International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

Carbon Property of Soil Prediction By VIS/NIR Spectroscopy Using DrSeqANN

[*1]Mr. Digambar Aggayya Jakkan.
Research Scholar: Indian Institute of Information Technology Nagpur (IIITN), Maharashtra, India.

[2]Dr. Pradnya Ghare.
Assistant Professor: Visvesvaraya National Institute of Technology (VNIT Nagpur),Maharashtra, India.

[3]Dr. Nirmal Kumar
Senior Scientist from the Division of Remote Sensing Application at the ICAR National Bureau of Soil Survey & Land Use Planning, Nagpur, India.

[4]Dr. Chandrashekhar Sakode.
Assistant Professor: Indian Institute of Information Technology Nagpur (IIITN), Maharashtra,India.

[*1]Corresponding author mail ID: d.Jakkan@gmail.com

Abstract: Carbon(C) levels have a direct impact on plant health and productivity. 200 soil samples from the Indian state of Uttar Pradesh were utilized in this study as a database to assess the efficacy of employing visible/near-infrared (VIS/NIR) spectroscopy data. The samples wavelengths ranged from 350 to 2,500 nm. The spectral features used to predict C were chosen using Ensemble Lasso Ridge Regression (ELRR), Random Forest (RF), and the more complicated Artificial Neural Network. The preprocessing employed the log derivative, Log10x derivative, and inverse derivative to replicate the wavelength of the spectrum. The essential feature wavelengths for C were discovered to be between 350 and 450 nm, according to the results. The recommended Dropout Sequential Artificial Neural Network (DrSeqANN) technique combined with the Log10x pre-processed data produced the most accurate results.

Keywords: Near Infrared Spectroscopy (NIR), Pre-processing, Log10x, Spectrum wavelength, ANN, Dropout Sequential Artificial Neural Network (DrSeqANN)

1. Introduction

Wetlands contain a large portion of the carbon reserves on the earth [1]. According to the UNEP World Conservation Monitoring Centre, wetlands encompass 6% of the earth. Wetland regions contain about 14% of the total carbon that is stored on land. Changes to the carbon stores in wetlands could have a big influence on global warming since they store a lot of carbon [2].

One of the established methods for determining the amount of Carbon(C) is dry combustion. Large-scale C detection, monitoring, and forecasting in dry environments is a challenging topic that calls for the creation of effective, quick, and precise techniques [3][4]. Although traditional methods are well known for their accuracy, they can be time-consuming to employ and risk damaging materials during processing, making it difficult to repeat laboratory results. However, recent research has demonstrated that spectroscopy in the visible infrared (VIS/NIR) region is a reliable, inexpensive, quantitative, and non-destructive tool for identifying the chemical composition of soil and its quality [5]. To ascertain the correlations between various soil components (organic phosphorus, organic carbon, and many more) and reflected spectra, scientists have created multiple empirical models [6]. Spectrum data can be utilized in a variety of ways, but machine learning techniques stand out due to their ability to analyze datasets fast and consistently [7][8].

Using traditional regression techniques such partial least squares regression or multivariate linear regression, the spectrum identification of soil with increasing C concentration was frequently successful [13][14]. Overestimation and underestimation are common issues with these techniques [15]. The results indicate that both support vector machines and random

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

forests may be able to produce accurate estimates of C. C concentrations were assessed using samples taken from China's middle and lower Yangtze River using VIS/NIR spectroscopy and SVM in [16].

This research have all shown promising outcomes when combining machine learning techniques with VIS/NIR spectroscopy data. How to evaluate C in wetlands has been the subject of several research [17][18]. Recently, a wide range of machine-learning techniques, such as interval partial least squares (iPLS) and ant colony optimization (ACO), have been presented and applied for feature selection. Inspired by previous studies and taking into account these considerations, the proposed research endeavor employs VIS/NIR spectroscopy and machine learning approaches to quantify estimates of the C content of wetlands. The study area was the Uttar Pradesh District in India, where 200 soil samples were collected at different depths and chemically analyzed. To give C contents over wetland zones, a DrSeqANN Model is employed to extract specific wavelengths from spectrum information.

