Journal of Environmental Biology
p-ISSN: 0254-8704 • e-ISSN: 2394-0379 • CODEN: JEBIDP

JEB®

Original Research Paper
Volume 45 Issue 1 January 2024 pp. 96-105
DOI: http://doi.org/10.22438/jeb/45/1/MRN-5134

J. Environ. Biol., 2024

Journal website : www.jeb.co.in ★ E-mail : editor@jeb.co.in

A machine learning based approach for prediction and interpretation of soil properties from soil spectral data

A. Divya, R. Josphineleela and L. Jaba Sheela*

Department of Computer Science and Engineering, Panimalar Engineering College, Chennai-600 123, India

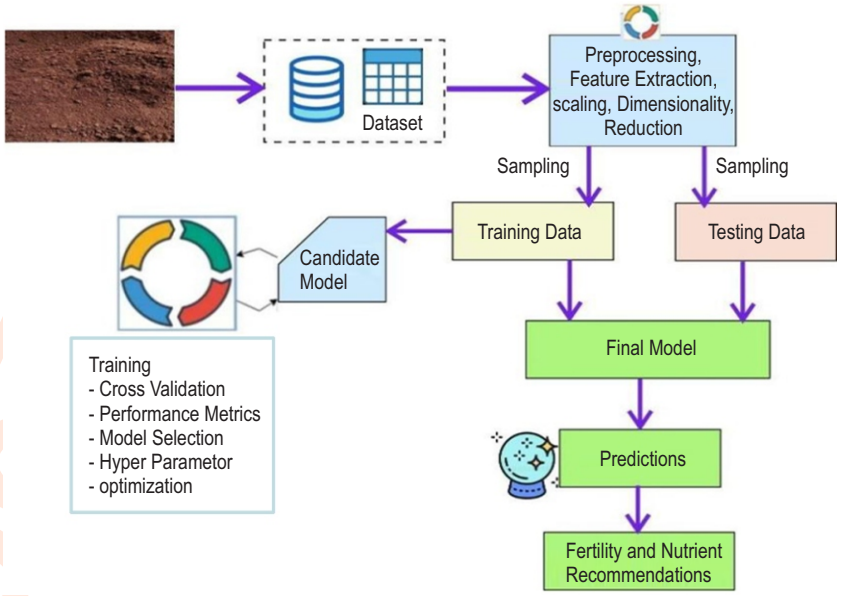
Received: 02 April 2023 **Revised:** 23 October 2023 **Accepted:** 04 November 2023

***Corresponding Author Email :** ljsheela@panimalar.ac.in ***ORCiD:** https://orcid.org/0000-0002-7182-5582

Abstract

Aim: An active agricultural sector depends on good soil quality, essential for sustained food cultivation. However, intensive farming and rising demand can lead to soil deterioration, affecting crop yields. Smart soil prediction driven by machine learning is crucial for precision farming and efficient nutrient distribution.

Methodology: Visible-near infrared Spectroscopy (vis-NIRS) is used to capture the soil's spectral data.Then, the spectral data is preprocessed with Savitzky-Golay Smoothing.The data that has been preprocessed is then used to train the machine learning model.The preprocessed data enhances model performance compared to spectral reflectance data in its unprocessed state.The machine learning model acquires data-based knowledge, identifies patterns, and predicts soil quality parameters. The Random Forest and Gradient Boosted Regression Tree are two algorithms employed in this study.



Results: The spectral reflectance data is used to train, validate, and evaluate the machine learning model.In determining soil properties, both algorithms demonstrated a high degree of prediction accuracy, as demonstrated by the results.Gradient Boosted Regression Tree out performs Random Forest, but is expensive and requires sequential data. Random forest algorithm works well with large datasets, but over-fitting issues arise in some instances.

Interpretation: The findings of the study indicate that machine learning can automate the current soil testing procedure in laboratories, thereby making it more efficient, affordable, and environmentally friendly.

Key words: Gradient Boosted Regression Tree, Machine learning, Random forest, Soil fertility, Soil moisture

How to cite: Divya, A., R. Josphineleela and L. Jaba Sheela: A machine learning based approach for prediction and interpretation of soil properties from soil spectral data. *J. Environ. Biol.*, **45,** 96-105 (2024).

Introduction

The pace of population increase and the requirement for a growing food supply must be proportionately related. So it is essential to act in order to boost agricultural productivity. Crop yield is significantly influenced by soil characteristics and plant interactions. Farmers must determine the requirements for soil fertility in order to produce crops more effectively and economically (Massawe *et al*., 2018; Paul *et al.,* 2020). Farmers may find it beneficial to quantify the nutrient concentration in the soil in order to replenish depleted nutrients and determine the optimal crop for a given soil sample (Motwani *et al.,* 2022). Potassium, nitrogen, phosphorous, calcium, and pH are among the soil nutrients that contribute to plant growth. Since it offers more details about various aspects of the fertility of the soil, the soil's pH is the most significant property (Zhang *et al*., 2017). Tools for monitoring soil quality are essential for addressing agronomic sustainability challenges in a populous nation like India (Paul *et al.,* 2020). The present approaches for assessing soil quality metrics rely on wet chemical techniques or physical measurements, including granulometry (de Santana *et al.,* 2018). These conventional analytical procedures are lengthy, and a few of them generate remnants that must be removed in a subsequent phase. To measure soil parameters, soil scientists have traditionally employed laborious, lengthy, and hazardous (because of hazardous substances) methods in laboratory (Abou Samra *et al.,* 2018).

There is a growing need for alternative analytical techniques that can quickly, accurately, and simply detect a variety of soil variables due to the expansion of precision agriculture. Several strategies are used in soil quality prediction to determine whether the soil is appropriate for a crop before it is planted based on spectral data collected from the soil (Benke *et al.,* 2020). Smart soil forecasting represents an economical approach to predict soil performance across diverse crop types. Smart farming uses artificial intelligence (AI) to automate soil and crop management (Folorunso *et al.,* 2023). AI emulates how humans acquire knowledge and resolve problems (Chen et al., 2019). Artificial intelligence algorithms offer the possibility of predicting soil fertility, aiding in crop selection based on factors such as soil pH, soil nutrients, and precipitation. In precision agriculture, machine learning and deep learning algorithms are commonly utilized artificial intelligence methods for this purpose (Jiang *et al*., 2021; Folorunso *et al.,* 2023). Despite its increasing use worldwide, the lack of widespread acceptance of digital creative solutions is an impediment to high agricultural production in developing nations.

