Journal of Information Systems Engineering and Management
2025, 10(8s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Predicting the Carbon Property of Soil Using DrSeqANN and VIS/NIR Spectroscopy

[1]Mr. Jakkan D A., [2]Dr. Pradnya Ghare, [3]Dr. Nirmal Kumar, [4]Dr. Chandrashekhar Sakode
[1]*Student Research: IIITN, Maharashtra, India.*
[2]*AP: India's VNIT Nagpur, Maharashtra.*
[3]*Sr Scientist Section of RSP at the ICAR NBSS & LUP, Nagpur, India.*
[4]*AP: IIITN, Maharashtra,India.*
[*1]*Author mail ID: d.Jakkan@gmail.com*

ARTICLE INFO

Received: 22 Oct 2024
Revised: 12 Dec 2024
Accepted: 22 Dec 2024

ABSTRACT

Plant productivity and health are directly impacted by carbon(C) levels. This study evaluated the potential of visible/near-infrared (V/NIR) spectroscopy (350-2,500 nm) for soil characterization, utilizing a dataset of 200 soil samples from Uttar Pradesh, India. The predictive performance of spectral data was compared across three modeling approaches: an Ensemble of Lasso and Ridge Regression models (ELRR), Random Forest (RF), and a more complex Artificial Neural Network (ANN) were employed to choose the spectral characteristics that were utilized in the C prediction.. To reproduce the spectrum's wavelength The log derivative, log to base 10 derivative log10x and inverse derivative were employed in the preprocessing.. The results showed that the availability of C was found to be between 350 and 450 nm. Using the Log10x pre-processed data and the suggested DrSeqANN-Dropout Sequential Artificial Neural Network technique, the most accurate results were obtained by accessing parameters with the aid of RMSE = 0.08, R2 = 0.82, and RPIQ = 4.32 for our suggested DrSeqANN model. Compared to the other two approaches

Keywords: Dropout-Sequential Artificial Neural Network (DrSeqANN) , data preparation techniques , Artificial Neural Network (ANN) , Spectral Wavelength , Logarithm (base 10) of the reciprocal of reflectance (Log10x),Near-Infrared(NIR)Spectroscopy

INTRODUCTION

Wetlands are vital carbon sinks, holding a significant share of the Earth's carbon reserves [1].make about 6% of the planet, The UNEP World Conservation Monitoring Center claims that. Approximately 14% of all carbon stored on land is found in wetland areas. Wetlands store a lot of carbon, Therefore, disruptions to the carbon stored in wetlands could have a major effect on the increase of global temperatures [2].One of the accepted methods for figuring out the amount of carbon (C) is dry combustion. The development of efficient, rapid, and accurate methods is necessary to address the difficult problem of extensive C monitoring, forecasting, and detection in arid conditions [3][4]. Despite their reputation for accuracy, traditional procedures can be laborious to use and run the risk of destroying materials during processing hinders the reproducibility of lab results. Still, recent research has demonstrated the non-destructive nature of visible infrared (V/NIR) spectroscopy. quantitative, affordable, and dependable method for determining the chemical makeup and quality of soil [5].

Scientists have developed several empirical models to determine the relationships between different soil elements (such phosphorus, or carbon, and several others) and reflected spectra [6]. Although there are many applications for spectrum data, machine learning approaches are particularly noteworthy for their quick and reliable dataset analysis [7][8]. The method generally performed well in identifying the spectra of soils as carbon concentration increased when using conventional regression approaches such multivariate linear regression or partial least squares regression [13][14]. With these methods, overestimation and underestimating are frequent problems [15]. The findings suggest that reliable estimates of C could be generated by random forests and support vector machines. In [16], V/NIR spectroscopy and SVM were used to measure the levels of C in samples from China's middle and lower Yangtze Rivers.

this research has demonstrated encouraging results when VIS/NIR spectroscopy data is combined with machine learning algorithms. Numerous studies have examined How to evaluate C in marshes [17] [18]. Several ML techniques, like iPLS-interval partial least squares and ACO-ant colony optimization, have recently emerged as methods for feature selection. Inspired by prior research and acknowledging these factors, the present study quantifies estimates of the C content of wetlands using V/NIR spectroscopy and machine learning approaches. The study took place in the Uttar Pradesh District of India. 200 soil samples, collected at multiple depths, were chemically analyzed. Spectral measurements were then made, and the data was pre-processed.

The Indian state of Uttar Pradesh was selected as the research area, and within it, soil samples were collected from the cities of Kanpur, Kanpur Dehat, Unnao, Raebareli, Amethi, Sultanpur, and Azamgarh , as shown in Figure 1. According to GPS coordinates, The pertinent place is located at latitude in Uttar Pradesh, India. 26.536938 and longitude 80.489960, or 26° 32' 12.9768' N and 80° 29' 23.8560' E. The results showed that For C, the key characteristic wavelengths fell between 350 and 450 nm. The best accurate results were obtained using the Log10x preprocessed data in conjunction with the suggested Dropout Sequential Artificial Neural Network (DrSeqANN) technique.