2. Spectral Measurements and Pre-Processing

Figure 1 illustrates the selection of the Uttar Pradesh state in India as the research area for the collection of soil samples from the cities of Kanpur, Kanpur Dehat, Unnao, Raebareli, Amethi, Sultanpur, and Azamgarh. The relevant area is located in Uttar Pradesh, India, at a latitude of 26.536938 and a longitude of 80.489960, or 26° 32' 12.9768' N and 80° 29' 23.8560' E using GPS coordinates.

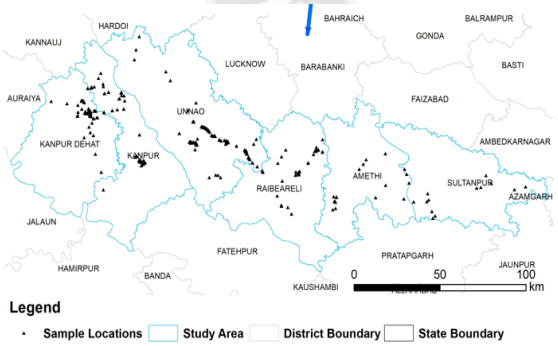


Figure 1: Study Area

Figure 1 depicts the research region. All 200 soil samples were systematically air dried, powdered, and separated by two-millimeter filter to remove any last-remaining plant remnants, roots, or stones.

200 soil samples were collected, having 2,151 characteristics between 350 and 2500 nanometers in wavelength. Spectral measurements can be affected by baseline settling, scattering anomalies, and high-frequency random disturbances. The dataset's spectral properties are enhanced [19] by utilizing Origin Pro version 9.0 [13]. In this study, we used the original spectrum of 200 soil samples and applied the First-Order Derivative (A'), inverse of First-Order Derivative (1/A'),First-Order Derivative's logarithm (lg(A')), and First-Order Derivative's log to base 10 (lg10(A')).

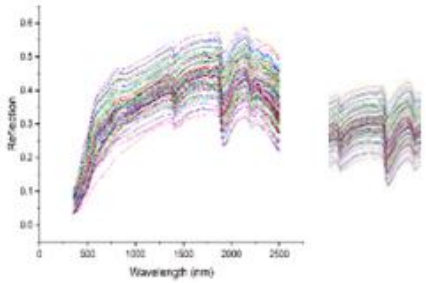


Figure 2 (a) Original Spectra

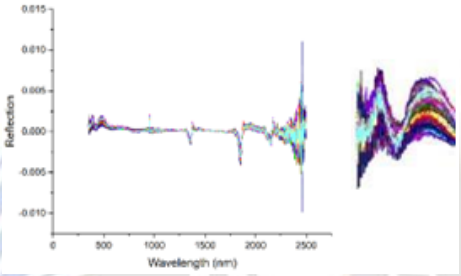


Figure 2 (b) First Derivative

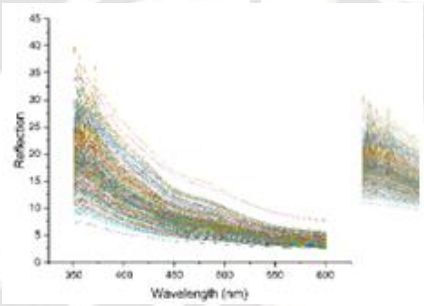


Figure 2 (c) Inverse Derivative

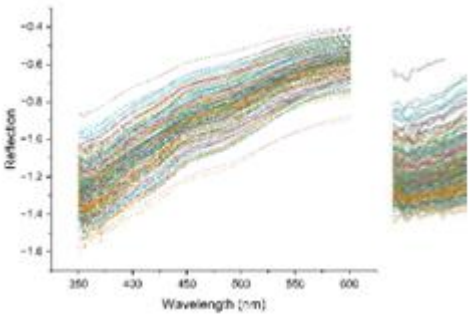


Figure 2 (d) Logarithmic derivative

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

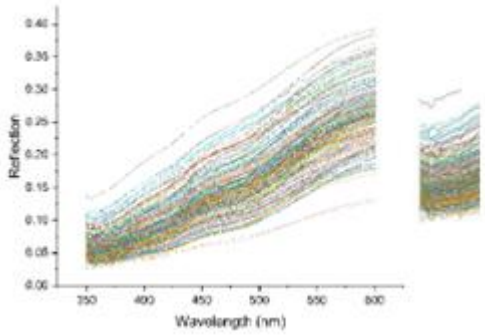


Figure 2 (e) Log to the base 10 derivatives of spectra

From Figure 2, it appears that log to the base 10 outperformed the other pre-processing methods. Absorption Peaks for Origin Spectra were at ~1450nm, ~1990nm, 2250.