Thus, the concept of Machine Learning Algorithms was used to generate predictive models. Machine learning is an automated data analysis approach that streamlines the construction of analytical models (Sahour *et al.,* 2021). This investigation employs both Gradient-Boosted Regression trees and Random Forest regression models. Three performance error metrics (R2, ME, and RMSE) were utilized to evaluate the plausibility of the results. This solution in particular utilizes current data (Vis-NIR spectra and soil characteristics) to train machine learning and estimate soil properties for newly collected samples based on Vis-NIR reflectance (Gholizadeh *et al*., 2017). This technology is a greener alternative to conventional laboratory techniques and can assist farmers in obtaining quicker and more precise results. Additionally, scientists may always add new data to future model training and improvement.

Materials and Methods

Data collection: This research utilizes the ICRAF-ISRIC (World Agroforestry (ICRAF) and International Soil Reference and Information Centre (ISRIC)) world soil spectral library as the data repository for experiments in Machine Learning. This collection includes samples from 58 nations in Asia, Africa, Europe, South America, and North America. Vis-NIR spectra wavelengths range between 350 and 2500 nanometers and are divided into 216 wavebands (Pham *et al.,* 2021). After removing missing data and duplicate information, the remaining samples are separated into categories for training, testing, and validation. The set used for validation is utilized to identify the most suitable model trained with the training data set.

Prediction of soil property using spectral data: This research proposes predicting soil properties by analyzing Vis-NIR spectra derived from soil samples. Visible and near-infrared spectroscopy (Vis-NIRS) is considered one of the most captivating alternatives to conventional methods for regular soil analysis, as it has the potential to either fully or partially replace them. The simultaneous multi component analysis performed by Vis-NIRS spectroscopy in conjunction with chemometric tools. The spectrum encodes the natural soil composition, which consist of nutrients, organic substances, and water (Rossel *et al.,* 2016). The spectra depict all of these encodings as absorption at particular electromagnetic radiation wavelengths. It is believed to be fast, cheap, non-destructive, requires minimal sample preparation, is harmless to the environment (it doesn't use poisonous reagents), and can be connected to a machine to produce a precise result (de Souza *et al.,* 2016). Visible and infrared spectroscopy is an efficient method for soil characterization. Available spectroscopic measurements are rapid, accurate, and reasonably priced. We can employ their measurements to both numerically and qualitatively assess the soil (Hengl *et al.,* 2017). It is done through machine learning algorithm (Bondi *et al.,* 2018). Machine learning is a specialized branch of artificial intelligence based on the notion that systems can gather knowledge based on data, identify trends, and arrive at decisions using little assistance from humans. Methods for machine learning are typically categorized into two task categories: classification and regression. The former is used to predict labels, while the latter is used to predict quantities (Sahour *et al.,* 2021). Since we needed to determine the soil's fertility and nutrients, a regression-based method was used (Folorunso *et al.,* 2023).

Spectrum Pre-processing: Data errors are unavoidable when using hardware devices are used to record spectra (e.g., Vis-NIR spectroscopy). In order to obtain more comprehensible soil

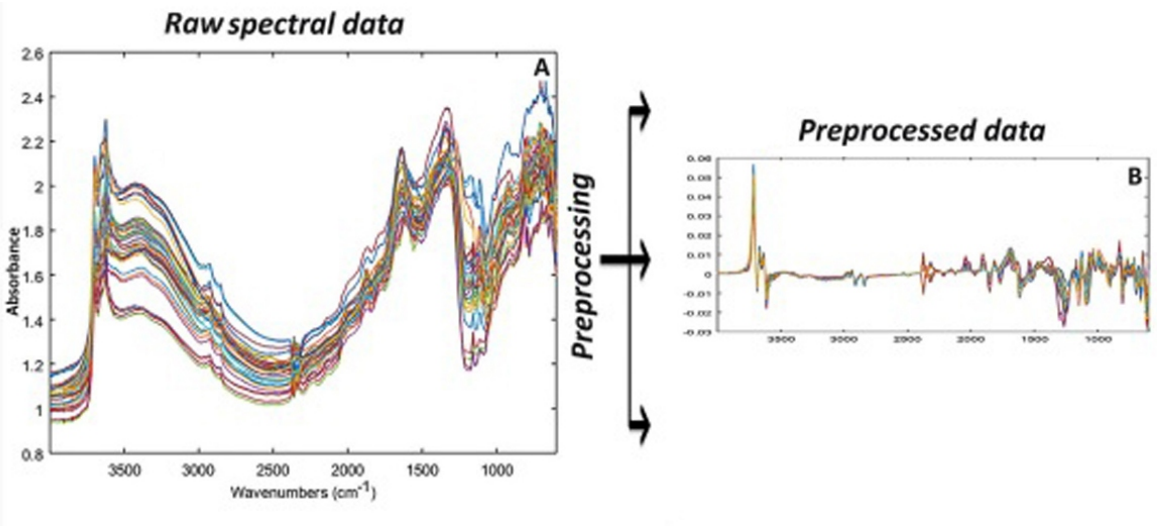


Fig. 1: Transformation of Raw Spectral data in to preprocessed data (using first derivative). Adapted from Barra *et al.* (2021).

