic Soil by VIS-NIR Spectroscopy: Effect of Variable Selection. Remote Sens. 2020 Oct 16;12(20):3394.

54. Wiesmeier M, Lungu M, Cerbari V, Boincean B, Hübner R, Kögel-Knabner I. Rebuilding soil carbon in degraded steppe soils of Eastern Europe: The importance of windbreaks and improved cropland management. Land Degrad Dev. 2018 Apr;29(4):875–83.

55. Lal R. Restoring Soil Quality to Mitigate Soil Degradation. Sustainability. 2015 May 13;7(5):5875–95.

56. Diacono M, Montemurro F. Long-term effects of organic amendments on soil fertility. A review. Agron Sustain Dev. 2010 Apr;30(2):401–22.

57. Rajan K, Natarajan A, Kumar KSA, Badrinath MS, Gowda RC. Soil organic carbon – the most reliable indicator for monitoring land degradation by soil erosion. Curr Sci. 2010;99(6):823–7.

58. Manlay RJ, Feller C, Swift MJ. Historical evolution of soil organic matter concepts and their relationships with the fertility and sustainability of cropping systems. Agric Ecosyst Environ. 2007 Mar;119(3–4):217–33.

59. Krupenikov IA, Boincean BP, Dent D. Humus – Guardian of Fertility and Global Carbon Sink. In: The Black Earth [Internet]. Dordrecht: Springer Netherlands; 2011 [cited 2022 Dec 19]. p. 39–50. Available from: http://link.springer.com/10.1007/978-94-007-0159-5_7

60. Lal R. Soil Carbon Sequestration Impacts on Global Climate Change and Food Security. Science. 2004 Jun 11;304(5677):1623–7.

61. Houghton RA. Balancing the Global Carbon Budget. Annu Rev Earth Planet Sci. 2007 May 1;35(1):313–47.

62. Chen S, Xu D, Li S, Ji W, Yang M, Zhou Y, et al. Monitoring soil organic carbon in alpine soils using in situ vis-NIR spectroscopy and a multilayer perceptron. Land Degrad Dev. 2020 May 15;31(8):1026–38.

63. Ramírez PB, Calderón FJ, Haddix M, Lugato E, Cotrufo MF. Using Diffuse Reflectance Spectroscopy as a High Throughput Method for Quantifying Soil C and N and Their Distribution in Particulate and Mineral-Associated Organic Matter Fractions. Front Environ Sci. 2021 May 17;9:634472.

64. Seema, Ghosh AK, Das BS, Reddy N. Application of VIS-NIR spectroscopy for estimation of soil organic carbon using different spectral preprocessing techniques and multivariate methods in the middle Indo-Gangetic plains of India. Geoderma Reg. 2020 Dec;23:e00349.

65. Li H, Jia S, Le Z. Prediction of Soil Organic Carbon in a New Target Area by Near-Infrared Spectroscopy: Comparison of the Effects of Spiking in Different Scale Soil Spectral Libraries. Sensors. 2020 Aug 5;20(16):4357.

66. Amin I, Fikrat F, Mammadov E, Babayev M. Soil Organic Carbon Prediction by Vis-NIR Spectroscopy: Case Study the Kur-Aras Plain, Azerbaijan. Commun Soil Sci Plant Anal. 2020 Mar 25;51(6):726–34.

67. Gruszczyński S. Prediction of soil properties with machine learning models based on the spectral response of soil samples in the near infrared range. Soil Sci Annu. 2019 Dec 1;70(4):298–313.

68. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, Massachusetts: The MIT Press; 2016. 775 p. (Adaptive computation and machine learning).

69. Chollet F. Deep learning with Python. Second edition. Shelter Island: Manning Publications; 2021. 478 p.

70. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, Massachusetts: The MIT Press; 2016. 775 p. (Adaptive computation and machine learning).

