RDNet: Deep Learning Model for Predicting pH_{H_2O} and pH_{KCl} from Soil Vis-NIR Spectra

1st Vung Pham
Department of Computer Science
Sam Houston State University
Huntsville, 77340, Texas, USA
vung.pham@shsu.edu

2nd David C. Weindorf
Department of Earth and Atmospheric Sciences
Central Michigan University
Mount Pleasant, 48859, Michigan, USA
weind1dc@cmich.edu

3rd Tommy Dang
Department of Computer Science
Texas Tech University
Lubbock, 79409, Texas, USA
tommy.dang@ttu.edu

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9671527

***Abstract*—Soil properties are vital to profiling and utilizing soil resources. Conventional approaches to measurements of soil properties often involve costly, environmental-unfriendly, and time-consuming laboratory procedures. Conversely, machine learning (ML) and deep learning (DL) are gaining traction in giving rapid, non-destructive, and cost-saving alternatives to predictions of soil properties. These ML/DL models are convenient and fast because they utilize spectral data, such as visible and near-infrared (Vis-NIR) spectra, that can be easily collected using proximal sensors for their training and prediction purposes. However, existing ML/DL approaches to this problem pose several limitations, such as having small sample sizes, needing to divide the sample data into local areas to increase accuracy, and having relatively low accuracy. Therefore, this work experiments various ML/DL methods that leverage Vis-NIR spectra collected from a rather large number of soil samples distributed all over the world to predict pH_{H_2O} and pH_{KCl}. We then propose a DL method, called RDNet, that outperforms the other existing approaches. We also utilize visualizations to verify if the proposed model learns legitimate information from the training data.**

***Index Terms*—machine learning, deep learning, soil Vis-NIR spectra, soil property predictions**

I. Introduction

Soil health analysis and soil utilization require the measurements of soil properties. For instance, the measured soil properties enable soil fertility analysis and environmental analysis (e.g., spills of chemicals in soil and heavy metal concentrations). Measuring these complex soil properties requires complicated, destructive (due to the use of chemicals), and time-consuming laboratory procedures [1]. Conversely, recent developments in the field of visible and near-infrared (Vis-NIR) spectrometer-related technologies and contemporary machine learning advancements offer rapid, cost-effective, and environmentally-friendly alternatives to analyzing soil properties [2].

Current work in this domain often involves small samples (from 40 to 500) localized in some particular areas. Even with small numbers of data samples, the data in this area often involve many data features (e.g., 216 to 1024 wavebands from Vis-NIR spectrum in the range of 350 nm to 2,500 nm). Due to these reasons, the conventional machine learning (ML) approaches to these problems have gained more popularity than their deep learning (DL) counterparts.

There are many soil properties (e.g., pH_{H_2O}, pH_{KCl}, clay content), and the same are true for types of spectral data that scientists often collect from soil samples (e.g., Vis-NIR or X-ray fluorescence spectra). However, without losing generality, this work focuses on leveraging Vis-NIR spectra to predict pH_{H_2O} and pH_{KCl}. There are two main reasons for this decision. First, the measurement procedures for pH_{H_2O} and pH_{KCl} are relatively easy and accurate compared to the other properties. Similarly, the collection of Vis-NIR spectra is relatively cheaper compared to other techniques. Therefore, there are more (more extensive) sample data available. Second, though they are relatively easy to measure, pH_{H_2O} and pH_{KCl} are still worth predicting in special cases. Examples of such instances include when there is no information for these data in the historical literature or when there is a need for measuring these properties from large numbers of measurements using conventional laboratory procedures [3].

This project explores different ML/DL approaches to predicting soil properties using larger data samples globally distributed worldwide. We then propose a novel DL model, called RDNet [4], that achieves results that are superior to existing methods in the same field while utilizing Vis-NIR spectra to predict pH_{H_2O} and pH_{KCl}. We also use visualizations to explain if the learned ML/DL models extract helpful information or merely memorize trivial data from the underlying training samples. The following section discusses the most related ML/DL approaches that utilize spectral data of soil samples to predict their characteristics.

II. Related work

A. Spectrum processing

Generating errors is inevitable while acquiring spectral data using proximal sensors (e.g., Vis-NIR spectroscopy). Examples of causes of such errors are background noise, scattering, baseline drift, electrical noise [5]. Soil structural properties also influence the soil spectra, such as light scattering effects. The raw reflectance spectra of soils could lead to unreasonable results. Thus, scientists often need to transform the acquired spectra to have more meaningful data [6]. Different spectral pre-processing methods resulted in different training/prediction results. However, Savitzky-Golay (SG) is a dominant transformation technique in this domain (e.g., [2],

[7], [8]). SG transformation deletes high-frequency, random noise, enhances the signal-to-noise ratio, and reforms the reflectance spectra [9]. Therefore, our work also applies SG as a data pre-processing technique before training ML/DL models for prediction purposes.

B. Machine learning approaches

There are several ML methods used in the literature for this purpose; however, many of the successful ones (e.g., [2], [5], [7], [8], [10]–[12]) use partial least squared regression (PLSR). The other commonly used ML methods include random forest (RF) (e.g., [13]–[15]), support vector regression (SVR) (e.g., [11], [12]) and multiple linear regression (MLR) (e.g., [5], [11]). PLSR and RF are the two most common for soil property prediction using spectral data. The reason is that they work well with a large number of features while having limited training data. PLSR is particularly useful when predicting a set of dependent variables from a large group of variables [7]. PLSR leverages successive latent variables on both the predictor (X) and the response (Y) variables. In other words, PLSR is suitable for managing multivariate data with high co-linearity in the independent variables, particularly when the sample size is small [8]. Conversely, the RF approach accounts for the nonlinearity of the soil spectral responses; hence, in many cases, it provides higher prediction accuracy [16]. Therefore, this work experiments with these ML approaches and compares the results to other DL techniques.

C. Deep learning approaches

Deep learning has gained successes in various fields from behavioral malware analysis [17], road damage detection and classification [18], graph adversarial attacks and defenses [19], and even solar flare predictions [20]. However, using DL techniques for predicting soil characteristics from spectral data is not as common as its ML counterparts. Still, many initial works explore the predictive powers of DL approaches in this specific domain. Thus, it is still out of the scope of this paper to review every related research in this area. Instead, this work covers several of the most recent studies that use DL to predict soil properties using spectra acquired from corresponding soil samples.

Specifically, Mousavi et al. [8] used a neural network for predicting Atterberg limits (the range of water content of soils in the plastic phase) using Vis-NIR data scanned from 60 forest soil samples. In the same vein, Khosravi et al., [21] experimented with neural networks and other calibration methods (e.g., PLSR, SVR) for predicting lead (Pb) and Zinc (Zn) concentrations from 120 solid samples collected from a waste dump. Similarly, Xu et al., [22] compared the performance of the backpropagation neural network (BPNN) with the other methods (e.g., PLSR, SVR) while predicting soil organic matter (SOM), total nitrogen (TN), total phosphorus (TP), and total potassium (TK) from 148 soil cores.