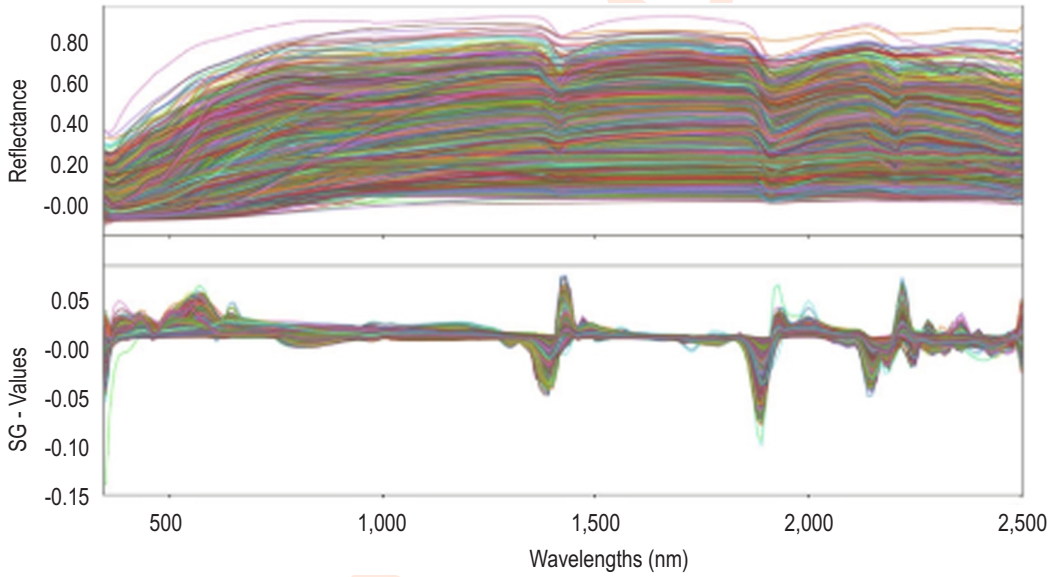


Fig. 2: Savitzky-Golay values- Visible and Near-Infrared(vis-NIR) Reflectance. Adapted from Pham *et al.* (2021).

spectral data, scientists frequently transform the data. Diverse spectral preprocessing methods generated distinct preprocessing outcomes. Nonetheless, numerous researchers in this field employ Savitzky-Golay (SG) for signal pre-processing (Ba *et al.,* 2020).

The SG smoothing method improves the signal-to-noise ratio, reduces the effects of high-frequency random noise from ground interposition, and changes the reflectance spectra. As a consequence, this effort also explored the use of SG to modify the data prior to employing Machine Learning techniques. This investigation employs an SG with the first derivative of order 5 polynomials and a window size of 11 as depicted in Fig. 1. Fig. 2 depicts the SG transformation applied to raw reflectance data. Additionally, the SG findings document the differences in reflectance measurements between each waveband. These SG transformation properties augment the training of machine learning (ML) models with additional information. Machine learning algorithms can perform better when trained on SG transformation data as opposed to raw reflectance data (Pham *et al.,* 2020).

Model description: Fig. 3 illustrates the architecture of the model. The preprocessed data is subsequently used for training machine learning algorithm. The preprocessed data enhances model performance compared to spectral reflectance data in its

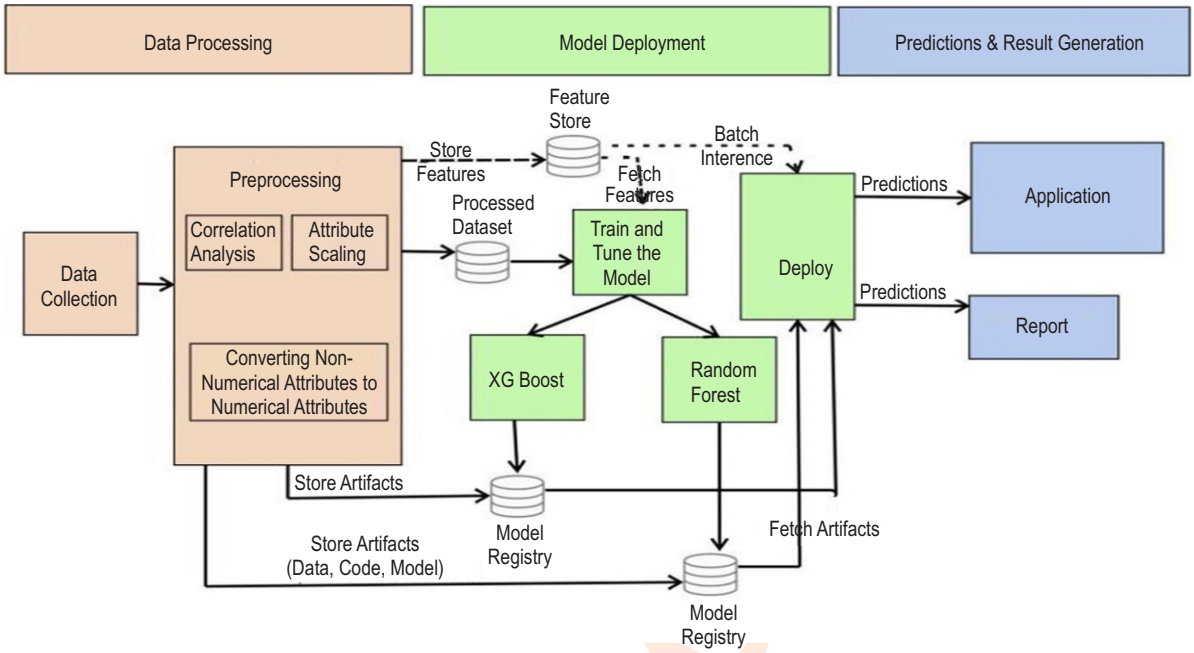


Fig. 3: Architecture Diagram for Soil Properties Prediction.

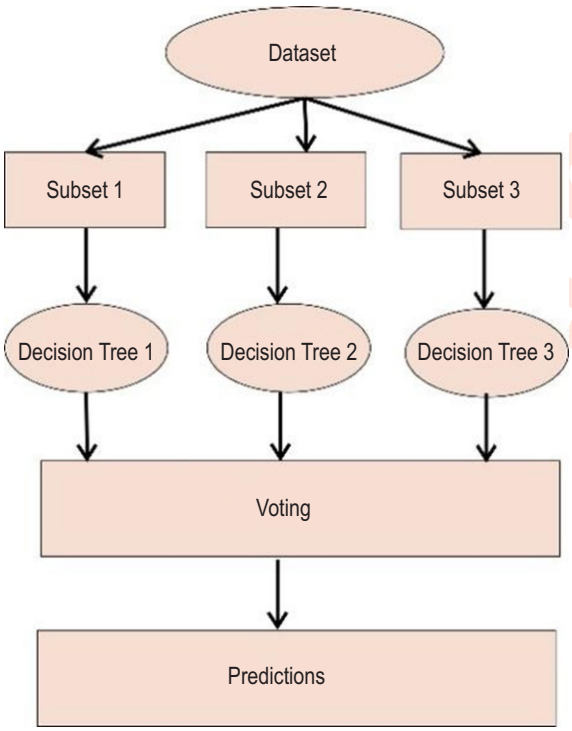


Fig. 4: Flow Diagram of Random Forest Algorithm.

unprocessed state.The machine learning model acquires data-based knowledge, identifies patterns, and predicts soil quality parameters (Gruszczyński *et al.,* 2022).