2.1 first derivative (A')

The main purposes of derivatives are peak overlap resolution and resolution improvement (i.e., elimination of linear and constant baseline drift across samples). At 450 nm, 480 nm, 950 nm to 1030 nm, 1300 nm, 1800 nm, and 2300 nm to 2500 nm were the absorption peaks of the spectra. The mathematical formulas utilized for the first derivative (A) pre-processing are shown 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

2.1. Inverse of the First Derivative (1/A')

Let f(x) be a function that is both invertible and differentiable. Let y=f-1(x) be the inverse of f(x) for all x satisfying f `(f-1(x)≠0, (1/A`)

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

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

Absorption Peaks for Spectra were at ~350nm to ~430nm. Equation 2 displays the mathematical formulas used for the Inverse of First Derivative (1/A`) preprocessing.

2.2. Log Derivative (Log A')

Equation 3. demonstrates the log derivative's mathematical formulas. (Log A') pre-processing. Absorption Peaks for Spectra were at ~350nm to ~400nm

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

where as

f is the function f(x)
x is a real variable

2.4 Log to Base 10 Derivative ($Log_{10}x$)

Spectra's absorption peaks were located between ~350nm and ~450nm. Equation 4 displays the mathematical formulas used for the pre-processing of the Log of Derivative (Log A}).

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

3. Machine Learning Techniques

Two machine learning techniques i.e. Random Forest (RF) and Ensemble Lasso -Ridge Regression (ELRR) are used, as well as the proposed DrSeqANN with dropout layers to eliminate unwanted data and construct a model using VIS/NIR to predict C content.

The 200-sample dataset was collected from the Indian state of Uttar Pradesh. The model is assessed on RMSE, RPIQ, and R^2 parameters for Carbon property after applying the pre-processing. Here, the three regression models were compared and contrast using the NIR spectroscopy soil data.

3.1. Ensemble Lasso-Ridge Regression (ELRR)

In the ensemble lasso and ridge regression (ELRR), ridge regression and Lasso are brought together. L1 and L2 penalty, or automated variable selection and continuous shrinkage, are both performed by the ELRR at the same time. Two separate penalty functions make up the ELRR penalty.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

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

Where as
Ridge Regression = L1, Regularization L1 and Lasso Regression = L2
regularization parameter = λ.
total sum for the theta vector = θ_i.
how many features there are = n.

The ridge penalty (L1), the first part of the penalty, while the lasso penalty is the second (L2). There is an effort to strike a middle ground between the two penalties via the penalty parameter, which takes values from [0,1]. The advantage of the ELRR penalty is that it combines the feature selection properties of the lasso penalty with the effective regularization of the ridge regularization.

3.2. Random Forest

renowned algorithm for machine learning One subset of the supervised learning methodology is Random Forest [1]. It is applicable to both regression and classification issues in machine learning (ML) [1]. It is predicated on the idea of group education. It is a technique for combining various classifiers to solve challenging problems and improve model performance. As the name suggests, Random Forest is a regressor that enhances the dataset's predictive accuracy by using many decision trees on various subsets and averaging the results. The random forest creates forecasts based on the opinions of the preponderance of projections rather than relying solely on a single decision tree. using forecasts from each tree. The random forest model will help to solve the overfitting issue to some extent.

3.3. Artificial Neural Network (ANN)

Artificial Neural Network [15], An effective computing system called an ANN is based on biological brain networks. In describing ANNs, terms like "connectionist systems," "parallel distributed processing systems," and "artificial neural systems" were used.
Each neuron has a connector that joins it to other neurons. Every connector is as Cited with some weights. Every neuron is considered to be in an inherent state, which is characterized by activation signals. Other components may receive the output signals that are produced by combining input signals with the activation algorithm.