71. Zhou, Chellappa. Computation of optical flow using a neural network. In: IEEE International Conference on Neural Networks [Internet]. San Diego, CA, USA: IEEE; 1988 [cited 2023 Apr 3]. p. 71–8 vol.2. Available from: http://ieeexplore.ieee.org/document/23914/

72. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. 2014 [cited 2023 Apr 3]; Available from: https://arxiv.org/abs/1412.6980

73. Fukushima K. Cognitron: A self-organizing multilayered neural network. Biol Cybern. 1975;20(3–4):121–36.

74. Rietveld HM. A profile refinement method for nuclear and magnetic structures. J Appl Crystallogr. 1969 Jun 2;2(2):65–71.

75. Doebelin N, Kleeberg R. *Profex* : a graphical user interface for the Rietveld refinement program *BGMN*. J Appl Crystallogr. 2015 Oct 1;48(5):1573–80.

76. Bergmann, J., Friedel, P., Kleeberg, R. BGMN - a new fundamental parameters based Rietveld program for laboratory X-ray sources, it's use in quantitative analysis and structure investigations. Comm Powder Diffr IUCr. 1998;20:5–8.

77. Barthès BG, Kouakoua E, Moulin P, Hmaidi K, Gallali T, Clairotte M, et al. Studying the Physical Protection of Soil Carbon with Quantitative Infrared Spectroscopy. J Infrared Spectrosc. 2016 Jun;24(3):199–214.

78. Viscarra Rossel RA, Hicks WS. Soil organic carbon and its fractions estimated by visible-near infrared transfer functions: Vis-NIR estimates of organic carbon and its fractions. Eur J Soil Sci. 2015 May;66(3):438–50.

79. Morellos A, Pantazi XE, Moshou D, Alexandridis T, Whetton R, Tziotzios G, et al. Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. Biosyst Eng. 2016 Dec;152:104–16.

80. Cortes C, Vapnik V. [No title found]. Mach Learn. 1995;20(3):273–97.

81. Morellos A, Pantazi XE, Moshou D, Alexandridis T, Whetton R, Tziotzios G, et al. Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. Biosyst Eng. 2016 Dec;152:104–16.

82. Zamanian K, Zhou J, Kuzyakov Y. Soil carbonates: The unaccounted, irrecoverable carbon source. Geoderma. 2021 Feb;384:114817.

83. Barthès BG, Kouakoua E, Moulin P, Hmaidi K, Gallali T, Clairotte M, et al. Studying the Physical Protection of Soil Carbon with Quantitative Infrared Spectroscopy. J Infrared Spectrosc. 2016 Jun;24(3):199–214.

84. Wang C, Pan X. Improving the Prediction of Soil Organic Matter Using Visible and near Infrared Spectroscopy of Moist Samples. J Infrared Spectrosc. 2016 Jun;24(3):231–41.

85. Aichi H, Fouad Y, Lili Chabaane Z, Sanaa M, Walter C. Prediction accuracy of local and regional soil total carbon models, calibrated based on visible-near infrared spectra, in the Djerid arid region. J Infrared Spectrosc. 2018 Oct;26(5):322–34.

86. Fontán JM, López-Bellido L, García-Olmo J, López-Bellido RJ. Soil Carbon Determination in a Mediterranean Vertisol by Visible and near Infrared Reflectance Spectroscopy. J Infrared Spectrosc. 2011 Aug;19(4):253–63.

87. Morón A, Cozzolino D. Application of near Infrared Reflectance Spectroscopy for the Analysis of Organic C, Total N and pH in Soils of Uruguay. J Infrared Spectrosc. 2002 Jun;10(3):215–21.

88. Reeves JB, Van Kessel JAS. Investigations into near Infrared Analysis as an Alternative to Traditional Procedures in Manure Nitrogen and Carbon Mineralisation Studies. J Infrared Spectrosc. 1999 Jun;7(3):195–212.

89. Reeves JB, McCarty GW, Meisinger JJ. Near Infrared Reflectance Spectroscopy for the Analysis of Agricultural Soils. J Infrared Spectrosc. 1999 Jun;7(3):179–93.

90. Peng Y, Knadel M, Gislum R, Deng F, Norgaard T, de Jonge LW, et al. Predicting Soil Organic Carbon at Field Scale Using a National Soil Spectral Library. J Infrared Spectrosc. 2013 Jun;21(3):213–22.