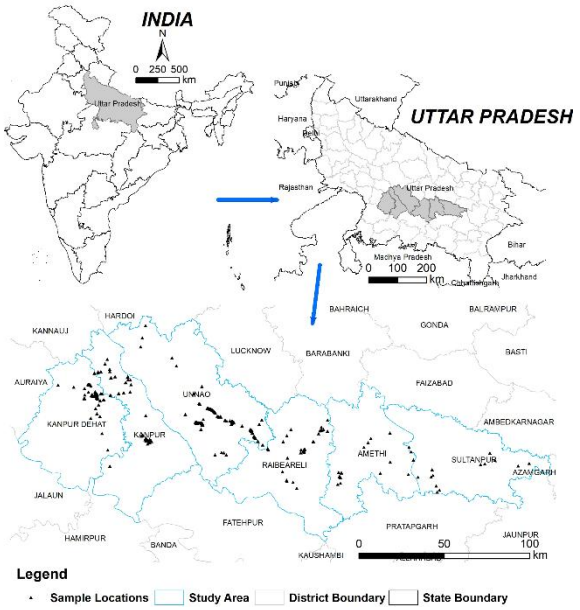


Figure 1: Study Area and sample points of Carbon(C).

A total of 35 sites were sampled (Figure-1). Samples were collected at four vertical depths of 5 cm, 20 cm, 40 cm, and 60 cm as well as five uniformly spaced locations with a grid of 30 to 30 m at each sampling site because Hyperion images have a spatial resolution of 30 m. In order to represent the soil for that sampling site (at that particular depth), the samples for each of the five locations (at that depth) were then uniformly mixed. 435 samples in all were gathered and delivered to the laboratory for chemical examination.

To get rid of any leftover plant matter, residues, roots, etc., all soil samples (n = 140) were completely air-dried, ground up, and sieved through a two mm mesh screen.The research region is shown in Figure-1. To get rid of any residual traces of plant remnants, roots, or stones, Each of the 200 soil samples was carefully allowed to air dry before being ground into a powder and sifted through a two-millimeter filter.200 soil samples with 2,151 properties ranging in wavelength from 350 to 2500 nanometers were gathered. High-frequency random disturbances, scattering anomalies, and baseline settling can all have an impact on spectral observations. The use of Origin Pro version 9.0 [13] improves the spectral features of the dataset [19]. The original spectrum of 200 soil samples was utilized in this investigation, and the First-Order Derivative (A'), its inverse (1/A'), its logarithm (lg(A')), and its log to base 10 (lg10(A')) were all employed.

Table-1: Matrix of Soil Properties by First and Third Quartile, Mean, Standard Deviation, and Maximum

Soil Property	Wavelength	Count	Mean	Std Deviation	Minimum	Ist Quartile	IIIrd Quartile	Maximum
Ph Extract	350	199	0.059554	0.017678	0.026468	0.046662	0.067182	0.13893
Ec Extract	351	199	0.060029	0.01739	0.02449	0.04878	0.068644	0.134328
CaCO3 Equivalent %	2499	199	0.356908	0.06267	0.225497	0.309888	0.393054	0.520923
C	2500	199	0.356783	0.062924	0.225079	0.308125	0.393934	0.524133

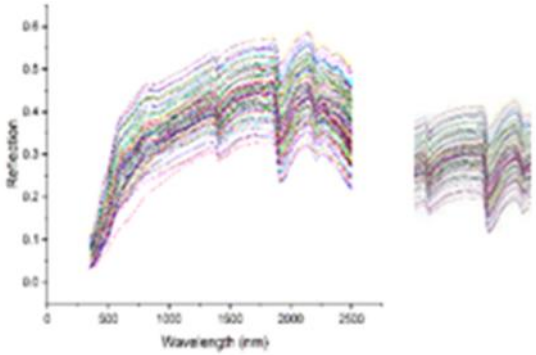
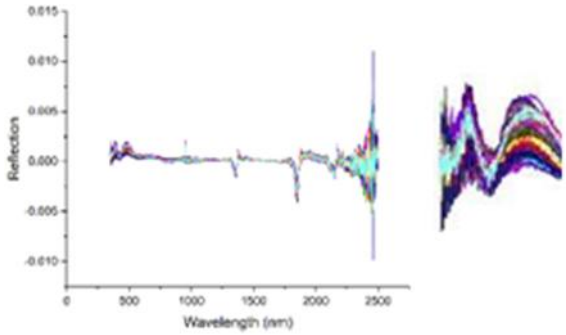
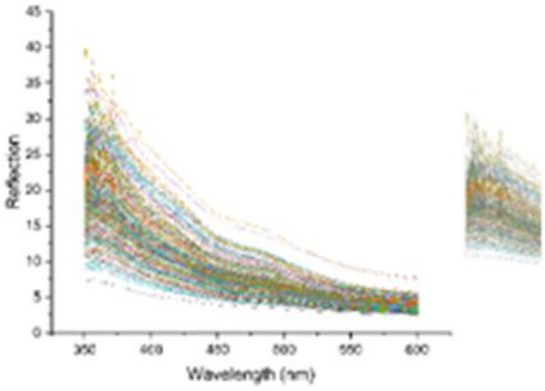


Figure 2 (a) Original Spectra



First-derivative analysis results are shown - Figure 2 (b)



The inverse derivative is displayed - Figure2(c)