Most of the previously mentioned studies compare simple neural networks such as multilayer perceptron neural networks (MLP) with other conventional ML approaches (e.g., PLSR, SVR, RF). Contrariwise, Xu et al. [23] focus on comparing different DL approaches to measuring SOM from Vis-NIR reflectance data scanned from 248 red soil samples. Notably, besides MLP, they also explored the relatively more advanced DL architectures such as Convolutional Neural Network (CNN) [24] and DenseNet [25]. These recent works also provide a good overview of previous work related to this area. Interested readers can refer to these papers for further details.

Existing DL techniques in the literature for this area have limitations. Such limitations include using small samples for training or DL architectures used are too simple or too complicated considering the small number of training samples and the large numbers of input features. Therefore, this work experiments with a larger dataset (at a global scale). This work also proposes an appropriate DL model with sufficient training weights to extract the best out of extensive input features.

III. Experiments and Results

Figure 1 depicts the main ideas of this approach. Conventionally, soil scientists need to undergo slow, complex, and environmental-unfriendly laboratory methods to acquire soil properties from soil samples (paths with thick arrows). Instead, we can train ML/DL models from existing Vis-NIR and soil properties (paths with dashed arrows) to predict soil properties of unseen soil samples. The trained ML/DL model then offers a rapid, cost-effective, and non-destructive alternative to measuring soil properties using spectral data acquired by Vis-NIR devices (paths with thin arrows).

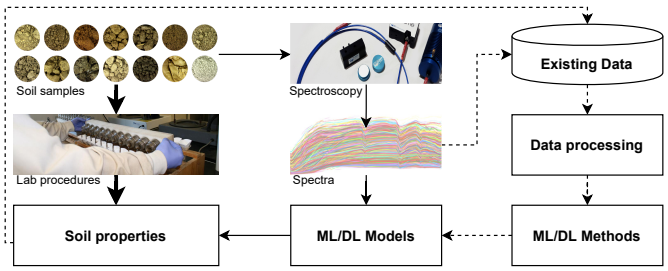

Fig. 1. Conventional laboratory procedures for measuring soil properties (thick arrows) versus using existing Vis-NIR spectral data and corresponding soil properties to train machine learning (ML) or deep learning (DL) model for predicting soil properties (dashed arrows). The trained ML/DL model can then be used to predict soil properties using Vis-NIR spectra acquired from soil samples (thin arrows).

This approach starts with the data gathering and splitting the gathered dataset into the train, validation, and test sets. Next, it discusses the evaluation methods commonly used in evaluating the performance of models for soil property predictions. After the data and evaluation methods, this work details the experiments with the common ML approaches (i.e., PLSR and RF) and DL approaches (i.e., CNN and DenseNet). Then, this work proposes a DL architecture for this problem. Finally, this work utilizes visualizations to explore whether the proposed models can extract legitimate information or merely memorize trivial characteristics from the underlying spectral data.

A. Data and data processing

Most current work in this domain involves using a small number of samples while having a large number of data features, as discussed in Section II. Conversely, DL approaches often have large numbers of weights to be trained. Thus, they need a large number of training samples for training purposes. Therefore, this work utilizes ICRAF-ISRIC soil Vis-NIR spectra [26] acquired from 4,437 soil samples distributed worldwide while experimenting with different ML/DL models for soil property predictions from spectral data. These 4,437 samples were collected from 785 geographical locations in various countries (58) in Africa, the Americas, Asia, and Europe. Specifically, there are 216 consecutive wavebands (10 nm per waveband) for Vis-NIR wavelengths from 350 nm to 2,500 nm.

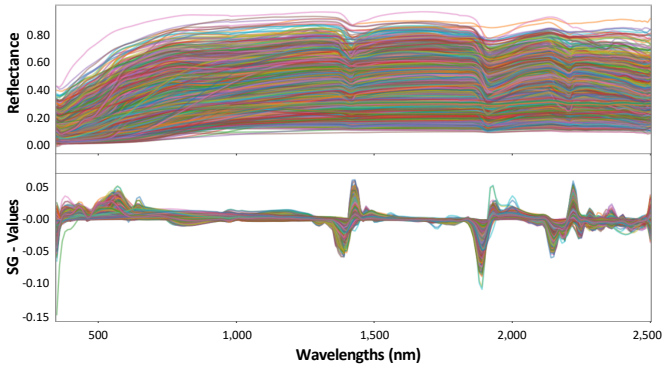

Fig. 2. Visible and near-infrared (Vis-NIR) reflectance and the corresponding Savitzky-Golay (SG) values for the ICRAF-ISRIC world soil spectral library.

Soil scientists often use SG to transform the spectral data before feeding them to further analysis models. In this case, this work also leverages this common practice. Precisely, we apply SG transformation to the acquired Vis-NIR spectra with the following configurations: the first derivative, window length equals 11, and polynomial order as 5. Figure 2 depicts the resulting data after applying such transformation on the ICRAF-ISRIC Vis-NIR spectra. Observably, SG transformed sequences capture the changes under reflectance values between every consecutive waveband. This characteristic of GS transformation makes the SG-transformed sequences more efficient while training ML/DL models. Specifically, concerning PLSR, using raw reflectance data gives similar results to using the SG data. Precisely, RPDs for both pH_{H_2O} and pH_{KCl} of 1.68 when using SG data vs. 1.71 for pH_{H_2O} and 1.70 for pH_{KCl} when using raw Vis-NIR reflectance, respectively. However, for all other experimented methods (RF and DL), using SG-transformed data outperforms the raw reflectance data results. Therefore, in the following sections, this work will report the results on SG-transformed data only for the sake of brevity.

Besides Vis-NIR spectra, there are 23 soil properties (measured by the conventional laboratory procedures) in this ICRAF-ISRIC dataset. The examples of these properties include pH_{H_2O}, pH_{KCl}, $CaCO_3$. Though there are 23 soil properties available, this work focuses on utilizing the acquired Vis-NIR spectral data to train ML/DL models and predict pH_{H_2O} and pH_{H_2O}. A recent work [3] inspired the selection of these two pH values in this work. Precisely, Wang et al. [3] utilized conventional laboratory procedures to acquire pH_{H_2O} from soil samples and trained ML models to predict pH_{KCl} from the measured pH_{H_2O}. They claimed that this approach could reduce time and expenses when many measurements are taken using conventional laboratory procedures. We hypothesize that using Vis-NIR to predict both pH_{H_2O} and pH_{KCl} would then save even more time and cost. Moreover, in the ICRAF-ISRIC dataset used in this project, these two soil properties have minimal missing data numbers.