4. Proposed DrSeqANN Model

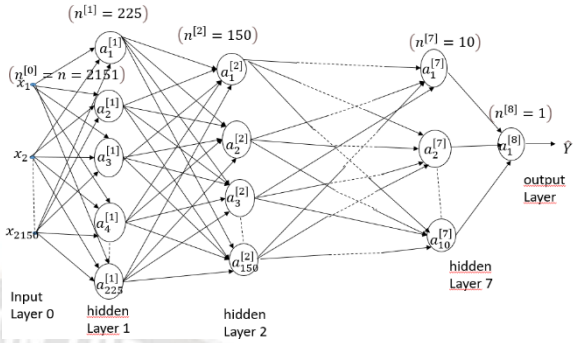


Figure 3: Neural Network Structure

According to Figure-3, seven hidden layers are used between input and output Layers. A total of 2151 spectral characteristics are employed in the input layer. A total of 225 neurons are employed in the first concealed layer. An input layer allows the input batch to enter the network. A sample feature is as Cited with each node in the input layer. A series of hidden layers stack up after the input layer until the final (output) layer. The "complex" nonlinear operations with connections are carried out by these levels. Despite being viewed as "complex," the fundamental operations are actually rather straightforward arithmetic calculations. A hidden layer is a stack of computing nodes. Each node extracts a feature from the input. The stack of output coming from a layer's nodes is called a feature map or representation. The size of the feature map, also equal to the number of nodes, is called the layer size. Intuitively, this feature map has results of various "sub-problems" solved at each node. They provide predictive information for the next layer up until the output layer to ultimately predict the response. The seventh layer is employed as the last layer using ten neurons. Tanh [20] is the activation function employed. The train test split ratio was considered as 70:30. The "ADAM" optimizer and Root Mean Squared Error (RMSE) as the loss function were used in the compilation of the model.
Neural network regularization methods like L1 and L2 weight penalties were introduced [15]. However, the overfitting problem was not entirely resolved by these regularizations.
Co-adaptation is a significant problem when learning big networks. If all the weights are learnt at once in such a network, it's common for some links to predict outcomes better than others. Here, as the network

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

is frequently taught, the weaker connections are ignored and the stronger connections learn more. Expanding the size of the neural network

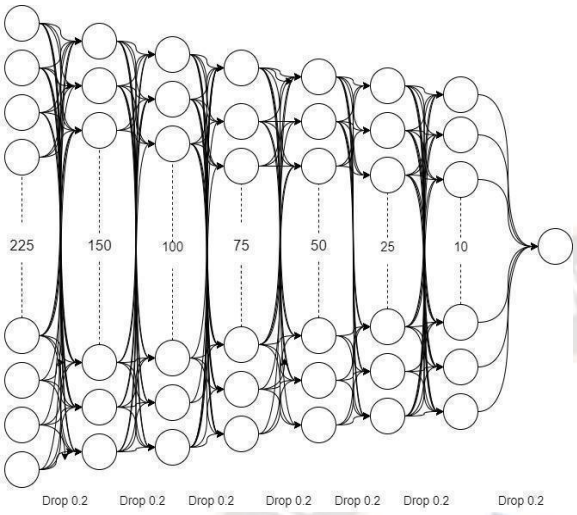


Figure 4: The Structure of Dropping the Layers in a Regularized Network

would not be advantageous. As a result, neural networks' accuracy and size were limited. Dropout followed. An innovative regularization strategy. It made the co-adaptation whole again. We could now create networks that were bigger and deeper. And make use of everything's predictive power.
As shown in Figure 4, The input to the network is a batch of samples. Each sample is a feature vector. The hidden layers in a Network are Dense. A Dense layer is characterized by a weight matrix W and bias b. They perform simple affine transformations (dot product plus bias: XW + b). The affine transforms extract features from the input. The transforms are passed through an activation function. The activations are nonlinear. Its nonlinearity enables the network to implicitly divide a complex problem into arbitrary sub-problems. The outputs of these sub-problems are put together by the network to infer the final output $\hat{y}$. Dropout changed the approach of learning weights. Instead of learning

all the network weights together, dropout trains a subset of them in a batch training iteration. With dropout, only a subset of nodes are kept active during batch learning.