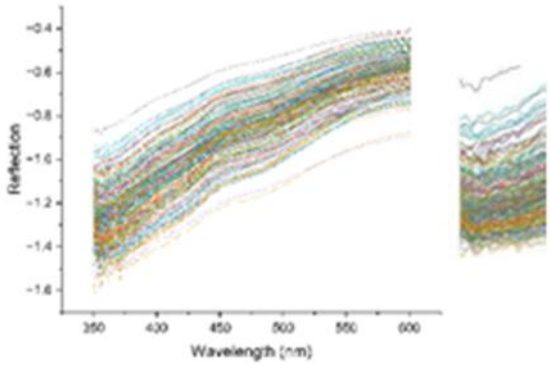


Figure 2 (d) Logarithmic-derivative

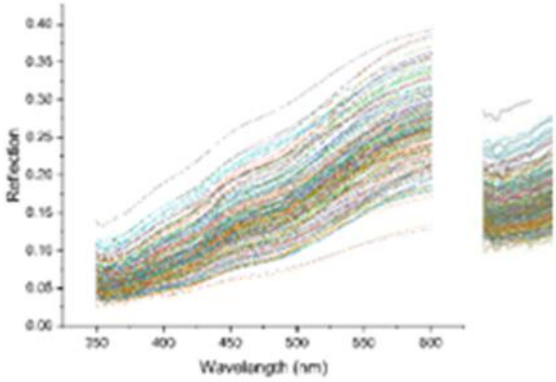


Figure 2 (e) Log10x derivatives of spectra

According to Figure 2, Log10x seemed to perform better than the other pre-processing techniques. The Origin Spectra's absorption peaks were located at ~1450 nm, ~1990 nm, and 2250 nm.

1.1. First-Derivative (A')

Derivatives are mainly used to improve resolution by separating overlapping peaks and removing linear and constant baseline variations across different samples. The spectra exhibited absorption peaks at 450, 480, 950, 1300, 1800 nm, and across the range from 2300 to 2500 nm. The mathematical expressions employed for pre-processing the first derivative (A) are presented in Equation-1.

$$\frac{d_y}{d_x} = \frac{f(x+h) - f(x)}{h} \qquad (1)$$

where

y-dependent variable

x-independent variable

d_x-change in value x

d_y-change in value y

h- limiting value

1.2. Inverse- First-Derivative (1/A')

Assume that f(x) is an invertible and differentiable function. For every x that satisfies f `(f-1(x)≠0, (1/A}), let y=f-1(x) inverse of f(x)

$$\frac{dy}{dx} = \frac{d}{d_x}\left(f^{-1}(x)\right) = (f^{-1})`(x)$$

$$= \frac{1}{f`(f^{-1}(x))} \qquad (2)$$

The absorption peaks in the spectra were situated in the 350–430 nm range. The formulas employed for inverse first-derivative preprocessing (1/A`) are detailed in Equation 2

1.3. Log A' (Log Derivative)

The mathematical formalism for the log derivative preprocessing step (Log A') is presented in Equation 3. Spectral absorption peaks were found to lie within the 350-400 nm range

$$\frac{d}{d_x} ln\, f(x) \quad \frac{1}{f(x)} \frac{df(x)}{dx} \qquad (3)$$

where as

f represents the function f(x).

One real variable is x.

1.4. *Derivative Log to Base 10 (Log10x)*

The absorption peaks in the spectrum were found between ~350 and ~450 nm. Equation 4 shows the mathematical techniques applied to Log Derivative (Log A) pre-processing

$$\log_{10} x_i^1 = x_i^1 - x_{i-1}^1 \qquad (4)$$

MACHINE LEARNING METHODS

To predict carbon (C) content using visible/near-infrared (V/NIR) spectroscopy, this study employed two machine learning techniques – Random Forest (RF) and Ensemble Lasso-Ridge Regression (ELRR) – in addition to the novel DrSeqANN model incorporating dropout layers. This analysis utilized a dataset of 200 soil samples collected from Uttar Pradesh, India. Model performance in predicting carbon content was evaluated based on RMSE, RPIQ, and R2 metrics, following spectral pre-processing. The results were then used to compare and contrast the three regression models using the NIR spectroscopy soil data

1.2. Ensemble Learning Using Lasso-Ridge Regression (ELRR)

Ensemble larning using lasso and ridge regression (ELRR) is a combination of Lasso and Ridge regression. The ELRR simultaneously executes automated variable selection and continuous shrinkage, or L1 and L2 penalty. The ELRR penalty consists of two distinct penalty functions.

$$L1 = \lambda \sum_{i=1}^{n} \left| \theta^2{}_i \right| \quad L2 = \lambda \sum_{i=1}^{n} \theta_i^2 \qquad (5)$$

In contrast,

Laso-Rgresion=L2, Ridge-Regression=L1

λ = Regularization parameter.

Θ = Total sum for the vector of theta

n = Number of features.

The ridge penalty (L1) and the lasso penalty (L2) are the first and second halves of the penalty, respectively. The penalty parameter, which accepts values between 0 and 1, is an attempt to find a compromise between the two penalties. The ELRR penalty has the advantage of combining the ridge regularization's successful regularization with the lasso penalty's feature selection capabilities.

1.3. Implementation of Random Forest (rf)

Random Forest (RF), a widely adopted supervised learning algorithm [1], is applicable to both regression and classification problems [1]. Grounded in the principles of ensemble learning, RF leverages the combined predictive power of multiple decision trees to improve overall model performance and address complex modeling challenges. The RF algorithm, acting as a regressor, enhances prediction accuracy by training multiple decision trees on diverse subsets of the data and averaging their respective outputs. This ensemble approach effectively reduces the risk of over-fitting and enhances the generalization capability of the model.

1.4. ANN- Artificial Neural Network

ANN [15], An ANN is a powerful computing system that is modeled after biological brain networks. Terms like "artificial neural systems," "parallel distributed processing systems," and "connectionist systems" were used to describe ANNs.

Every neuron is connected to other neurons by a connection. Each connector has some weights and is cited. It is believed that each neuron exists in an innate state that is distinguished by activation signals. When input signals are combined with the activation method, output signals may be sent to other components.