At the data cleaning stage, we removed 599 soil samples due to missing data or duplication. The remained 3,838 samples are further divided into a training set (3,070 samples), a validation set (384 samples), and a testing set (384 samples). Notably, in most of the reviewed works, the data are often divided into training and validation sets only due to the limited number of samples. Authors often use the validation set to find the best results and report the results based on this validation set. Arguably, this approach is too optimistic, and the reported results do not generalize well. Therefore, this work uses the validation set to search for the optimal model trained using the training dataset. However, the results reported in this paper are evaluated on the independent dataset (the test dataset) instead of the optimistic results evaluated on the validation set.

B. Model selection methods

Mean squared error (MSE), R^2 correlation coefficient, and residual prediction deviation (RPD) are the performance evaluation metrics used in this work to select the best ML/DL model from the experimented ones. It is prominent practice for regression problems to use MSE as a loss function. Conversely, R^2 and RPD are important scores in this specific domain; thus, they are also reported by scientists in this area. Specifically, R^2 indicates the relationship between actual values and the corresponding predicted ones. Moreover, different research has different sample sizes and difficulty levels (i.e., deviations of the values to be predicted). Thus, it is not easy to compare the results from different models trained on different data samples. In this case, RPD enables scientists to compare models of models that are trained and validated on various datasets with different characteristics. Precisely, given an RPD score, there is a corresponding predictive capability, as shown in Table I. Scientists can use these predictive capabilities to compare the performances of models trained and evaluated on different data samples.

TABLE I
RESIDUAL PREDICTIVE DEVIATION (RPD) AND CORRESPONDING PREDICTIVE CAPABILITIES.

Capability	Bad	Rough	Moderate	Good	Excellent
RPD	< 1.5	1.5 – 2.0	2.0 – 2.5	2.5 – 3.0	> 3.0

C. Common machine learning methods

PLSR and RF are the prominent ML methods in this specific domain because these two techniques can leverage the characteristics (small sample sizes and large input features) of the training samples in this domain. Consequently, we experimented with these two methods in this project. Regarding PLSR, one important hyperparameter to tune is the number of PLSR components. Figure 3 depicts the process of searching for the best number of PLSR components on validation data. Consequently, Figure 4 reports the PLSR results with this best number of PLSR components (69) evaluated on the test set. In this reporting chart, X-axis represents the actual values while Y-axis designates the predicted values. The solid line reflects the perfect predictions (where all actual values are the same as predicted ones). Concomitantly, the regression between these two quantities is represented as the dashed line. Observably, PLSR does not provide good evaluation results on the independent test set. These low evaluation scores suggest that though PLSR is dominant in this specific domain, it is not adequate when there is a massive number of heterogeneous samples distributed at the global scale as in this scenario.

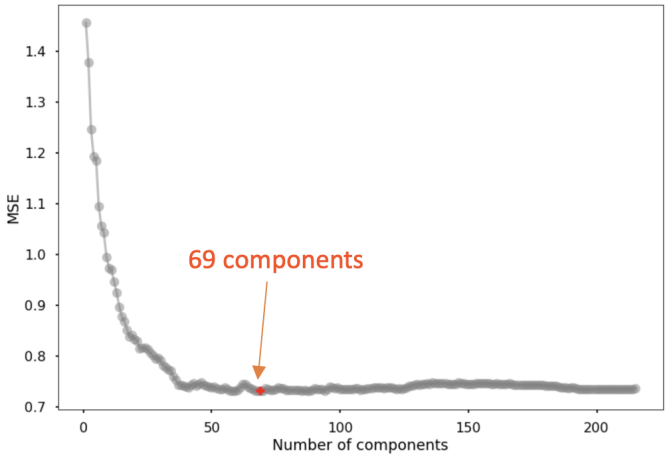

Fig. 3. Experiments with different numbers of partial least squares regression (PLSR) components. PLSR hyperparameter with lowest mean squared error (MSE) is at 69 components.

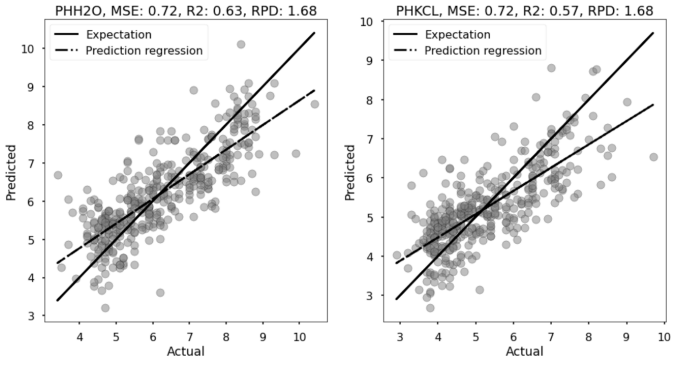

Fig. 4. Partial least squares regression results on the test set.

RF method has several hyperparameters to tune. However, two critical hyperparameters include the number of trees (*n_estimators*) and the trees' maximum depth (*max_depth*). Therefore, we create a function to find the set of hyperparameters for the model trained on the training data that achieves the best performance on the validation set. Specifically, this work searched on $max\_depth = [10, 20, 40, 80, 150, 200, 400, 500, 600, 700, 1000, 2000, 3000]$ and $n\_estimators = [50, 100, 200, 300, 1000]$. Observably, these hyperparameters start with small values and skips, then end with exponentially large values. This search value change pattern is the common practice in hyperparameter search. The combinations of these hyperparameters add up to 65 configurations.

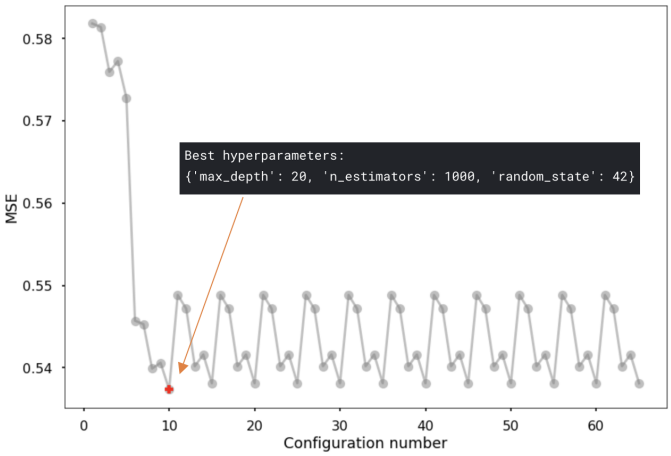

Fig. 5. The search for random forest hyperparameters that provide lowest mean squared error (MSE). The best model has 1,000 estimators (*n_estimators*) and each has a maximum depth (*max_depth*) of 20.