Algorithm Description: Gradient-boosted regression trees and Random forests are the Machine Learning methods utilized in this study due to small sample size and diverse data attributes.

Random Forest (Regression) Algorithm: Random Forest (RF) is a technique based on the ensemble method that makes conclusions using classification or regression trees. The decision trees are tree-based techniques for addressing both regression and classification issues with binary-split data. The trees are instructed to solve regression problems by minimizing the sum of squared variances from the average. Breiman's classification and regression trees (CART) are among the available algorithms. Nonetheless, the CART algorithm may generate models with no viable outcomes, necessitating pruning to improve the model of regression. Breiman devised the random forest, which employs a collection of trees despite aggregation, to address this issue. It aids in obtaining more precise outcomes while simultaneously making the model less vulnerable to over-fitting problem (de Santana *et al.,* 2018). Fig. 4 shows the process involved in creation of Random Forest model. The ensemble of decision trees is constructed as a set of well-organized trees denoted by T1(θ), T2(θ), ..., Tb(θ), where b represents the total number of trees, and θ refers to the collection of bootstrap samples with replacement from the initial training data, which consists of s samples and p variables per sample. The set θ = { θ1, θ2,... θB} has a dimension of m*n, with m being approximately two-third of s, and p representing the number of stochastic variables used at each node of the ensemble tree (de Santana *et al.,* 2018). The regression tree is constructed using two-third of the sample. In conjunction with the training phase, one-third of the bootstrap samples are employed for cross-validation methods. These specimens are known as out-of-bag (OOB) samples, and they are

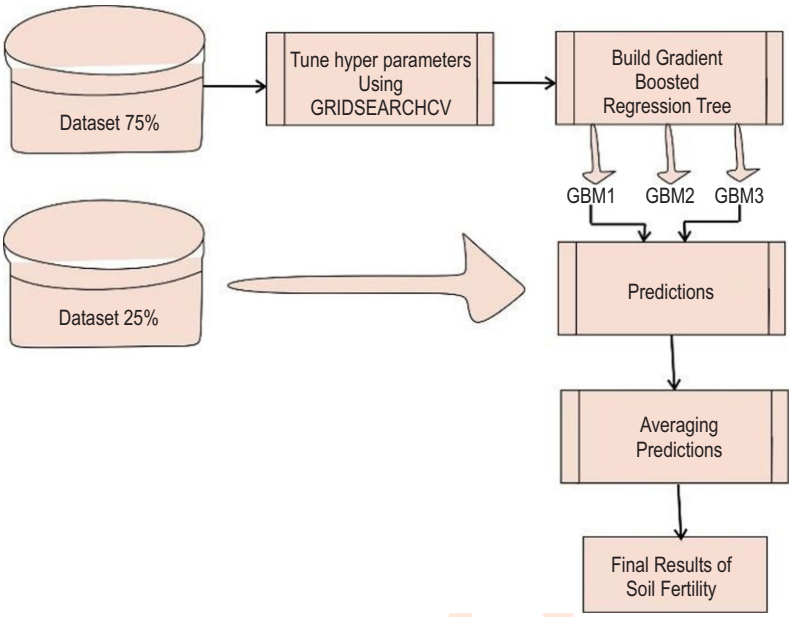


Fig. 5: Process Flow Diagram of Gradient Boosted Regression Tree.

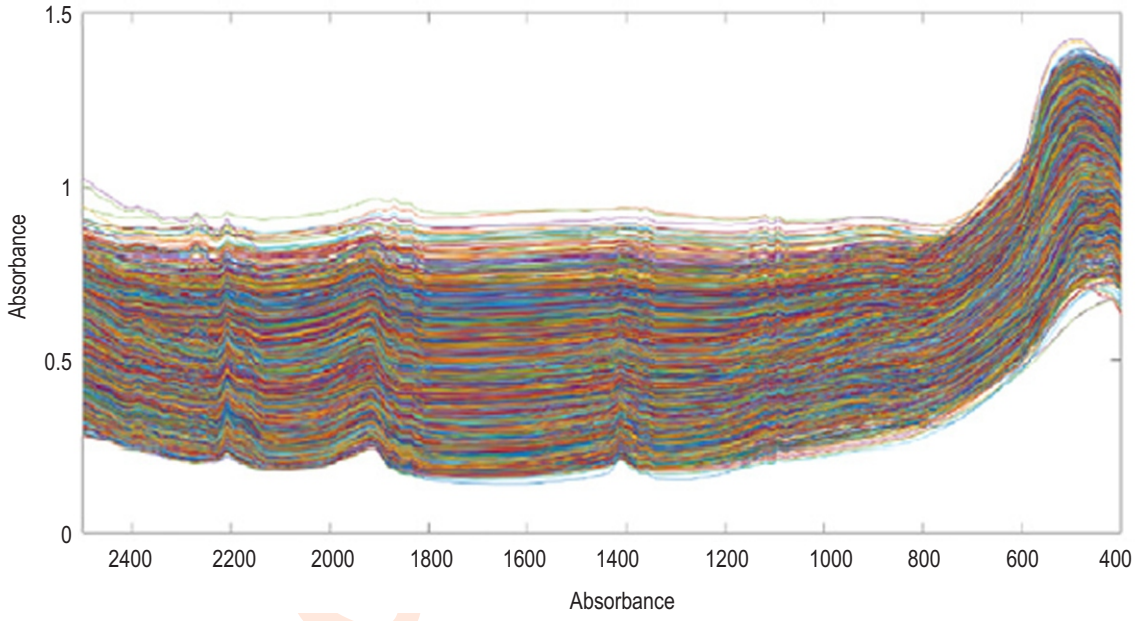


Fig. 6: vis-NIR Spectra of Soil samples (Adapted from De Santana *et al.* (2018).