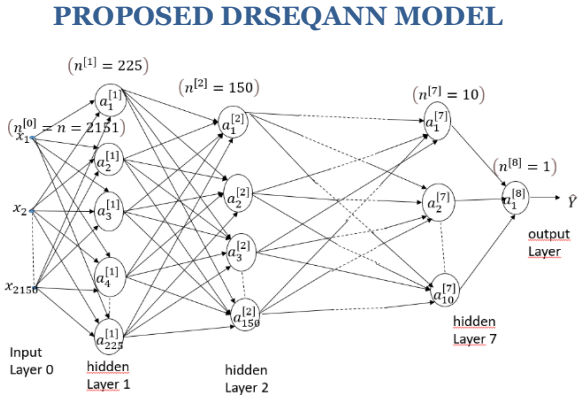


Figure 3: Structure of Neural Network

Between the input and output layers, seven hidden layers are used as shown in Figure 3. The input layer uses a total of 2151 spectral characteristics. There are 225 neurons used in the first hidden layer. The input batch can enter the network through an input layer. For every node in the input layer, a sample feature is as Cited. Up to the final (output) layer, several hidden layers are added after the input layer. These levels perform the "complex" nonlinear procedures with connections. Although considered "complex," the basic procedures are essentially quite simple mathematical computations. A stack of computational nodes is referred to as a hidden layer. From the input, each node extracts a feature.

A feature map, or representation, Up to the final (output) layer, several hidden layers are added after the input layer. This feature map intuitively shows the outcomes of different "sub-problems" that have been resolved at each node. In order to predict the reaction, they supply predictive data to the subsequent layer and all the way up to the output layer. Ten neurons are used in the seventh layer, which is the final layer. The hyperbolic tangent (Tanh) function was used as the activation function [20]. The data was divided into training and testing sets with a 70:30 split ratio The Root Mean Squared Error (RMSE) was used as the loss function and the "ADAM" optimizer was used to construct the model. L1 and L2 weight penalties were introduced as regularization techniques for neural networks [15]. These regularizations did not, however, completely eliminate the overfitting issue.

One major issue in learning large networks is co-adaptation. It is common for some links in such a network to forecast results more precisely than others. if all of the weights are learned simultaneously. In this case, the stronger connections learn more and the weaker connections are ignored because the network is taught frequently. Increasing the neural network's size

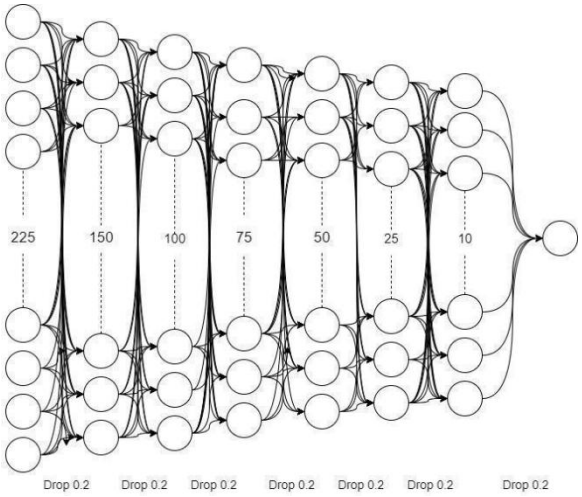


Figure 4: The Framework for Layer Dropping in a Regularized Network.

would not be beneficial. Consequently, the size and precision of neural networks were constrained. Dropout came next. A novel approach to regularization. The co-adaptation was restored. Now we could build deeper and larger networks. And exploit the predictive power of everything.

A batch of samples serves as the network's input, as seen in Figure 4. Every sample is a vector of features. A network's hidden layers are dense. Weight matrix W and bias b are characteristics of a dense layer. They carry out basic affine transformations (XW + b, dot product plus bias). Features are extracted from the input via the affine transforms. An activation function is applied to the transformations.

Nonlinear activations are present. The network can implicitly break down a difficult problem into arbitrary sub-problems because of its nonlinearity. The network combines the outputs of several sub-problems to deduce the ultimate output, $\hat{y}$. Dropout altered how weights were learned. It is common for some links in such a network to forecast results more precisely than others.

$$E_r = \underbrace{\frac{1}{2}(t - \sum_{i=1}^{n} p_i w_i l_i)^2}_{L_1} + \underbrace{\sum_{i-1}^{n} p(1-p_i) w_i^2 l_i^2}_{L_2} \qquad (6)$$

Equation 6 above illustrates the existence of two regulizers, L1 and L2. There seems to be a distinction between dropout's random weight suppression and L1's data-driven weight suppression, even though L1 regularization promotes sparsity by shrinking small weights to zero.

Dropout, however, is a regularization technique. This regularization resembles an L2 more. This is shown mathematically by Pierre and Peter (Baldi and Sadowski, 2013). They showed that, Under linearity assumptions regarding the activation function, the loss function's shape when using dropout mirrors that of L2 regularization. The dropout rate (p) represents the fraction of nodes randomly excluded during each batch iteration.

There is a penalty factor p(1-p) in the regularization term of equation 6. The component p(1 - p) reaches its maximum at p = 0.5. Consequently, for p = 0.5, the dropout regularization is greatest. Under linearity assumptions, dropout is a regularization method that is comparable to L2 regularization. For maximal regularization, a dropout rate of p = 0.5 is the best option. Thus, for hidden layers, Generally speaking, A dropout rate of 0.5 is recommended. Our model's dropout rate of 0.2 indicates its good performance. The activation function is tanh, the kernel initializer is normal, the bias value is zero, and the weights are automatically modified based on the input features. This process begins with the first input layer, which has 225 applied nodes, and continues to the seventh layer, which has 10 nodes.