Figure 5 shows the MSE results on these configurations. The best configuration is the 10^{th} one with $max\_depth = 20$ and $n\_estimators = 1000$. Notably, for reproducibility, for all experiments, this work sets the random state to a constant (42 is the default value). Consequently, Figure 6 reports the results of fitting an RF model with these hyperparameters on the training data and were evaluated on the independent test dataset. Observably, RF model provides better results for all MSE, R^2, and RPD for both pH_{H_2O} and pH_{KCl}. These improved results compared to PLSR's indicate that RF performs better with a larger amount/more diverse data because it can extract information from the nonlinear relationship between the data features and the outputs.

D. Common deep learning methods

RF approaches often work well in modeling the nonlinear relationship between input and output. Therefore, better RF results (vs. PLSR's) suggests that the relationship between Vis-NIR input and pH_{H_2O} and pH_{KCl} outputs is somewhat nonlinear. Concomitantly, DL approaches also have great abilities to model nonlinearity in input/output relationships. Therefore, it should be beneficial to experiment with DL approaches for this specific scenario. Specifically, Section II indicates

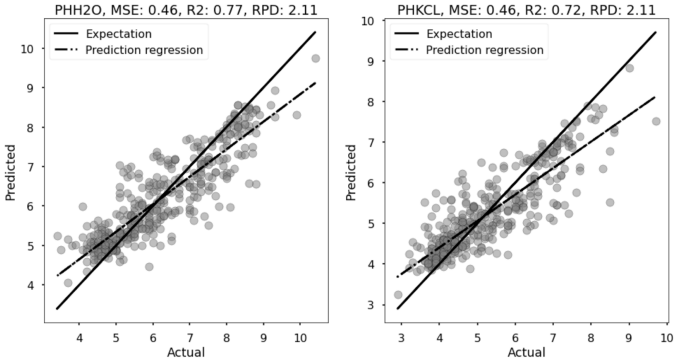

Fig. 6. Random forest results on the test set.

that two dominant DL methods used in this specific domain (predicting soil properties from Vis-NIR spectra) include MLP and CNN. Therefore, we also experimented with these two methods. It is nearly impossible to exhaustively explore all the potential architectures and hyperparameters to get the best DL model for this dataset. Therefore, in the following sections, the experimented DL models are based on observations from the training/validation information and repeated trials. Furthermore, this work uses early stopping to search for the best models using validation data. However, to be consistent and assure generalization, it only reports the evaluation results on the test data.

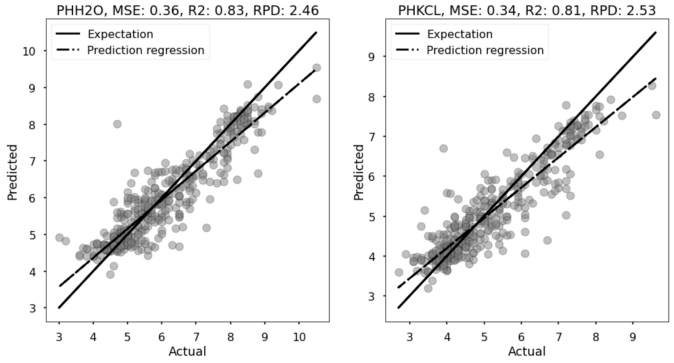

Fig. 7. Multilayer perceptron results on the test set.

The best found MLP neural network architectures or hyperparameters include five dense (i.e., fully conncted) layers with 512, 256, 128, 64, and 32 hidden units correspondingly. There is a dropout layer (rate equals 0.1) after each of these five layers to tackle overfitting. Figure 7 reports the MLP model performance assessed on the independent test dataset. Observably, this MLP model evaluated on the test dataset provides better performances for all performance evaluation metrics (MSE, R^2, and RPD) for both measurements (pH_{H_2O} and pH_{KCl}), compared to the experimented ML models (PLSR and RF). These improved results indicate great potentials for using neural networks and Vis-NIR spectra to predict pH_{H_2O} and pH_{KCl} at a global scale like in this specific scenario.

In their recent work [23], Xu et al. utilized DenseNet and Vis-NIR spectral data to predict SOM from soil samples. We assumed that the same architecture described in this work has the ability to extract useful information from input Vis-NIR spectra. Therefore, this work also implemented the same DenseNet architecture described in their original paper. We refer interested readers to this paper [23] for further information about the architecture of the experimented DenseNet. Figure 8 reports the DenseNet performance assessed on the independent test data. Observably, DenseNet provides MSE, R^2, and RPD scores that are somewhat similar to RF's performance. In other words, it is not as good as the MLP model for this scenario.

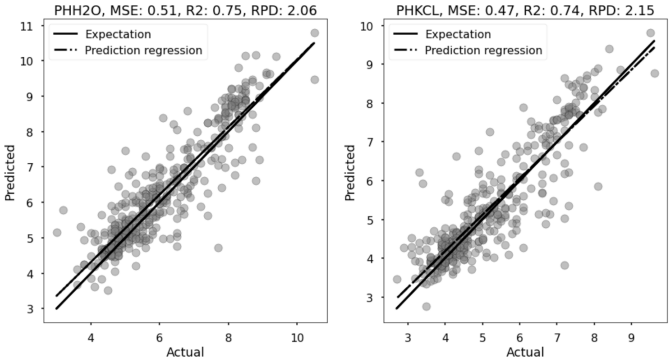

Fig. 8. DenseNet results on the test set.

DenseNet (a specific architecture of CNN) does not perform well in this particular case. However, the Vis-NIR spectra in this dataset are consecutive wavebands. Thus, they are sequential data. Concomitantly, CNN can extract salient features from data sequences well. Therefore, we continued to explore another CNN-based model for this purpose. Specifically, this work experimented with using VGG [27] blocks (each block has two convolutional layers followed by a pooling layer) to build a neural network for soil property prediction using Vis-NIR spectra. The best experimented VGG-based neural network has 4 VGG blocks (the first block has 32 hidden units, and the other three have 64 hidden units each). Also, there is a dense layer (128 hidden units) as a buffer before the final output layer. Like other cases, every VGG block is followed by a dropout layer (with a dropout rate equal to 0.1) to tackle overfitting issues. Figure 9 presents the results of this model assessed on the independent test dataset. Observably, this VGG-based model performs better than DenseNet. However, its performance is similar to the experimented MLP model.

Notably, we also applied recurrent neural network (RNN) and long-short term memory neural network (LSTM) to tackle this problem (soil property predictions using Vis-NIR spectra). The reason is that these types of neural networks can work well with sequential data. However, the Vis-NIR spectra, in this case, has 216 wavebands. This long sequence makes these two types of neural networks inefficient. Due to space limitations, we do not report the performance of these two experimented models in this paper.