used to determine the model's accuracy. The primary tuning parameters for a random forest are p, the total count of trees, and the lowest node size. The value of "p" parameter can vary from 1 to n (the total quantity of variables). The total number of trees must be adequate to stabilize the Out-of-bag sample error under second alternative. In broad terms, 500 trees are adequate; however, if many trees are preferred, the outcome will not be statistically distinct, but the creation of the model will take longer. The minimum node size specifies the minimal number of nodes for which no attempt is made to divide; the standard settings for regression and classification are 5 and 1, respectively (de Santana *et al.,* 2018). Random forests possess enticing characteristics, such as the ability to utilize data sets with latent values (Zhao *et al.,* 2016). In addition, the proximity matrix

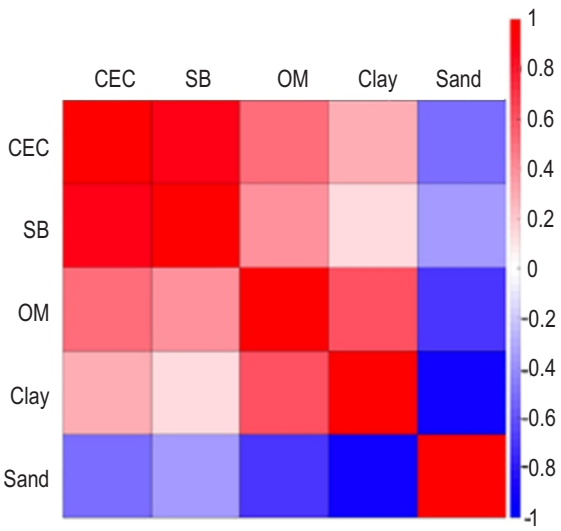


Fig. 7: Correlogram depicting Soil Properties (Adapted from De Santana *et al.* (2018)

(Afanador *et al.,* 2016) can be used for determining the significance of parameters and identifying anomalies. The proximity matrix is a measure of sample similarity, determined by dividing the total number of trees by the count of instances where two samples follow the same path in the tree and reach the same terminal node. This metric ranges from 0 to 1, with 1 indicating a high similarity between samples, and 0 indicating a lack of similarity between samples (de Santana *et al.,* 2018).

Gradient Boosted Regression Tree (GBRT): The gradient-boosted regression algorithm is an error function-based optimization algorithm. Gradient-boosted Regression trees are a machine learning technique used to solve regression as well as classification issues. The combination of weak models for forecasting, like decision tree, yields a robust prediction model. The GBRT algorithm was then created by Friedman. As depicted in Fig. 5, every calculation is carried out by a fundamental model, and each subsequent calculation reduces the remaining parameters of the previous model and generates a completely new fundamental model with fewer gradient residuals (Wei *et al.,* 2019). As a result, the loss function can be reduced and enhanced by perpetually modifying and improving the weak learner's weight in order to make it a better performer.

Performance Metrics: The model's accuracy is assessed using statistical coefficients, including R^2 (R Square), Root Mean Square Error (RMSE), and Mean Squared Error (MSE).The closer R^2 is to 1, the better the model's stability and degree of fit. The Root Mean Squared Error and Mean Squared Error are used to assess the model's prediction capabilities. As the Root Mean Squared Error (RMSE) and Mean Squared Error (MSE) decrease, the prediction accuracy improves (Wei *et al.,* 2019).

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

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

$$MAE = \frac{1}{m}\sum_{i=1}^{n}|y_i - \hat{y}_i| \longrightarrow 3$$

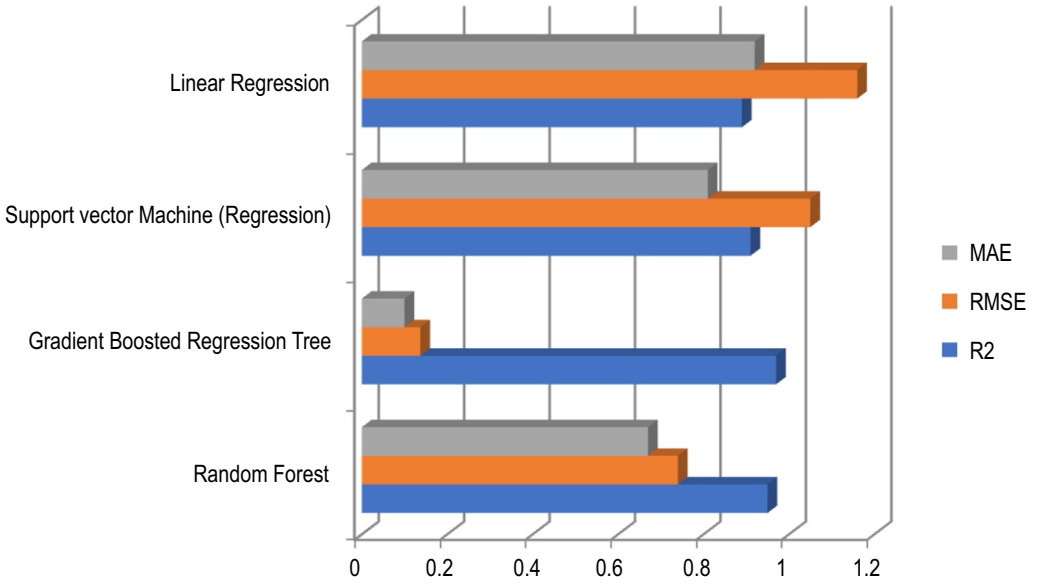


Fig. 8: Comparative Analysis of Various Machine Learning Models.

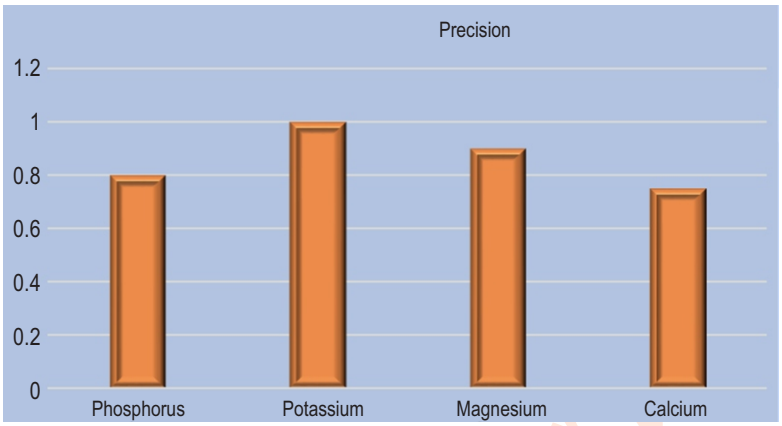


Fig. 9: Precision Chart for Random Forest Algorithm.