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

As shown in above equation 6, there are two regulizers L_1 and L_2. regularization L1 pushes the small weights to zero, Still, there is an apparent difference¸L1 does a data-driven suppression of weights while dropout does it at random. Nevertheless, Dropout is a method of regularization. This regularization is more akin to an L2. Baldi and Sadowski (2013) provide a mathematical demonstration of this by Pierre and Peter. They proved that, under linearity (activation) assumptions, the loss function with dropout has the same form as regularization L2. The dropout rate(p) is the fraction of nodes that are dropped at a batch iteration. The regularization term in Equation 6 has a penalty factor p(1-p). The factor p(1 - p) is maximum when p = 0.5. Therefore, the dropout regularization is the largest at p = 0.5. Dropout is a regularization technique equivalent to L2 regularization under linearity assumptions. A dropout rate p = 0.5 is an ideal choice for maximum regularization. Therefore, a dropout rate of 0.5 is usually a good choice for hidden layers. Our model achieves good performance with dropout rate of 0.2. starting with first input layer number of nodes applied is 225 continued up to seventh layer number of nodes is 10, weights are automatically adjusted with the input features, bias value is 0, kernel initializer is normal, activation function is tanh.

4.1 Model Calibration Evaluation and comparison

DrSeqANN Model processes data in 5 batches with 1000 epochs. One iteration of the entire training dataset constitutes a training period. The weights are updated depending on a gradient-based optimization technique during training.
To verify the efficiency of three algorithms (DrseqANN, RF & ELRR), Out of 200 samples, 140 were used for training and 60 for testing. Ratio of performance to interquartile distance (RPIQ), Root Mean Squared Error (RMSE), and Coefficient of determination (R2) were used to gauge the algorithm's precision., which measures how well one measure performs concerning another. RPIQ considers prediction error and fluctuation of detected measures to give more consistent and objective evaluation of model validity. Better predicting ability is indicated by a higher RPIQ value. [22]. A more stable model has a higher R^2, lower RMSE and highest RPIQ.

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 $y_i = Actual\ values$
$\hat{y} = Predicted\ values$
$\bar{y} = Mean\ of\ the\ values$
$y_i - \bar{y} = Deviation\ of\ y\ from\ mean\ of\ y$

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

Root Mean Squared Error (RMSE):

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

$$y_{\propto} = Actual\ value$$
$$y_{est} = Estimated\ value$$
$$n = number\ of\ data\ points$$

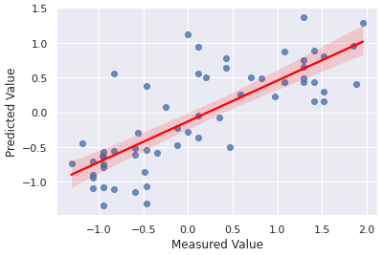
Ratio of Performance to Interquartile (RPIQ):
Quartiles: The values that divide a List of numbers into Quarters
Upper Quartile: The median of the Upper half of a set of datasets.
Range: is the difference between Greatest and Least values of the dataset
Interquartile Range: The range in the middle of the data, it is the difference between Upper Quartile and the Lower Quartile of the dataset RPIQ=IQ/RMSE or RPIQ = Q3-Q1/RMSE

5. Results and Discussions

5.1. Comparative Analysis

After comparing the three models, it appears that the DrSeqANN model achieves an RPIQ of 8.42, RF model RPIQ of 8.25, and ELRR model of 8.11. When applying the ELRR model to raw data (before Preprocessing), it yields the greatest RPIQ value. The values of the coefficient of determination, mean squared error, and root means square error for each model were nearly identical. The scatter plot of all three models on the original data is displayed in the figures 5 to 9.
The first step is to pre-process the data using Derivative functions. Pre-processing is finished by using the first derivate, Inverse Derivative, Log Derivate, and $Log_{10}x$. This makes it easier to assess if these samples are present in the training sets that were used to create the prediction models.