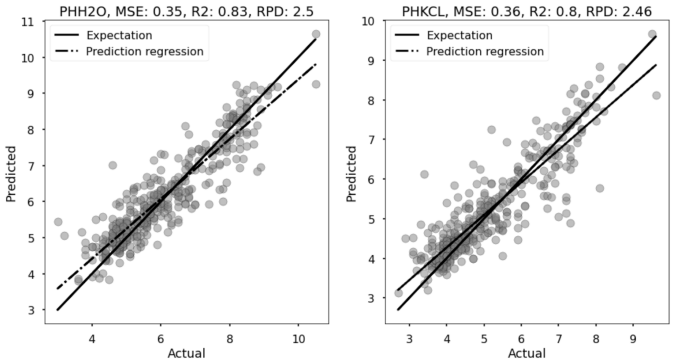

Fig. 9. Visual Geometry Group based neural network results on the test set.

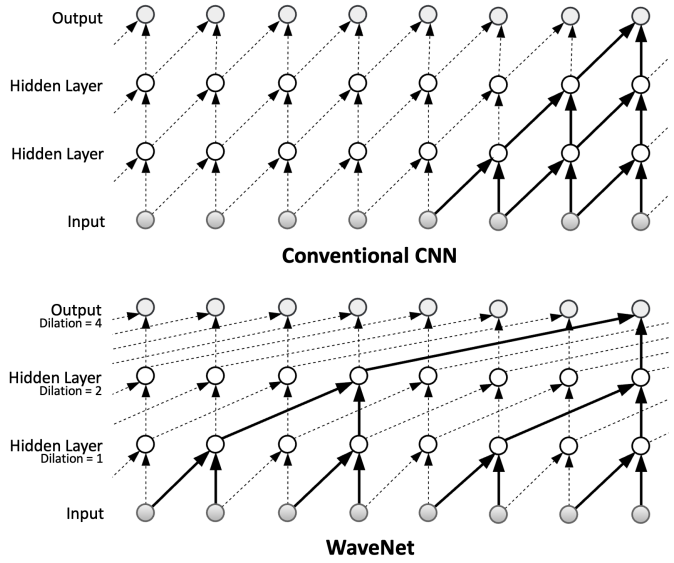

Fig. 10. Conventional convolutional neural network (CNN) vs. WaveNet. WaveNet has the ability to learn useful features from the input sequence with less training weights by leveraging dilations.

E. WaveNet and RDNet

CNN has achieved great performances in the image analysis domain because this domain often has large numbers of images available for training. For instance, ImageNet [28] has 1,281,167 and 50,000 images in its training and validation sets respectively. In this specific case, by using the ICRAF-ISRIC global soil Vis-NIR spectral library, we have relatively more data samples to train our models than the other studies in the same domain. However, this sample size is relatively modest for efficiently training the conventional CNN models. Due to the smaller number of training samples, conventional CNN (with many weights to be trained) does not perform well in this specific case. Concomitantly, WaveNet [29] is another type of CNN architecture that utilizes a dilation to decrease the number of training parameters while still begetting the capacity to learn useful features from a long sequence of data. This reduction in the number of training weights means that it requires fewer data samples to train, which is required in this specific scenario. Specifically, Figure 10 depicts the difference between two experimented CNN architecture types (conventional CNN vs. WaveNet).

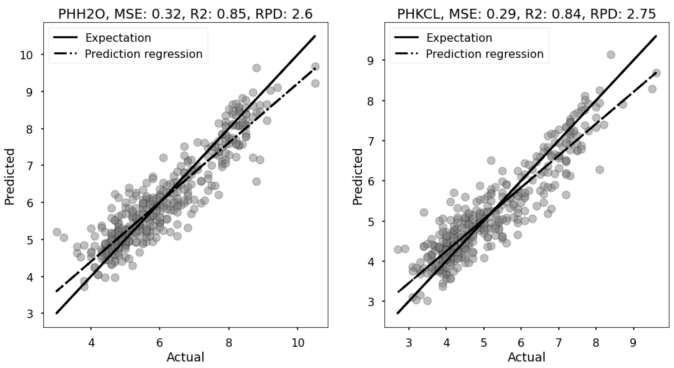

Fig. 11. WaveNet results on the test set.

To the best of our knowledge, this is the first work in the literature that uses WaveNet to predict soil properties from Vis-NIR spectral data. Figure 12 depicts the architecture for the WaveNet-based neural network model experimented for this specific scenario. Specifically, there are six convolutional layers with dilation rates of 1, 2, 4, 8, 16, and 32 (32 filers each). These six layers make a WaveNet block. Notably, a dropout layer (with a dropout rate of 0.2) is added after the WaveNet block to tackle overfitting.

Figure 11 depicts the results for the experimented WaveNet-based model for this scenario evaluated on the independent test set. Observably, the WaveNet results are encouraging and outperform all other experimented models reported so far. However, this model can accommodate only one WaveNet block. In other words, we cannot stack more such WaveNet blocks to create deeper neural networks and gain performance, or adding more such blocks to the model induces the performance degradation issue.

We leverage a skip connection [30] for each WaveNet block that allows stacking more WaveNet blocks to build a deeper neural network and gains predictive performance. Combining the skip connection and WaveNet block creates a residual dilated block (called RD block). Figure 13 depicts the architecture of an RD block. Specifically, this block has the same number of convolutional layers and hyperparameters as the described WaveNet block (the dashed arrow means that there are more convolutional and activation layers). However, there is a skip connection to add the block's input to the block's output before the activation of the last convolutional layer (as suggested in the original ResNet paper). Also, due to this requirement, we split the activation out as a separate layer after every convolutional layer.

Figure 14 depicts the neural network architecture for this scenario utilizing RD block called Residual Dilated Neural Network or RDNet for short. Notably, the architecture is similar to that of the experimented WaveNet architecture. However, there are two main differences. First, the skip connection in the RD Block allows stacking up to ten such blocks together and produce a deeper neural network model. Second, there is a