3.1. Evaluation and comparison of model calibration

The DrSeqANN model trains for 1000 epochs, processing the data in five batches. A single epoch represents one complete iteration over the entire training dataset. Weight adjustments are made during training using a gradient-based optimization method.

The dataset was split into training and testing sets (140 and 60 samples, respectively) to evaluate the performance of the three algorithms (DrseqANN, RF, and ELRR). The coefficient of determination (R2), root mean squared error (RMSE), and ratio of performance to interquartile distance (RPIQ) were used to evaluate the algorithm's accuracy. which quantifies how well one measure performs in relation to another. To provide a more consistent and impartial assessment of model validity, RPIQ takes prediction error and volatility of detected measurements into account. A higher RPIQ value indicates a better capacity to forecast [22]. greater R2, lower RMSE, and greater RPIQ are all indicators of a more stable model.

Coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y})^2}{\sum_i (y_i - \bar{y})^2} \qquad (7)$$

where $y_i = Real\ values$

$\hat{y} = Estimated\ values$

$\bar{y} = The\ average\ of\ the\ values$

$y_i - \bar{y} = Y's\ deviation\ from\ its\ mean$

RMSE-Root Mean Squared Error $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{\propto} - y_{est})^2}$

$y_{\propto} = Real\ value$

$y_{est} = Predicted\ value$

$n = quantity\ of\ data\ points$

Ratio of Performance to Interquartile-RPIQ

Quartiles: The values that produce quarters in an integer list are called quartiles.

Upper Quartile: the middle value of a dataset's upper half.

Range: The range of a dataset is the difference between its highest and lowest values.

Interquartile Range: Interquartile Range: The difference between the dataset's upper and lower quartiles is represented by the interquartile range

RPIQ = Q3-Q1/RMSE

Or

RPIQ=IQ/RMSE

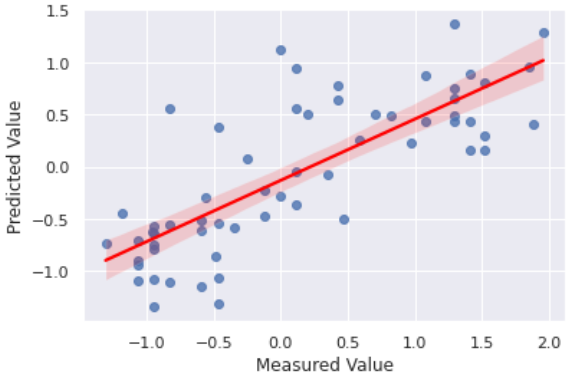
FINDINGS AND DISCUSSION

1.5. Evaluation via Comparison

DrSeqANN achieved a higher RPIQ (8.42) compared to RF (8.25) and ELRR (8.11), indicating better performance. The RPIQ value is highest when the ELRR model is applied to raw data (prior to preprocessing). Each model's mean squared error, The values of the root means square error and coefficient of determination exhibited very similar characteristics. Scatter plots illustrating the performance of the three models on the original data can be found in Figures 5 through 9. Pre-processing the data with derivative functions is the initial stage. The first derivative,

inverse derivative, log derivative, and log10x are used to complete the pre-processing. This facilitates determining whether these samples are included in the training sets from which the prediction models were developed.

(a)



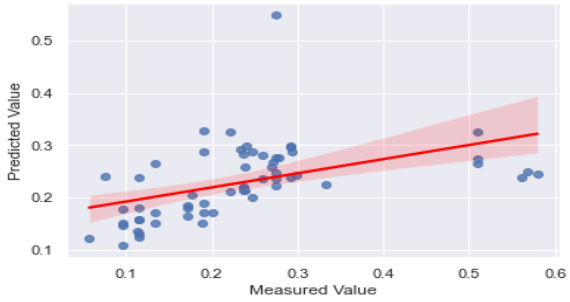
(b)

(c)

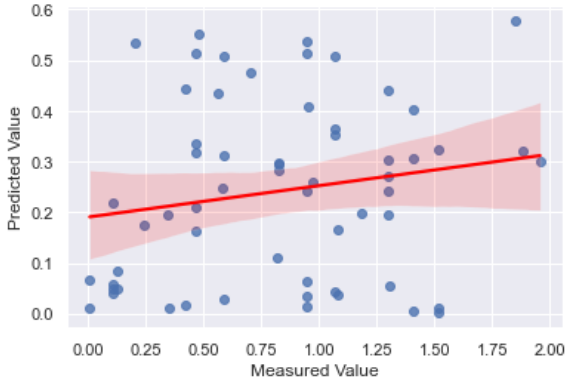
Figure 5: Scatter plots showing the original raw data (pre-preprocessing) for (a) DrSeqANN, (b) RF, and (c) ELRR.

The scatter plot for each of the three models prior to preprocessing is displayed in Figure 5.

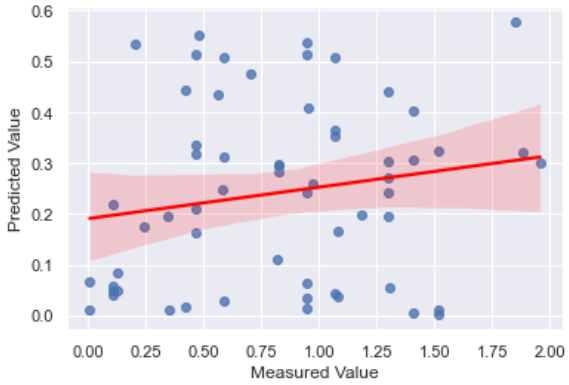
It can be shown from Figure 6 that the DrSe-qANN model performs better on Log10x pre-processing data. In contrast to the RF and ELRR models, the DrSeqANN model has less dispersed data points. The test data set yielded an R2 value of 0.82, an RMSE value of 0.08, and an RPIQ of 4.32.