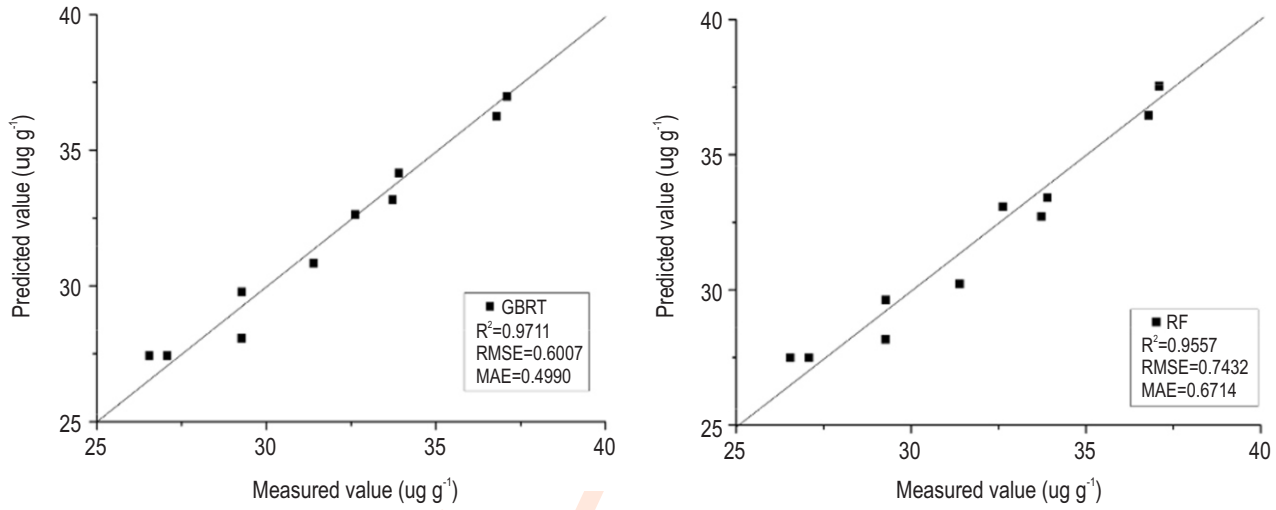


Fig. 10: Comparison of prediction accuracies of Gradient Boosted Regression Tree and Random Forest. Adapted from Wei *et al.* (2019).

Where n denotes the count of samples, y_i indicates observed value, $\hat{y_i}$ iindicates the predicted value, and y represents the mean of values measured.

Results and Discussion

The purpose of this investigation was to ascertain the viability of spectral information for assessing soil properties. In comparison to conventional analytical methods (Gangotry *et al.,* 2022), the advantages of soil spectroscopy include the speed and affordability of observations and soil property estimates(Shen *et al.,* 2021). The spectrum of soil reflection is rich and comprehensive, reflecting its chemical as well as physical characteristics. Visible-near-infrared diffuse reflectance (vis-NIR) spectroscopy has proven to be a cost-effective, environmentally friendly, non-invasive, and reproducible measurement technique that is optimal for soil property analyses (Ba *et al.,* 2020).

Vis-NIR Spectral Information: Visible and near-infrared reflectance spectroscopy (vis-NIRS) spectra of all soils exhibit minor visible and near-infrared differences (Fig. 6). Vis-NIRS spectra comprises fewer absorption regions than MID-IR (mid-infrared) spectra as a result of their broad and contiguous bands, making them difficult to interpret (Suchithra *et al.,* 2018). As previously reported by Bellon-Maurel and McBratney (2011); Viscarra Rossel *et al.* (2006), absorption at 400–600 nm, 1100 nm, 1400 nm, 1800–2000 nm and 2200–2400 nm are observed in the spectra of majority of soils. Iron-containing minerals, such as hematite and goethite, account for the majority of visible-range (400–780 nm) absorptions. Due to chromophores and obscurity

Table 1: Performance comparison chart

Machine Learning Model	RMSE	R^2
Gradient Boosted Regression Tree	0.25	0.97
Random Forest (Regression Model)	0.27	0.749

Table 2: Accuracy rate of Gradient Boosted Regression Tree

Soil Attribute	Training Accuracy (%)	Testing Accuracy (%)	Inference of Model	Justification for Inference
Calcium	96	88.8	✓ Brilliant (with Fine tuning)	✓ GBRT's sequential learning enables it to efficiently fine-tune predictions.
pH	94.7	79.1		✓ It performs well on validation and testing.
Phosphorous	84.9	82		✓ It can handle complex interactions successfully, and with proper tuning, it may be stable, ensuring consistent performance.

Table 3: Accuracy rates of Random Forest Algorithm

Soil Attribute	Training Accuracy (%)	Testing Accuracy (%)	Inference of Model	Justification for Inference
Calcium	97.1	94.4	✓ Brilliant	✓ Because of its ensemble nature, Random Forest is stable.
pH	94.7	79.1	✓ Over-fitting	✓ It excels in accuracy.
Phosphorous	83.8	27		✓ Random Forest identifies feature relevance.
				✓ But while predicting phosphorous, the model overfits.Overfitting means the model outperforms in the training set but fails to generalize to the testing set.

of organic matter, organic matter in certain soils can exhibit wavelengths of absorption in that region (de Santana *et al.,* 2018) . Absorptions in the NIRS region are caused by the harmonics of hydroxide, suphate, carbonate, nitrogen hydride, methylydine and carbon monoxide groups, as well as certain basic properties such as carbon dioxide and water. (de Santana *et al.,* 2018). Owing to significant relationships between these parameters and a number of additional parameters that exhibit reflectance in vis-NIRS (Fig. 7), it's feasible to construct regression algorithms known as second-order estimates for calculating the values of cation exchange capacity and sum of exchange bases. Important to observe is that relationships between parameters vary by region and soil type (Terra *et al.,* 2015).

Comparing Random Forest (RF) vs Partial Linear Squared Regression Model (PLS): According to de Santana *et al.* (2018), random forest and PLS yield substantially different results, suggesting that the RF algorithm is preferable for quantifying clay, CEC, OM and sand in soil samples. PLS forecasted negative results for certain variables for certain samples whereas models based on random forests failed to demonstrate such a problem. The number of discrepancies omitted from the random forest model in each set had been substantially less when compared to PLS, which is advantageous for practical applications.