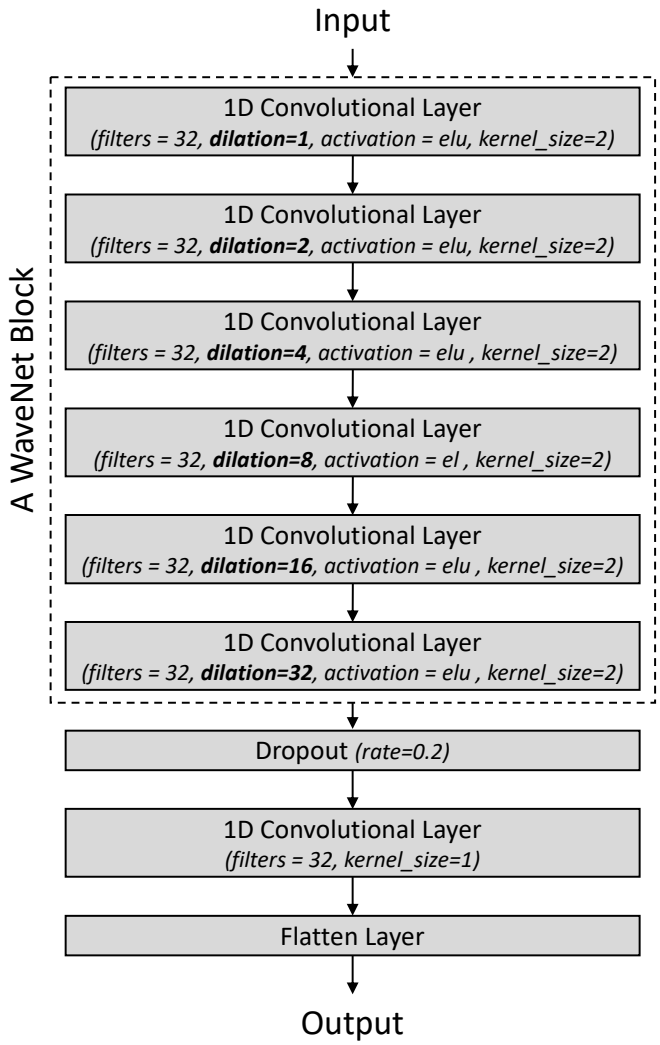

Fig. 12. The experimented WaveNet-based neural network model.

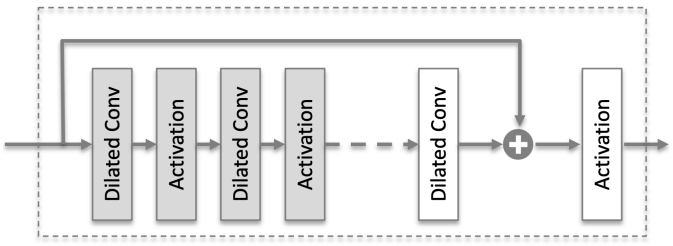

Fig. 13. Architecture of a residual dilated block (RD block). RD block has an architecture similar to that of a WaveNet block, however, there is a skip connection to add the input to the output right before the last activation.

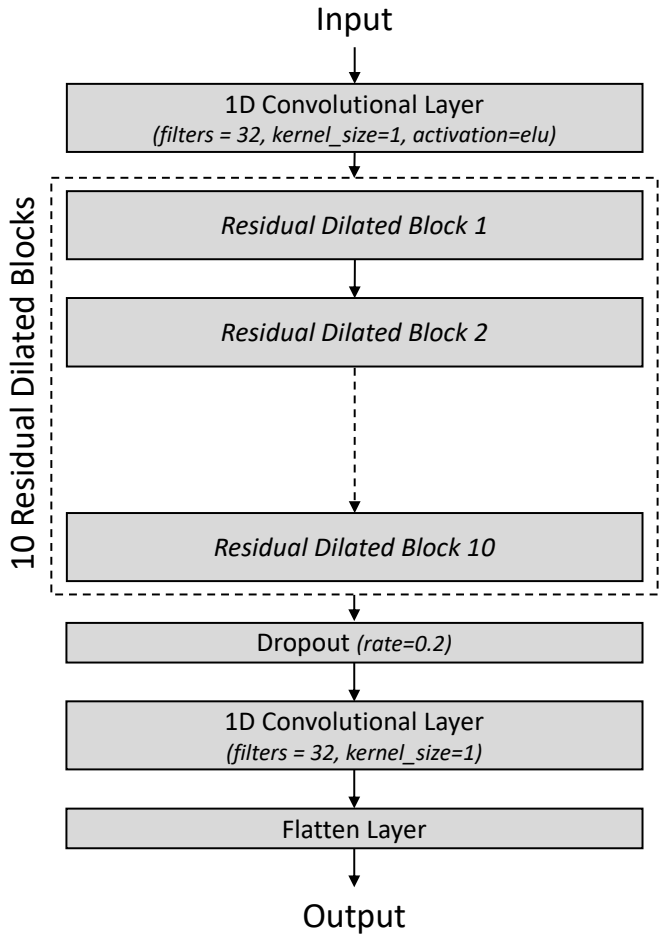

Fig. 14. The experimented residual dilated neural network (RDNet). With the skip connection, we can add up to 10 residual dilated blocks in this model.

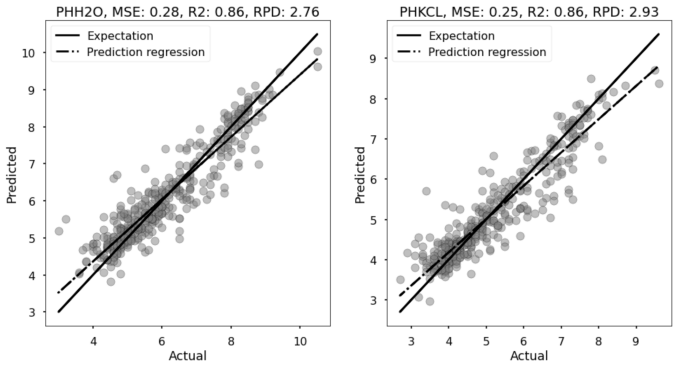

Fig. 15. Residual Dilated Neural Network (RDNet) model performance evaluated on the independent test dataset.

convolutional layer added right after the input to avoid having to broadcast the input (with 32 filters) before feeding to the first RD block.

Figure 15 depicts the experimented RDNet model performance assessed on the independent test data. Observably, the RDNet model's performance is superior to all other experimented models in all evaluation metrics (i.e., MSE, R^2, and RPD). Specifically, with RPD values of 2.76 and 2.93 for pH_{H_2O} and pH_{KCl} respectively, this model has the good predictive capability (as shown in Table I). Notably, the prediction of pH_{KCl} gets very close to the excellent predictive ability (i.e., RPD of 3.0).

F. Visualization and Explainability

Though having superior predictive capabilities, ML models, especially DL models, often have many parameters (i.e., weights). Therefore, they may use this massive number of weights to memorize the inputs and outputs trivially instead of learning useful salient features from the training samples [31], [32]. The RDNet model proposed in this work is also a relatively deep neural network (with 62 hidden CNN layers and one dense layer for the output). Thus, it might face the same issue. That said, we utilize different methods to inspect what wavebands (input features) each experimented model pays attention to when it makes its predictions. Knowing what input features have significant impacts on model outputs helps validate if the model can learn legitimate salient features from the training samples.

