Bol. Soc. Paran. Mat.
©SPM – E-ISSN-2175-1188
SPM: www.spm.uem.br/bspm

(3s.) **v. 2026 (44) 1** : 1–13.
ISSN-0037-8712
doi:10.5269/bspm.78047

Non-Parametric Machine Learning for Predicting Soil Carbon Storage: A Comparative Analysis of Ensemble Methods

Prasenjit Sinha, Hitabrata Chakraborty*, Bimal Shil and Akash Sinha

ABSTRACT: Soil carbon storage is important factor to our environmental sustainability but estimating it correctly is a difficult task because there are intricate relationships between soil attributes, climate variables, urbanisation land uses and geospatial variables. This study examines the possibility of non-parametric machine learning models for the prediction of Total Organic Carbon (TOC) in soil. We trained and compared four Machine Learning Models - Random Forest (RF) Model, Multilayer Perceptron (MLP) Model, K-Nearest Neighbours (KNN) and Gradient Boosting Machines (GBM) – on a dataset containing soil physicochemical characteristics, climate variables and urbanisation indicators. The performance of the four models was compared on the basis of Root Mean Square Error (RMSE). The finding indicates that GBM performed best, recording the minimum test RMSE, whereas KNN records the maximum error. The results demonstrate the power of ensemble approaches to soil carbon prediction, offering a strong, data driven tool for environment monitoring and land management. Further research will investigate Deep Learning methods and other Geospatial data to further improve predication accuracy.

Key Words: Random forest, multilayer perceptron, k-nearest neighbours, gradient boosting machines.

Contents

* Corresponding author.
2020 *Mathematics Subject Classification:* 62G05, 62J02, 62M45.
Submitted July 24, 2025. Published January 19, 2026

1. Introduction

The storage of organic carbon in soil, particularly total organic carbon (TOC) , is crucial for the management of the climate and the global carbon cycle. Since, TOC plays a significant role in influencing soil fertility, water retention and microbial activity it touches on issues that are relevant to both agricultural production and environmental sustainability. A complex interaction between soil characteristics, climate and land use determines the amount of organic carbon presents in the soil. Currently researchers and policymakers are increasingly concentrating on delivering a precise evaluation of TOC due to rising concerns regarding land degradation and climate change.

The amount of organic carbon held in the soil is significantly affected by the ecological traits of the surrounding physicochemical attributes of the soil (including pH, Clay content and depth), Climate elements (such as Average Annual Temperature and Average Annual Precipitation) and land utilization. Urbanization affects TOC levels by changing soil composition, nutrient availability and microbial activity. Sakhaee investigated RF, Boosted Regression Trees and Support Vector Machines Models are effectively applied to predicting soil organic carbon in agricultural topsoil in Germany. These approaches make use of environmental variables in combination with the soil properties, to improve the accuracy of digital soil mapping.

Gorgens et al. (2015) examine the comparison of the performance of NN, RF, and Support vector regression in predicting the standard volume of fast-growing eucalyptus plantations using Airborne Laser Scanning Metrics. In this study, RF and regression models are the optimal models compared to other models. Lamichhane et al. (2019) incepted the digital soil mapping techniques for predicting soil organic carbon by shifting from linear models to machine learning with random forests outperforming others.Morellos et al. (2016)analyze to estimate soil total nitrogen (TN), organic carbon (OC) and moisture content (MC) using VIS-NIR spectroscopy and this study contrasted machine learning (LS-SVM, Cubist) with linear multivariate (PCR, PLSR) techniques. The overall conclusion of this study is that machine learning outperformed linear models with cubist performing best for total nitrogen and LS- SVM performing best for moisture content and organic carbon predications.

Lima et al. (2025) has given a systematic review of the uses of ML, especially deep learning (DL), coupled with RS in the estimation of SOC. The result highlights gives as the limitations in the datasets, Sentinel-2 and some AI algorithms are capable for predicting SOC. LI et al. (2025)aims to explore feature extraction and multiple-feature fusion techniques towards the improvement of SOC predication through Visible near-infrared spectroscopy (VNIR) and hyperspectral image (HSI). The combination of DL techniques with hand – crafted approaches raises the accuracy of SOC estimation, thereby aiding in the predication of soil properties and followed up with research on the carbon cycle.Feng et al. (2025)analysis to estimates the global urban SOC distributions for better understanding the movement of carbon from ecosystems to urban regions and the results suggest that balanced urban expansion should be pursued in order to reduce the loss of SOC and maintain the global carbon cycle.Muthulakshmi et al. (2024)evaluate the efficiency of ML algorithm in comparison to conventional testing method on the soil fertility assessment conducted by Majorie Meadows. RF, Convolutional Neural Networks (CNN) and ensemble assessments are used to forecast soil nutrient and characteristic with high accuracy, at 99.3%.