Comparison of Random Forest (RF), Gradient Boosted Regression Tree (GBRT) and other models: Fig. 8 demonstrates that the GBRT regression method produces favorable inversion results (Svetnik *et al*., 2003; Wang *et al*., 2020). In conclusion, the R^2 coefficients of the validation collections for every model were greater than 0.84, indicating that the total precision adheres to the actual requirements, in comparison, GBRT demonstrates superior robustness and predictive power compared to other three models, and it also yields a more favorable inversion effect (Wei *et al.,* 2019; Liu *et al*., 2017).

Our Findings: The tables that follow provide an overview of the prediction accuracies generated during training and evaluation phases of the model based on the objective variable.

Gradient Boosted Regression Tree: Table 2 demonstrates that the gradient-boosted regression tree revealed excellent training and testing accuracy for the prediction of all soil properties, including phosphorus. But it is expensive to compute, difficult to comprehend, and largely depends on the training data's order (or sequence).

Random Forest Algorithm: Table 3 shows that Random Forest also exhibits good prediction capabilities. But in predicting phosphorous, the model frequently over-fits. The precision attained by Random Forest Algorithm for predicting few soil properties is depicted in Fig. 9. The greater the accuracy, the closer the actual and predicted values are to the 1:1 line on the scatter plot, as depicted in Fig. 10.The GBRT predictive model demonstrates a minimal shift from the 1:1 line as well as the greatest degree of fit. (Wei *et al.,* 2019). As a result of their excellent training and testing accuracy, both Random Forest Machine Learning algorithms and Gradient Boosted Regression Tree are the best algorithms to employ for predicting soil properties. The remarkable accuracy of random forests for forecasting soil properties, vital nutrients, and fertility is possibly attributable to a variety of these benefits. Because Random Forest (RF) is less susceptible to noise and anomalies than other algorithms, it may be capable of producing accurate forecasts even when dealing with equivocal or insufficient soil data. The ensemble feature of Random Forest, which incorporates numerous decision trees, boosts forecast accuracy and minimizes bias by utilizing the pooled knowledge of numerous trees (Padarian *et al.,* 2019). RF is a fitting choice for forecasting soil quality values influenced by multiple factors, such as pH levels, humidity, and nutrient amounts, as it can effectively handle large datasets with diverse characteristics. RF automatically selects the most important aspects for forecasts, which can support in recognizing the most important soil characteristics and nutrients for determining soil fertility (Folorunso *et al.,* 2023) In predicting phosphorus, the model frequently over-fits. This is because target variables and dataset feature correlate only sometimes. By including additional model attributes, the issue can be addressed in future.

The research presented here contributes to the comprehension of Vis-NIRS technique by providing a productive, economical, and hygienic approach in addition to accurate results. The methodology is likely to be implemented in regular soil research facilities to assess the condition of soil metrics using an automatic analytical procedure. Our future work focus on improving the accuracy further to 98% and, we aim to predict the pollutant traces in soil using hyper spectral estimation.The act of analyzing and interpreting data acquired across a large range of wavelengths, often beyond the visible and near-infrared (Vis-NIR) spectrum, is known as hyperspectral estimation. While Vis-NIR spectra encompass wavelengths from 350 to 2500 nm and are separated into 216 wavebands, hyperspectral data frequently comprises hundreds or even thousands of contiguous small spectral bands. These multiple bands provide highly detailed information, allowing the distinction of minor spectral properties that might not be apparent in a broader wavelength range.

Authors' contribution: A. Divya: Identified problem statement and proposed methodologly; **R. Josphineleela:** Carried out extensive literature survey; **L. Jaba Sheela:** Wrote results and discussion and evaluated the manuscript.

Funding: No funding received.

Research content: The research content of manuscript is original and has not been published elsewhere.

Ethical approval: Not applicable.

Conflict of interest: The authors declare that there is no conflict of interest.

Data availability: No permission required for data usage.

Consent to publish: All authors agree to publish the paper in *Journal of Environmental Biology.*

References

Abou Samra, R.M. and R.R. Ali: The development of an overlay model to predict soil salinity risks by using remote sensing and GIS techniques: a case study in soils around Idku Lake, Egypt. *Environ. Monito. Assess.,* **190**, 1-16 (2018).

Afanador, N.L., A. Smolinska, T.N. Tran and L. Blanchet: Unsupervised random forest: a tutorial with case studies. *J. Chemom.,* **30**, 232-241 (2016).

Ba, Y., 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,* **229**, 117863 (2020).

Barra, Stephan M. Haefele, R. Sakrabani and F. Kebede: Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advances–A review. *TrAC Trends in Analytical Chemistry*, **135**, 116166 (2021).

Bellon-Maurel, V. and A. McBratney: Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils – Critical review and research perspectives. *Soil Biol. Bioche.,* **43**, 1398-1410 (2011).

Benke, K.K., S. Norng, N.J. Robinson, K. Chia, D.B. Rees and J. Hopley: Development of pedotransfer functions by machine learning for prediction of soil electrical conductivity and organic carbon content. *Geoderma*, **366**, 114210 (2020).

Bondi, G., R. Creamer, A. Ferrari, O. Fenton and D. Wall: Using machine learning to predict soil bulk density on the basis of visual parameters: Tools for in-field and post-field evaluation. *Geoderma,* **318**, 137-147 (2018).

Chen, D., N. Chang, J. Xiao, Q. Zhou and W. Wu: Mapping dynamics of soil organic matter in croplands with MODIS data and machine learning algorithms. *Sci. Total Environ.,* **669**, 844-855 (2019).

de Santana, F.B., A.M. de Souza and R.J. Poppi: Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters. *Spectrochimica Acta* Part A: *Molecular and Biomolecular Spectroscopy,* **191**, 454-462 (2018).

de Souza, A.M., P.R. Filgueiras, M.R. Coelho, A. Fontana, T.C.B. Winkler, P. Valderrama and R.J. Poppi: Validation of the near infrared spectroscopy method for determining soil organic carbon by employing a proficiency assay for fertility laboratories. *J. Near Infra. Spectro.,* **24**, 293-303 (2016).