(a)

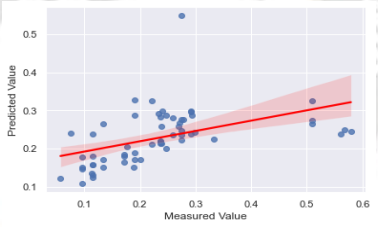


(b)

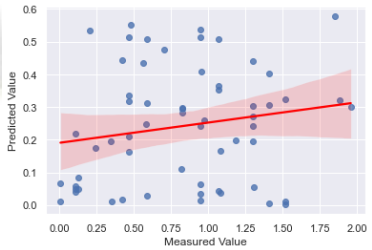
(c)

Figure 5: The scatter plot of original Raw data (before Preprocessing) (a) DrSeqANN, (b) RF, (C) ELRR

Figure 5 shows the scatter plot for all three models before Preprocessing.
From Figure 6, it is analyzed that the DrSeqANN model is giving better results on $Log_{10}x$ pre-processing data. The data points are less scattered in the DrSeqANN model compared to RF and ELRR models. The R^2 value on the test data set was obtained as 0.82 with the RMSE value as 0.08 and RPIQ as 4.32.

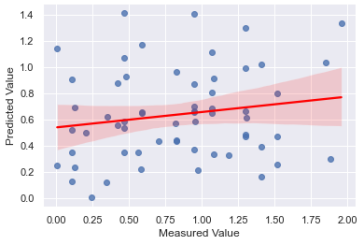


(a)



(b)

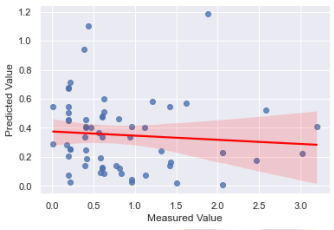
International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*



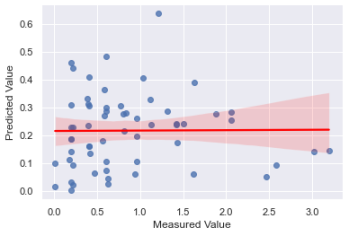
(c)

Figure. 6: Scatter plot of on $Log_{10}x$ pre-processing data (a) DrSeqANN, (b) RF, (c) ELRR

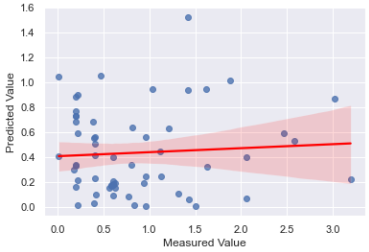
Figure 7 shows the scatter plot for all three models on the inverse derivative pre-processing data for regression analysis. From Figure 7, it is analyzed that the RF model is giving better results on inverse pre-processing data having R^2 measures of about 0.55, RMSE of 0.07 on test dataset. Data is less scattered in RF model compared to DrSeqANN and ELRR Models which is also reflected in the scatter plot shown in figure 7(b).



(a)



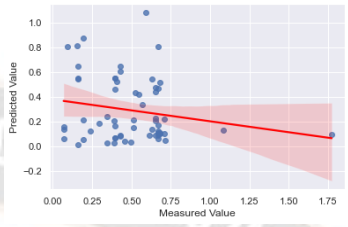
(b)



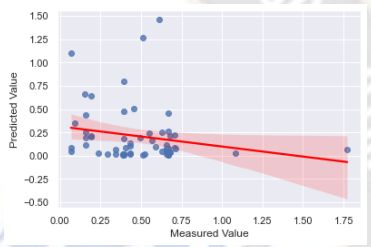
(c)

Figure 7: The scatter plot on the inverse derivative pre-processing data (a) DrSeqANN, (b) RF, (c) ELRR

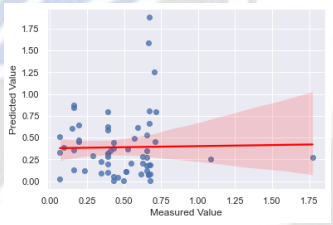
Figure 8 represents the scatter plot for all three models on an inverse derivative pre-processing dataset. The evaluation parameter values (RMSE, RPIQ and R^2) are very close to each other and there is a very marginal difference between these values. From figure 8 a, b & c it appears that the DrSeqANN model is performing slightly better with RMSE of 0.08 and R^2 measure of around 0.69.