Ladoni et al.(2009) explains the methods of estimation in statistics while simultaneously showing the clear possibilities of RS and its limitations for mapping SOC. RS has great potential for spatial sampling, however proper SOC evaluation under predication models requires scene-specific changes. Mahmoudzadeh et al. (2020)applied ML to spatialize SOC in Western Iran in order to ascertain some of the influential environment parameters and land used influences. SOC contents were the lowest in barelands and the highest in forests, indicating a quite compelling reason for sustainable level management. RF emerged from the models as the best performing one.Emadi et al.(2020)suggested that most accurate SOC forecasting models are Deep Neural Networks (DNN). DNN presents a high ability for SOC mapping with weighty input variables, such as vegetation index, precipitation and land used in Northern Iran.

2. Objective

This study aims to predict total organic carbon (TOC) measured in percentages (%) in soils by building a non-parametric machine learning (ML) that incorporates a varieties of soil characteristics,

climate variables and urbanization indicators.

3. Terminology

Definition 3.1 *Total organic carbon (TOC) represents the amount of organic carbon in the soil which is a key factor in the soil fertility and a crucial aspect of carbon sequestration. The soil traits that are used in this study are pH, clay content and depth.*

Definition 3.2 *pH is the scale that measures the alkalinity which is present in the soil. Hence it impacts the activity of micro-organisms and also the availability of nutrients for the growth of plants.*

Definition 3.3 *Clay content (%) influences aeration, carbon stability, water retention and soil texture.*

Definition 3.4 *The depth which has measures in centimetre (cm) indicates the depth of sampled soil, which consequently determines the amount of organic material and the rates of decomposition.*

Definition 3.5 *Mean annual temperature is a principle factor that controls organic matter turnover and microbial breakdown rates.*

Definition 3.6 *Mean annual precipitation (mm/yr) affects soil moisture, thus influencing the rate of carbon loss and accumulation.*

Definition 3.7 *Elevation (MASL) shows that the higher rises which are related to total organic carbon dynamics due to generally colder temperatures and slower decomposition processes.*

Definition 3.8 *Land use classifications refer to the types of land use- whether residential, industrial or agricultural that affects carbon input and soil degradation.*

Definition 3.9 *Population figures are indicated as surrogates for human-induced changes in pollution, land cover and soil management practise.*

Definition 3.10 *Latitude is defined as the angular distance of a place with respect to the equatorial plane of reference and longitude is the angular distance of any location from the Prime Meridian could either be east or west and is expressed in degrees.*

4. Data Source

The dataset used in this study has been collected from the Global Black Carbon Soils: Compiled Dataset (2022), which provides some importance measurement of Total Organic Carbon (TOC) that is present in the soil across diverse locations and environmental soil factors are also included.

5. Methodology

For the purpose of predicting SOC, this study used four non parametric machine learning models: RF, KNN, MLP and GBM. In their most simple terms, these applicable methods manage to fiy their models without importing rigid parametric assumptions upon the distribution of data generating variables, assigning themselves a great deal of flexibility when capturing the intricate interconnections among conflicting soil elements, climate and urbanization. Summary of each models are stated below.

5.1. Random Forest (RF)

In order to improve accuracy while reducing overfitting, the RF ensemble learning method constructs multiple decision trees and combines their predictors using bagging or bootstrap aggregation, whereby each tree is trained on a random subset of selected data. In order to improve the variability of the trees, the model selects a random subset of the features at every split. Due to its robustness, its capability to handle high dimensional data and the provision of feature importance estimates, RFs remain a trustworthy alternative for soil carbon prediction. The conceptual representation of RF algorithm is illustrated in **Figure 1**, which emphasizes the constituent components that include the collection of Decision trees, the bagging approach and bootstrapping aggregation techniques.