It is common to use PLSR's coefficient for each input feature as its significance to the final output prediction. Concomitantly, there is a built-in feature importance method for the RF model. Specifically, this method utilizes SHAP (SHapley Additive exPlanations) [33] for this purpose. SHAP is a unified method to interpret the model prediction outcomes of machine learning models using game theory and local explanations [20]. Similarly, this work uses a SHAP Kernel Explainer to produce input feature importance for the experimented RDNet model. Out of the experimented DL model, this section only reports RDNet's input feature importance for brevity. The reason is that RDNet is the deepest DL model among the experimented ones. That means it may have a high potential to learn trivial and unexplainable salient features.

Figure 16 depicts that wavebands (x-axis) with more information entropy (the red bands with more SG value differences among samples), and the corresponding PLSR, RF, and RDNet models' feature importance values (y-axis). Notably, PLSR does not focus on any specific sets of wavebands, while RF only focuses exceptionally high at about the wavelength of 2,200 nm. On the other hand, RDNet utilizes all the relevant, high-entropy wavebands (red shaded regions and especially at the wavelengths indicated by the dashed lines) while predicting pH_{H_2O} and pH_{KCl} outputs.

Specifically, RDNet utilizes well the important wavelengths around 700 nm, 1,400 nm, 1,900 nm, 2,200 nm, and 2,400 nm. The 700 nm reflects the red color in the soil, an essential indicator for soil pH values. Furthermore, the other wavelengths are related to the C-H group, O-H stretching, H-O-H blending, and clay lattice Al-OH absorption band [15], [34]. One evidence to support that RDNet utilizes valuable features from the underlying Vis-NIR spectra is that predictive results assessed on the independent test dataset are similar to that on the validation dataset. This similar performance on the two evaluation sets means the predictive capability is generalized well.

IV. Discussion

Table II reports the summary of the evaluation results (MSE, R^2, and RPD) for all the experimented ML/DL methods (PLSR, RF, MLP, DenseNet, VGG, WaveNet, and RDNet) while evaluated on the test set. Notably, our proposed solution (RDNet) has evaluation results that are superior to those from all other experimented ML/DL models. The superior performance is explained by its capability to leverage valuable features from the relevant Vis-NIR wavebands (as depicted in Figure 16).

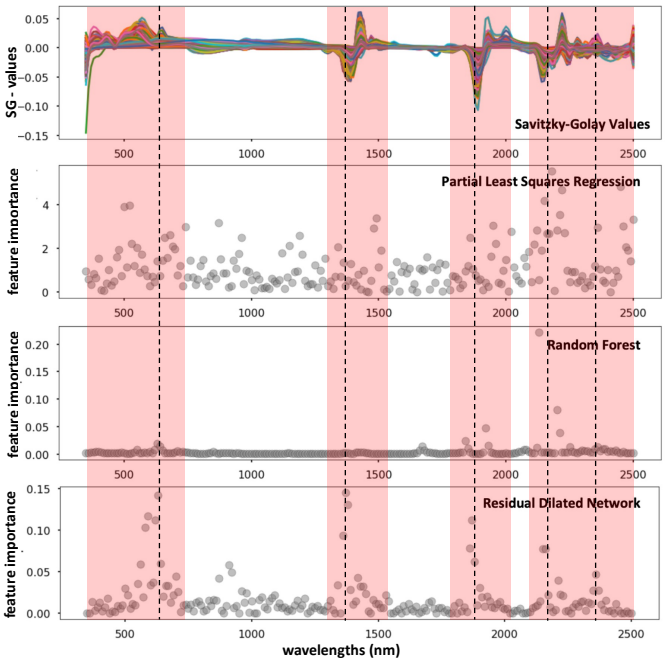

Fig. 16. Savitzky-Golay (SG) values and respective feature importance for the experimented partial least squares regression, random forest, and residual dilated neural network models. The wavebands at the red shaded regions (especially at the dashed lines) are those with information more relevant to the pH_{H_2O} and pH_{KCl} predictions.

TABLE II
Summary of the Evaluation Results on the Test Set for All the Experimented Models.

	pH	PLSR	RF	MLP	DenseNet	VGG	WaveNet	RDNet
MSE	H2O	0.72	0.46	0.36	0.51	0.35	0.32	**0.28**
	HKL	0.72	0.46	0.34	0.47	0.36	0.29	**0.25**
R2	H2O	0.63	0.77	0.83	0.75	0.83	0.85	**0.86**
	HKL	0.57	0.72	0.81	0.74	0.80	0.84	**0.86**
RPD	H2O	1.68	2.11	2.46	2.06	2.50	2.60	**2.76**
	HKL	1.68	2.11	2.53	2.16	2.46	2.75	**2.93**

Observably, Figures 4, 6, 7, 8, 9, 11, and 15 reveal the trend that the experimented models give greater outputs for the extremely low pH values and lesser outputs for the extremely high ones. These tendencies can further be leveraged to optimize the ML/DL models to gain better accuracy in the future. Additionally, tangential adaptations of this proposed approach can help explore and discover models to predict other important soil properties (e.g., organic carbon, soil cation exchange capacity) in the future. Finally, source codes, visualizations, and data of this project are available on its Github page at https://github.com/iDataVisualizationLab/V/tree/master/globalsoilspectral. Also, as an attempt to assure reproducibility, we set the random states for the execution environment (Numpy and Tensorflow) to a default value (42).

V. Conclusions

This work reports the results on experiments with different existing methods to train machine learning and deep learning models using Vis-NIR spectra to predict pH values (pH_{H_2O} and pH_{KCl}) of a set of globally distributed soil samples. From these experiments and the observations from the training data characteristics (large number of features/Vis-NIR wavebands and a relatively small number of training samples), this work proposes a neural network called RDNet that outperforms the other experimented models. RDNet reduces the number of training parameters using convolutional layers with dilations. Moreover, it uses skip connections to allow stacking more layers in the neural network, thus enhancing the predictive power.

Having a deeper neural network implies the potential for the model to memorize trivial features from training data. Therefore, this work utilizes visualization to explore how the proposed model learns from the input spectral data. The visualization indicates that RDNet learns from useful/high-information-entropy Vis-NIR wavebands while making pH_{H_2O} and pH_{KCl} predictions. In the future, we will investigate more models to predict other important soil properties available in this globally distributed soil spectral dataset. Also, the number of training samples in this work is relatively large compared to the other existing works; however, it is still small for deep learning models to be effective. Therefore, we will continue to explore and add more data sources for training efficient models.

Acknowledgments

This work was supported by the National Science Foundation under the I-Corps award number 2017018. The authors gratefully acknowledge the contributions of the reviewers that make this paper better.

References

[1] D. C. Weindorf, S. Chakraborty, J. Moore-Kucera, B. Li, L. Fultz, V. Acosta-Martinez, and C. Li, "Advanced modeling of soil biological properties using visible near infrared diffuse reflectance spectroscopy," *International Journal of Bioresource Science*, vol. 5, no. 1, pp. 01–20, 2018.