(a)



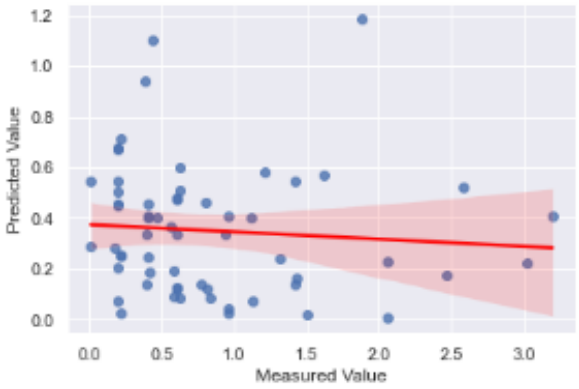
(b)



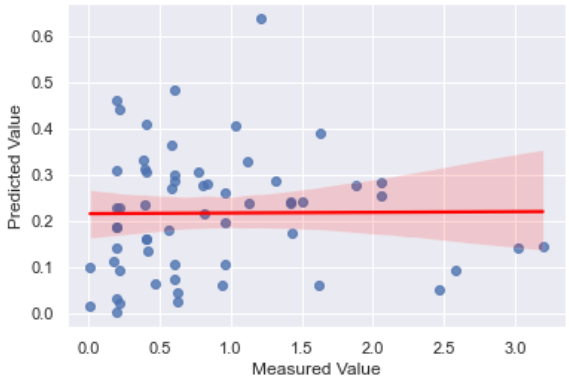
(c)

Figure 6: Log10x pre-processing data with the scatter plot (a) DrSeqANN, (b) RF, and (c) ELRR

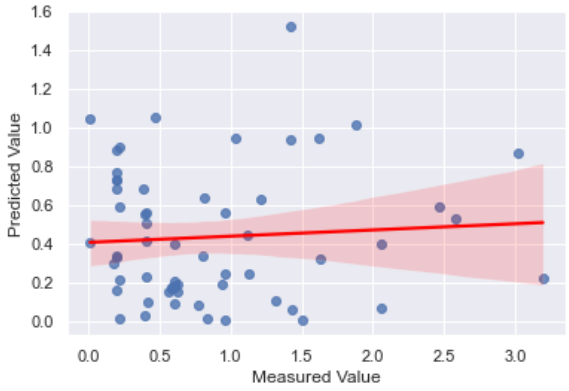
Figure 7 shows the scatter plots of the three models for regression analysis after inverse derivative pre-processing. The RF model exhibits improved performance with an R2 of approximately 0.55 and RMSE of 0.07 on the test data. As the scatter plots illustrate, the RF model displays less data dispersion compared to the DrSeqANN and ELRR models. .



(a)



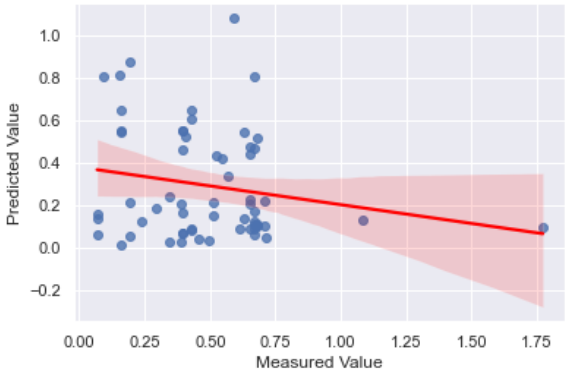
(b)



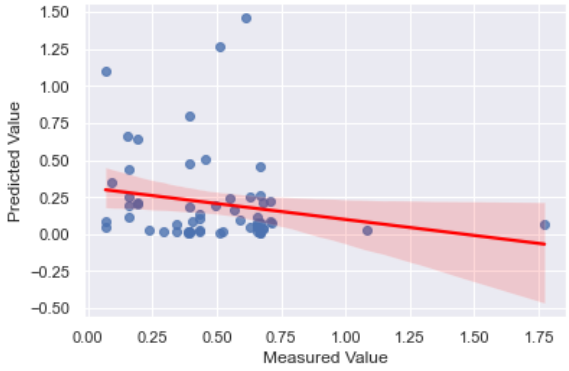
(c)

Figure 7: Model performance on inverse derivative pre-processed data, visualized using scatter plots: (a) DrSeqANN, (b) RF, (c) ELRR.

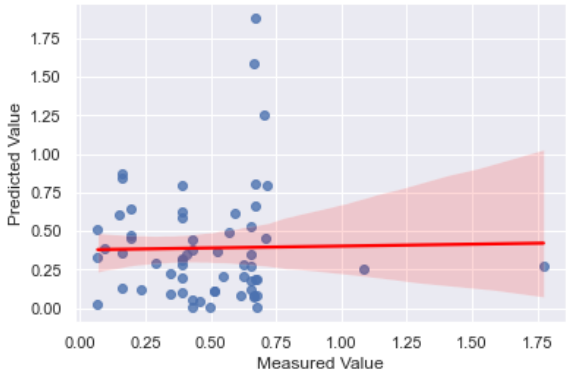
The scatter plots comparing the three models on a dataset that has been pre-processed using inverse derivative is shown in Figure 8. The values of the evaluation parameters (RMSE, RPIQ, and R2) are extremely similar to one another, with just a slight variation between them. With an RMSE of 0.08 and an R2 measure of roughly 0.69, the DrSeqANN model seems to be doing marginally better, according to figures 8a,b and c.