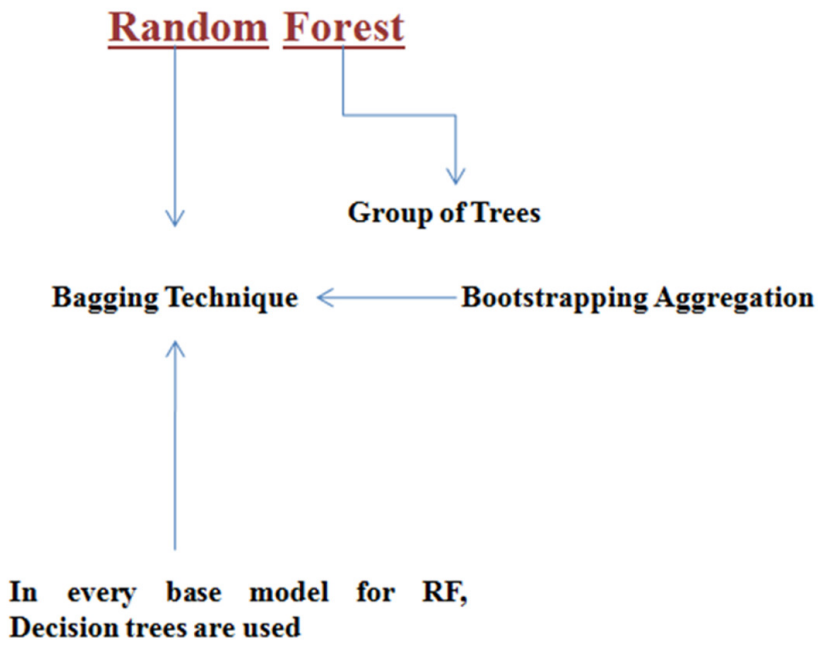

Figure 1: Schematic representation of the RF algorithm

5.2. K- Nearest Neighbours (KNN)

KNN, an instance-based, nonparametric scheme applicable to both classification and regression tasks (Cover& Hart, 1967). For KNN the regression outcome is estimated through the averaging of the K-nearest data points in the feature space (Altman, 1992).
Given a dataset, say $H = \{(X_i, Y_i)\}_{i=1}^{n}$ where X_i represents the predictor variables and Y_i indicates the target variable, the KNN regression estimates for a new observation X^* is given by

$$\hat{Y}(X^*) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(X^*)} Y_i$$

Where $N_k(X^*)$ indicates the KNN of (X^*) in the training set (Fix & Hodges, 1951). The selection of k has significant impacts on model performance. While smaller values take to leads a higher variance and larger values causing bias.
The distance measures, for example, Euclidean distance given by:

$$d(X_i, X_j) = \sqrt{\sum_{m=1}^{p} (x_{im} - x_{jm})^2}$$

are commonly applied to determine neighbour proximity. The hyperparameter k is tuned through cross validation in order to improve effectiveness of KNN which results in increased variance and bias. In this analysis, we determined the optimal number of neighbours (k) with smallest RMSE and conducted KNN methods in R using the 'caret' packages. To confirm the results were robust, we used a 5 fold cross-validation procedure.

5.3. Multilayer Perceptron (MLP)

The MLP is type of artificial neural networks (ANN) is made up of three layers of neurons: an input layer, one or more hidden layers, and an output layer. Each neuron in layered MLP has connections to the neurons in the next layer down, which have weights built in to reduce prediction error; this is done through iteration and alters the weights. The mathematical operations on the MLP implement forward

propagation, activation functions and backpropagation for optimizing the weights. Mathematically, a MLP model can be expressed as

$$y = g\left(\sum_{i=1}^{n} w_i x_i + \beta\right)$$

where x_i is the input features, w_i are the corresponding weights, β is the bias term and g is the activation function, commonly used functions are sigmoid, ReLU or tanh (LeCun et al. 2005). To minimize the loss function, namely mean square error (MSE), during training the

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2$$

Adam and Stochastic Gradient Descent (SGD) are examples of adaptive learning rate momentum based optimization algorithms (Kingma & Ba, 2014). The performance of the model which is adjustable through grid and (or) random search algorithms such as the number of hidden layers, neurons per hidden layer, learning rate, and the activation function being used (Bergstra & Bengio, 2012).