(a)



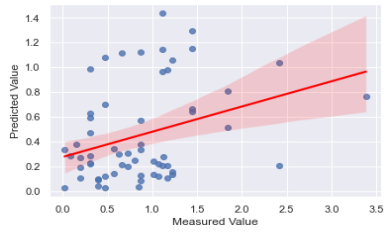
(b)



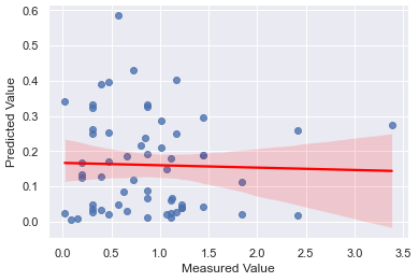
(c)

Figure 8: The scatter plot on an inverse derivative pre-processing(a)DrSeq ANN (b) RF (c) ELRR

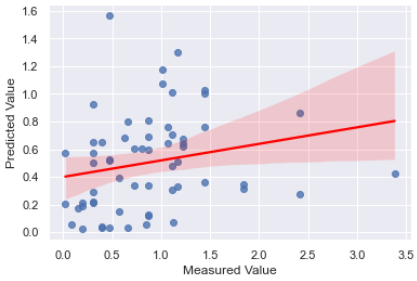
From figure 9 of the scatter plot for all three models, it is evident that the DrSeqANN model outperformed RF and ELRR model. The training set's R^2 value is 0.98, whereas the testing set is 0.67. In the case of the logarithmic derivative, the RPIQ is likewise the highest value as 3.87.



(a)



(b)



(c)

Figure 9: The scatter plot on an logarithmic derivative pre-processing (a) DrSeqANN, (b) RF (c) ELRR

The DrSeq-ANN model's RMSE performance was enhanced by the application of $Log_{10}x$ pre-processing. The R^2 value is good for Random Forest Model compared to the other two algorithms on original data. After applying the pre-processing techniques as discussed in earlier sections, it is observed that $Log_{10}x$ pre-processing gives the best results. The RMSE is 0.08, R^2 is 0.82 and RPIQ is 4.32 for our proposed DrSeqANN model. In contrast to the remaining two methods, the suggested DrSeqANN prototype performs well in the majority of the pre-processed data. The RPIQ value we found is always greater compared to RF and ELRR.

6. Conclusion

The regression machine learning and proposed DrSeq-ANN models were demonstrated for predicting C contents, utilizing Inverse Derivative, First Derivative, and Logarithmic Derivative, as well as $Log_{10}x$. The suggested DrSeq-ANN model outperformed the Random Forest and ELRR models for soil properties (C) on the presented dataset. In particular, applying a $Log_{10}x$ during pre-processing greatly enhanced the model's accuracy for R^2 by 17.55% when compared with previous models in the literature Future research can be carried out for other types (such as salt, silt and sand) of soil samples using DrSeqANN Model.

Acknowledgement:

The authors would like to thank Dr. Nirmal Kumar, Senior Scientist at the ICAR-National Institute of Soil Survey & Land Use Planning, Nagpur, for his ongoing technical assistance in accessing and maintaining the spectrum library. He works in the Department of remote sensing.

Data Availability:

The State of Uttar Pradesh, India's C data is available upon request and has all the information required to build prediction models.

Funding:

The Visvesvaraya Ph.D. Scheme for Electronics and IT Ministry of Electronics and IT (MeitY), Indian Government is funding this research in Phase II (Unique Awardee Number: MEITY-PHD-3080). Contributors of money have no influence over the preparation of manuscripts, study designs, data collecting and analysis, or publishing decisions.

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 measure-

ment 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.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

Table 1: Statistics of Soil Properties by Parameters of Mean, standard Deviation, First & Third Quartile and Maximum

Soil Property	**Wave-length**	**Count**	**Mean**	**Std Deviation**	**Mini-mum**	**Ist Quar-tile**	**IIIrd Quartile**	**Maximum**
Ph Extract	350	199.0	0.059554	0.017678	0.026468	0.046662	0.067182	0.138930
Ec Extract	351	199.0	0.060029	0.017390	0.024490	0.048780	0.068644	0.134328
CaCO3 Equiva-lent %	2499	199.0	0.356908	0.062670	0.225497	0.309888	0.393054	0.520923
C	2500	199.0	0.356783	0.062924	0.225079	0.308125	0.393934	0.524133