(a)



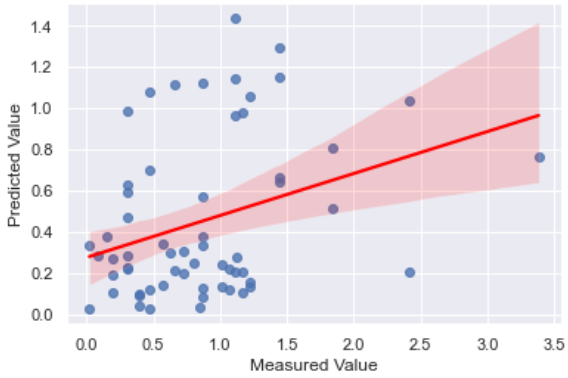
(b)



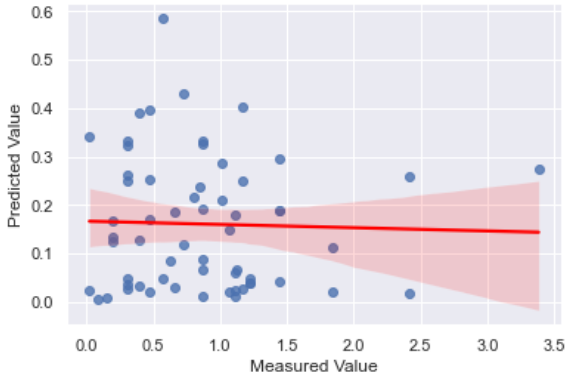
(c)

Figure 8: Scatter plots of data after inverse derivative pre-processing: (a) DrSeqANN, (b) RF, (c) ELRR

As shown in Figure 9, which presents the scatter plots for all three models that the DrSeqANN model performed better than the RF and ELRR models. R2 is 0.98 for the training set and 0.67 for the testing set. The RPIQ is likewise the greatest value, 3.87, in the case of the logarithmic derivative.



(a)



(b)

(c)

Figure 9: Scatter plots illustrating model performance on data following logarithmic derivative pre-processing: (a) DrSeqANN, (b) RF, and (c) ELRR.

The use of Log10x pre-processing improved the RMSE performance of the DrSeq-ANN model. For Random, the R2 value is good. Forest Model on the original data in comparison to the other two methods. It is found that Log10x pre-processing produces the best results when the pre-processing approaches covered in previous sections are applied. Our suggested DrSeqANN model has an RMSE of 0.08, an R2 of 0.82, and an RPIQ of 4.32. For most of the pre-processed data, the proposed DrSeqANN prototype outperforms the other two approaches. In comparison to RF and ELRR, the RPIQ value that we discovered is consistently higher.

CONCLUSION

The suggested DrSeq-ANN and regression machine learning models were shown to be able to predict C contents using Log10x, Inverse Derivative, First Derivative, and Logarithmic Derivative. On the given dataset, the recommended DrSeq-ANN model performed better for soil characteristics (C) than the Random Forest and ELRR mod-

els. Specifically, using a Log10x during pre-processing significantly improved the model's accuracy for R2 by 17.55%. This was achieved by assessing parameters using our suggested DrSeqANN model's RMSE = 0.08, R2 = 0.82, and RPIQ = 4.32. compared to the other two approaches. The DrSeqANN Model can be used in future studies for several kinds of soil samples, including sand, silt, and salt.

ACKNOWLEDGEMENT:

The authors extend their sincere appreciation to Dr. Nirmal Kumar, Senior Scientist at the Department of Remote Sensing, ICAR-National Institute of Soil Survey & Land Use Planning, Nagpur, for his continuous technical guidance. His assistance in accessing and maintaining the spectral library was crucial to the success of this work.

Data Availability:

The C data from the State of Uttar Pradesh, India, has all the information needed to create prediction models and is accessible upon request.

Funding:

This project is being funded in Phase II by the Indian Government's Ministry of Electronics and IT (MeitY) under the Visvesvaraya Ph.D. Scheme for Electronics and IT (Unique Awardee Number: MEITY-PHD-3080). Money donors have no say in how manuscripts are prepared, how studies are designed, how data is gathered and analyzed, or if a work is published.

REFERENCES

[1] M.-H. Hu, J.-H. Yuan, X.-E. Yang, and Z.-L. He, “Effects of temperature on purification of eutrophic water by floating eco-island system,” *Acta Ecol. Sin.*, vol. 30, no. 6, pp. 310–318, Dec. 2010, doi: 10.1016/J.CHNAES.2010.06.009.

[2] Y. Wang, L. Zhang, and Y. Haimiti, “Study on Spatial Variability of Soil Nutrients in Ebinur Lake Wetlands in China,” *https://doi.org/10.2112/SI73-011.1*, vol. 73, no. sp1, pp. 59–63, Jan. 2015, doi: 10.2112/SI73-011.1.

[3] M. Vohland, J. Besold, J. Hill, and H. C. Fründ, “Comparing different multivariate calibration methods for the determination of Carbonpools with visible to near infrared spectroscopy,” *Geoderma*, vol. 166, no. 1, pp. 198–205, Oct. 2011, doi: 10.1016/J.GEODERMA.2011.08.001.