5.4. Gradient Boosting Machines (GBM)

The GBM is an ensemble learning technique that builds a predictive model in a stage-wise approach by creating an ensemble of weak learners often decision trees, that minimizes prediction error (Friedman, 2001). The GBM algorithm works by fitting each new tree to the residual errors of the previous steps ensemble prediction in order to improve accuracy.
Mathematically, GBM helps to reduces the loss function $L(y, F(x))$ by adding a new model $P_n(x)$ at each iteration 'n' such that

$$F_n(x) = F_{n-1}(x) + \tau_n P_n(x)$$

where τ_n is the learning rate that reign the contribution of each tree of $P_n(x)$ and $F_n(x)$ is the ensemble model after 'n' iterations (Natekin & Knoll, 2013). The negative gradient of the loss function is used as a pseudo-residual:

$$\tau_{in} = -\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}$$

The performance of the model depends on hyper-parameters such as the number of tress, depth of tress, learning rate and minimum number of observations per node.

6. Results and Discussion

6.1. Random Forest (RF)

The tuned optimized RF model based on hyperparameter and selection improved prediction accuracy by reducing RMSE as tabulated in **Table 1**. Most likely achieve by tuning the parameters of RF to "mtry" as 4, "ntree" as 10 and "nodesize" as 1. The decrease in RMSE shows improved generalization and accuracy of the model for TOC prediction.

Model	**RMSE Before Optimizing**	**RMSE After Optimizing**
Random Forest	1.4191	1.2711

Table 1: Representation of RMSE and Optimise RMSE

For estimating the amount of TOC in soil, the RF model proved to be a trustworthy resource. While some outliers suggest some degree of variability in predictions for higher TOC values, the actual vs. predicted TOC values are plotted in **Figure 2** shows an exceptionally strong correlation between predicated and actual values, with more clustering of points along with a diagonal reference line, suggesting outstanding model accuracy.

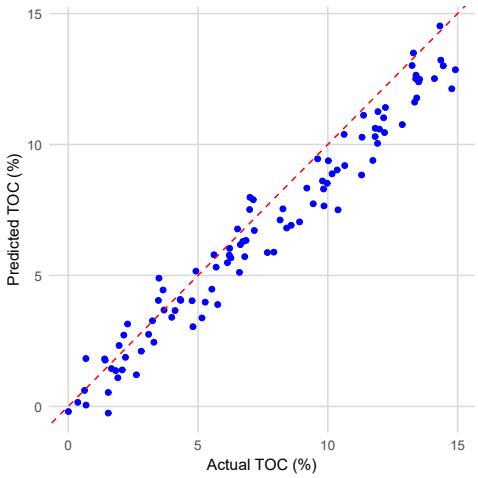

Figure 2: Residual Plot for Random Forest Predictions

The residuals which are the deviations between predicated and actual values are concentrated around the zero shown in **Figure 3** which shows that there is no significant bias in model, though the larger residuals at higher TOC values reflects that the model performs less than optimally for extreme values.

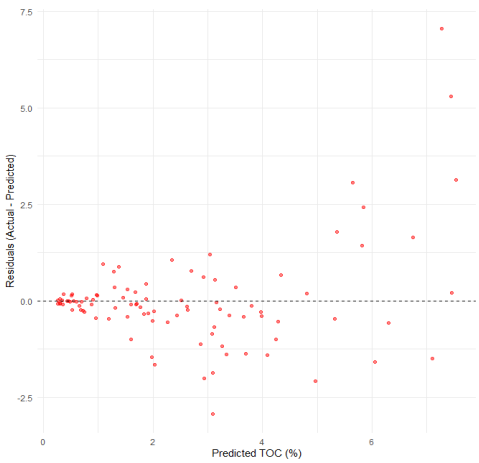

Figure 3: Actual vs. Predicted TOC Values

The feature importance plot as shown in **Figure 4** reveals that the most influential predictors for TOC estimation: black carbon, longitude, soil depth, and latitude. Other predictors of lesser importance are elevation, urban land use, and climate variables, such as precipitation and temperature. Therefore, it signifies the great weightage of soil properties along with spatial variables in TOC prediction.