[2] Y. Ba, J. Liu, J. Han, and X. Zhang, "Application of vis-nir spectroscopy for determination the content of organic matter in saline-alkali soils," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 229, p. 117863, 2020.

[3] A. Wang, D. Li, B. Huang, Y. Lu, *et al.*, "A brief study on using ph h 2 o to predict ph kcl for acid soils," *Agricultural Sciences*, vol. 10, no. 02, p. 142, 2019.

[4] V. Pham, D. C. Weindorf, and T. Dang, "Soil profile analysis using interactive visualizations, machine learning, and deep learning," *Computers and Electronics in Agriculture*, vol. 191, p. 106539, 2021.

[5] B. Lu, N. Liu, H. Li, K. Yang, C. Hu, X. Wang, Z. Li, Z. Shen, and X. Tang, "Quantitative determination and characteristic wavelength selection of available nitrogen in coco-peat by nir spectroscopy," *Soil and Tillage Research*, vol. 191, pp. 266–274, 2019.

[6] A. D. Vibhute, K. V. Kale, S. C. Mehrotra, R. K. Dhumal, and A. D. Nagne, "Determination of soil physicochemical attributes in farming sites through visible, near-infrared diffuse reflectance spectroscopy and plsr modeling," *Ecological Processes*, vol. 7, no. 1, p. 26, 2018.

[7] J. Liu, J. Han, J. Xie, H. Wang, W. Tong, and Y. Ba, "Assessing heavy metal concentrations in earth-cumulic-orthic-anthrosols soils using vis-nir spectroscopy transform coupled with chemometrics," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 226, p. 117639, 2020.

[8] F. Mousavi, E. Abdi, A. Ghalandarzadeh, H. A. Bahrami, B. Majnounian, and N. Ziadi, "Diffuse reflectance spectroscopy for rapid estimation of soil atterberg limits," *Geoderma*, vol. 361, p. 114083, 2020.

[9] A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures.," *Analytical chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[10] M. Allo, P. Todoroff, M. Jameux, M. Stern, L. Paulin, and A. Albrecht, "Prediction of tropical volcanic soil organic carbon stocks by visible-near-and mid-infrared spectroscopy," *CATENA*, vol. 189, p. 104452, 2020.

[11] Y. Zhang, M. Li, L. Zheng, Q. Qin, and W. S. Lee, "Spectral features extraction for estimation of soil total nitrogen content based on modified ant colony optimization algorithm," *Geoderma*, vol. 333, pp. 23–34, 2019.

[12] H.-D. Li, Q.-S. Xu, and Y.-Z. Liang, "libpls: An integrated library for partial least squares regression and linear discriminant analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 176, pp. 34–43, 2018.

[13] Y. Hong, S. Chen, Y. Chen, M. Linderman, A. M. Mouazen, Y. Liu, L. Guo, L. Yu, Y. Liu, H. Cheng, *et al.*, "Comparing laboratory and airborne hyperspectral data for the estimation and mapping of topsoil organic carbon: Feature selection coupled with random forest," *Soil and Tillage Research*, vol. 199, p. 104589, 2020.

[14] Y. Zhang, W. Ji, D. D. Saurette, T. H. Easher, H. Li, Z. Shi, V. I. Adamchuk, and A. Biswas, "Three-dimensional digital soil mapping of multiple soil properties at a field-scale using regression kriging," *Geoderma*, vol. 366, p. 114253, 2020.

[15] Y. Hong, S. Chen, Y. Liu, Y. Zhang, L. Yu, Y. Chen, Y. Liu, H. Cheng, and Y. Liu, "Combination of fractional order derivative and memory-based learning algorithm to improve the estimation accuracy of soil organic matter by visible and near-infrared spectroscopy," *Catena*, vol. 174, pp. 104–116, 2019.

[16] R. K. Douglas, S. Nawar, M. C. Alamar, A. Mouazen, and F. Coulon, "Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-nir spectroscopy and regression techniques," *Science of the Total Environment*, vol. 616, pp. 147–155, 2018.

[17] A. McDole, M. Gupta, M. Abdelsalam, S. Mittal, and M. Alazab, "Deep learning techniques for behavioral malware analysis in cloud iaas," in *Malware Analysis using Artificial Intelligence and Deep Learning*, pp. 269–285, Springer, 2021.

[18] V. Pham, C. Pham, and T. Dang, "Road damage detection and classification with detectron2 and faster r-cnn," in *2020 IEEE International Conference on Big Data (Big Data)*, pp. 5592–5601, IEEE, 2020.

[19] C. Pham, V. Pham, and T. Dang, "Graph adversarial attacks and defense: An empirical study on citation graph," in *2020 IEEE International Conference on Big Data (Big Data)*, pp. 2553–2562, IEEE, 2020.

[20] C. Pham, V. Pham, and T. Dang, "Solar flare prediction using two-tier ensemble with deep learning and gradient boosting machine," in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 5844–5853, IEEE, 2019.

[21] V. Khosravi, F. D. Ardejani, S. Yousefi, and A. Aryafar, "Monitoring soil lead and zinc contents via combination of spectroscopy with extreme learning machine and other data mining methods," *Geoderma*, vol. 318, pp. 29–41, 2018.

[22] S. Xu, Y. Zhao, M. Wang, and X. Shi, "Comparison of multivariate methods for estimating selected soil properties from intact soil cores of paddy fields by vis–nir spectroscopy," *Geoderma*, vol. 310, pp. 29–43, 2018.

[23] Z. Xu, X. Zhao, X. Guo, and J. Guo, "Deep learning application for predicting soil organic matter content by vis-nir spectroscopy," *Computational Intelligence and Neuroscience*, vol. 2019, 2019.

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[26] D. Garrity and P. Bindraban, "A globally distributed soil spectral library visible near infrared diffuse reflectance spectra," *ICRAF (World Agroforestry Centre)/ISRIC (World Soil Information) Spectral Library: Nairobi, Kenya*, 2004.

[27] Visual Geometry Group, "Visual Geometry Group." http://www.robots.ox.ac.uk/~vgg/. Accessed: 2021–03-11.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.
[29] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
[31] T. Dang, H. Van, H. Nguyen, V. Pham, and R. Hewett, "Deepvix: Explaining long short-term memory network with high dimensional time series data," in *Proceedings of the 11th International Conference on Advances in Information Technology*, pp. 1–10, 2020.
[32] D. D. Le, V. Pham, H. N. Nguyen, and T. Dang, "Visualization and explainable machine learning for efficient manufacturing and system operations," *Smart and Sustainable Manufacturing Systems 3*, 2019.
[33] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.
[34] R. V. Rossel and T. Behrens, "Using data mining to model and interpret soil diffuse reflectance spectra," *Geoderma*, vol. 158, no. 1-2, pp. 46–54, 2010.