[4] R. Kinoshita, B. N. Moebius-Clune, H. M. van Es, W. D. Hively, and A. V. Bilgilis, “Strategies for Soil Quality Assessment Using Visible and Near-Infrared Reflectance Spectroscopy in a Western Kenya Chronosequence,” *Soil Sci. Soc. Am. J.*, vol. 76, no. 5, pp. 1776–1788, Sep. 2012, doi: 10.2136/SSSAJ2011.0307.

[5] B. Kuang, Y. Tekin, and A. M. Mouazen, “Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon, pH and clay content,” *Soil Tillage Res.*, vol. 146, no. PB, pp. 243–252, Mar. 2015, doi: 10.1016/J.STILL.2014.11.002.

[6] R. A. Viscarra Rossel, 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*, vol. 131, no. 1–2, pp. 59–75, 2006, doi: 10.1016/j.geoderma.2005.03.007.

[7] S. Nawar and A. M. Mouazen, “Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques,” *CATENA*, vol. 151, pp. 118–129, Apr. 2017, doi: 10.1016/J.CATENA.2016.12.014.

[8] J. Wang *et al.*, “Desert soil clay content estimation using reflectance spectroscopy preprocessed by fractional derivative,” *PLoS One*, vol. 12, no. 9, Sep. 2017, doi: 10.1371/JOURNAL.PONE.0184836.

[9] G. M. Vasques, S. Grunwald, and W. G. Harris, “Spectroscopic Models of Soil Organic Carbon in Florida, USA,” *J. Environ. Qual.*, vol. 39, no. 3, pp. 923–934, May 2010, doi: 10.2134/JEQ2009.0314.

[10] D. Summers, M. Lewis, B. Ostendorf, D. C.-E. Indicators, and undefined 2011, “Visible near-infrared reflectance spectroscopy as a predictive indicator of soil properties,” *Elsevier*,doi:10.1016/j.ecolind.2009.05.001.

[11] B. Kayranli, M. Scholz, A. Mustafa, Å. H.- Wetlands, and undefined 2010, “Carbon storage and fluxes within freshwater wetlands: a critical review,” *Springer*, vol. 30, no. 1, pp. 111–124, Feb. 2010, doi: 10.1007/s13157-009-0003-4.

[12] P. T. Guo, M. F. Li, W. Luo, Q. F. Tang, Z. W. Liu, and Z. M. Lin, “Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach,” *Geoderma*, vol. 237–238, pp. 49–59, Jan. 2015, doi: 10.1016/J.GEODERMA.2014.08.009.

[13] T. Shi, L. Cui, J. Wang, T. Fei, Y. Chen, and G. Wu, "Comparison of multivariate methods for estimating soil total nitrogen with visible/near-infrared spectroscopy," *Plant Soil*, vol. 366, no. 1–2, pp. 363–375, May 2013, doi: 10.1007/S11104-012-1436-8/METRICS.

[14] Z. Shi *et al.*, "Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations," *Sci. China Earth Sci.*, vol. 57, no. 7, pp. 1671–1680, Feb. 2014, doi: 10.1007/S11430-013-4808-X/METRICS.

[15] K. Were, D. T. Bui, Ø. B. Dick, and B. R. Singh, "A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape," *Ecol. Indic.*, vol. 52, pp. 394–403, May 2015, doi: 10.1016/J.ECOLIND.2014.12.028.

[16] H. Xiaowei, Z. Xiaobo, Z. Jiewen, S. Jiyong, Z. Xiaolei, and M. Holmes, "Measurement of total anthocyanins content in flowering tea using near infrared spectroscopy combined with ant colony optimization models," *Food Chem.*, vol. 164, pp. 536–543, Dec. 2014, doi: 10.1016/J.FOODCHEM.2014.05.072.

[17] "Comparison of Discrete Wavelet Transform (DWT), Discrete Cosine Transform (DCT) and Stationary Wavelet Transform (SWT) based Satellite Image Fusion Techniques," *Int. J. Curr. Res. Rev.*, 2017, doi: 10.7324/ijcrr.2017.9129.

[18] A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, Jul. 1964, doi: 10.1021/AC60214A047/ASSET/AC60214A047.FP.PNG_V03.

[19] V. Svetnik, A. Liaw, C. Tong, J. Christopher Culberson, R. P. Sheridan, and B. P. Feuston, "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, pp. 1947–1958, Nov. 2003, doi: 10.1021/CI034160G/SUPPL_FILE/CI034160GSI20031008_041202.ZIP.

[20] J. Shunk, "Neuron-Specific Dropout: A Deterministic Regularization Technique to Prevent Neural Networks from Overfitting & Reduce Dependence on Large Training Samples," pp. 1–19, 2022, [Online]. Available: http://arxiv.org/abs/2201.06938.

[21] V. Bellon-Maurel, E. Fernandez-Ahumada, B. Palagos, J. M. Roger, and A. McBratney, "Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy," *TrAC Trends Anal. Chem.*, vol. 29, no. 9, pp. 1073–1081, Oct. 2010, doi: 10.1016/J.TRAC.2010.05.006.

[22] "[Comparative analysis of soil organic matter content based on different hyperspectral inversion models] - PubMed." https://pubmed.ncbi.nlm.nih.gov/23586255/ (accessed Jan. 11, 2023).

[23] J. Song *et al.*, "Estimation of Soil Organic Carbon Content in Coastal Wetlands with Measured VIS-NIR Spectroscopy Using Optimized Support Vector Machines and Random Forests," *Remote Sens.*, vol. 14, no. 17, 2022, doi: 10.3390/rs14174372.