Folorunso, O., O. Ojo, M. Busari, M. Adebayo, A. Joshua, D. Folorunso, C. Ugwunna, O. Olabanjo and O. Olabanjo: Exploring machine learning models for soil nutrient properties prediction: A systematic review. *Big Data Cogni. Comput.,* **7**, 113 pages (2023).

Gangotry, M., S.U. Nandhini, G. Vijayalashmi, K. Manigundan, B. Abirami and M. Radhakrishnan: Isolation and bioactive potentials of Streptomyces from Tripura Forest soil, North-east India. *J. Environ. Biol.,* **43**, 764-770 (2022).

Gholizadeh, A., N. Carmon, A. Klement, E. Ben-Dor and L. Borůvka: Agricultural soil spectral response and properties assessment: effects of measurement protocol and data mining technique. *Rem. Sens.,* **9**, 1078 (2017).

Gruszczyński, S. and W. Gruszczyński: Supporting soil and land assessment with machine learning models using the Vis-NIR spectral response. *Geoderma,* **405**, 115451 (2022).

Hengl, T., J.G. Leenaars, K.D. Shepherd, M.G. Walsh, G.B. Heuvelink, T. Mamo, H. Tilahun, E. Berkhout, M. Cooper, E. Fegraus and I. Wheeler and N.A. Kwabena: Soil nutrient maps of Sub-Saharan Africa: assessment of soil nutrient content at 250 m spatial resolution using machine learning. *Nutri. Cycl. Agroecosy.,* **109**, 77-102 (2017).

Jiang, Z.D., P.R. Owens, C.L. Zhang, K.R. Brye, D.C. Weindorf, K. Adhikari, Z.X. Sun, F.J. Sun and Q.B. Wang: Towards a dynamic soil survey: Identifying and delineating soil horizons *in-situ* using deep learning. *Geoderma,* **401**, 115341 (2021).

Keerthan Kumar, T.G., C.A. Shubha and S.A. Sushma: Random Forest algorithm for soil fertility prediction and grading using machine learning. *Int. J. Innov. Technol. Explor. Eng.,* **9**, 1301-1304 (2019).

Liu, L., M. Ji and M. Buchroithner: Combining partial least squares and the gradient-boosting method for soil property retrieval using visible near-infrared shortwave infrared spectra. *Rem. Sens.,* **9**, 1299 (2017).

Massawe, B.H., S.K. Subburayalu, A.K. Kaaya, L. Winowiecki and B.K. Slater: Mapping numerically classified soil taxa in Kilombero Valley, Tanzania using machine learning. *Geoderma,* **311**, 143-148 (2018).

Motwani, A., P. Patil, V. Nagaria, S. Verma and S. Ghane: Soil Analysis and Crop Recommendation using Machine Learning. In: 2022 International Conference for Advancement in Technology (ICONAT). IEEE, pp. 1-7 (2022).

Padarian, J., B. Minasny and A.B. McBratney: Using deep learning to predict soil properties from regional spectral data. *Geode. Regio.,* **16**, e00198 (2019).

Paul, G.C., S. Saha and K.G. Ghosh: Assessing the soil quality of Bansloi river basin, eastern India using soil-quality indices (SQIs) and Random Forest machine learning technique. *Ecol. Indica.,* **118**, 106804 (2020).

Pham, V., D.C. Weindorf and T. Dang: Soil profile analysis using interactive visualizations, machine learning, and deep learning. *Comp. Electro. Agricul.,* **191**, 106539 (2021).

Rossel, R.V., T. Behrens, E. Ben-Dor, D.J. Brown, J.A.M. Demattê, K.D. Shepherd, Z. Shi, B. Stenberg, A. Stevens, V. Adamchuk, H. Aïchi, B.G. Barthès, H.M. Bartholomeus, A.D. Bayer, M. Bernoux, K. Böttcher, L. Brodský, C.W. Du, A. Chappell, Y. Fouad and W. Ji: A global spectral library to characterize the world's soil. *Earth-Sci. Revi.,* **155**, 198-230 (2016).

Sahour, H., V. Gholami, M. Vazifedan and S. Saeedi: Machine learning applications for water-induced soil erosion modeling and mapping. *Soil Tilla. Res.*, **211**, 105032 (2021).

Shen, S. L., A. Njock, A. Zhou and H.M. Lyu: Dynamic prediction of jet grouted column diameter in soft soil using Bi-LSTM deep learning. *Acta Geotech*., **16**, 303-315 (2021).

Suchithra, M.S. and M.L. Pai: Improving the performance of sigmoid kernels in multiclass SVM using optimization techniques for agricultural fertilizer recommendation system. In: Soft Computing Systems: Second International Conference, ICSCS 2018. Springer Singapore, pp. 857-868 (2018).

Svetnik, V., A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan and B.P. Feuston: Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inform. Comp. Sci.,* **43**, 1947-1958 (2003).

Terra Fabrício S., José, A.M. Demattê, A. Raphael and Viscarra Rossel: Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis–NIR and mid-IR reflectance data, Geoderma, **255–256**,81-93 (2015).

Viscarra Rossel, R. A., D.J.J. Walvoort, A.B. McBratney, L.J. Janik and J.O. Skjemstad: Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties,*Geoderma*,**131**, 59-75 (2006).

Wang, J., J. Ding, D. Yu, D. Teng, B. He, X. Chen, X. Ge, Z. Zhang, Y. wang, X. Yang, T. Shi and F. Su: Machine learning-based detection of soil salinity in an arid desert region, Northwest China: A comparison between Landsat-8 OLI and Sentinel-2 MSI. *Sci. Total Environ.*, **707**, 136092 (2020).

Wei, L., Z. Yuan, Y. Zhong, L. Yang, X. Hu and Y. Zhang: An improved gradient boosting regression tree estimation model for soil heavy metal (Arsenic) pollution monitoring using hyperspectral remote sensing. *Appl. Sci.,* **9**, 1943 (2019).

Zhao, T. and Y. Wang: Interpolation and stratification of multilayer soil property profile from sparse measurements using machine learning methods. *Engine. Geol.,* **265**, 105430 (2020).

Zhang, Y., A. Biswas and V.I. Adamchuk: Implementation of a sigmoid depth function to describe change of soil pH with depth. *Geoderma*, **289**, 1-10 (2017).