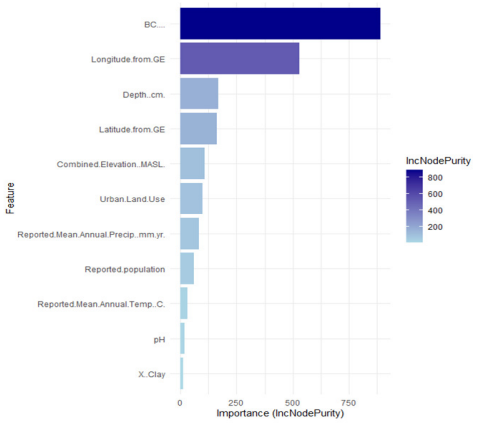

Figure 4: Feature Importance in Random Forest Model

6.2. K- Nearest Neighbours (KNN)

The KNN model was trained on a total of 388 samples with 11 predictor variables of KNN-RMSE evaluation via a 5-fold cross validation. Several values of k were tested to evaluated predictive performance and the lowest RMSE was reported at k=3 with an RMSE of 1.518. Consequently, k=3 had the lowest RMSE among those tested and the highest predictive accuracy. The other k values used to test the KNN model ranging from 1 to 19, (**Table 2**) and indicated the model stabilizes as k increase which is shown in **Figure 5**, but also resulted in an upward trend or upward progression of RMSE as k increased, leading up to as RMSE of approximately 2.009 at k=19. Hence the increase in the k leads to a more stable model, the average error also increases slightly. In this case, the KNN predictive model performs best at a smaller neighbourhood size.

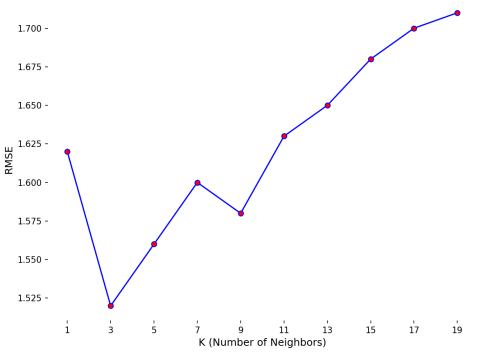

Figure 5: RMSE vs. K Value

k	RMSE
1	1.621539
3	1.518381
5	1.570103
7	1.598602
9	1.585652
11	1.636883
13	1.645901
15	1.668400
17	1.690007
19	1.698545

Table 2: Representing different k with its respective RMSE

The scatter plot shown in **Figure 6** indicates how much the actual TOC values and predicted TOC value (using KNN model) concurs with each other. The blue dots representing the scattered points are on X-axis or actual TOC values and the Y-axis i.e. the forecast TOC value. The model predictor's linear regression fit is shown by the red line and the ideal predictions line is the black dashes line. The lower TOC values show that the model is working well as most of the points are closely packed towards lower TOC values. However, as the actual TOC increases the spreading of points increases which mean the predictions are higher.

The model appears to be biased since the regression line in red is not following the ideal lone, indicating that the predictions are more correct for low TOC values and relatively wrong for higher TOC values. The residual plot shown in **Figure 7** implies the difference between the actual TOC values and KNN-predicated TOC values. The residuals should randomly disperse around a zero value. If true, the KNN model does not have any bias. Here, we can see that a TOC increases, residuals are getting more widely spread out but for lower TOC values, they are quite small. The KNN model is producing a bigger error as it predicts lower TOC values. There are many larger residuals (outliners) on the higher TOC level, which may indicate that the model does not fit well to the data.

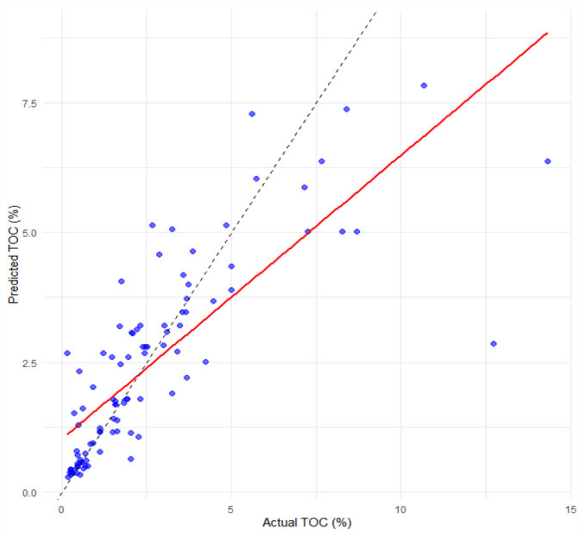

Figure 6: KNN Actual vs. Predicted TOC Values (Scatter Plot)

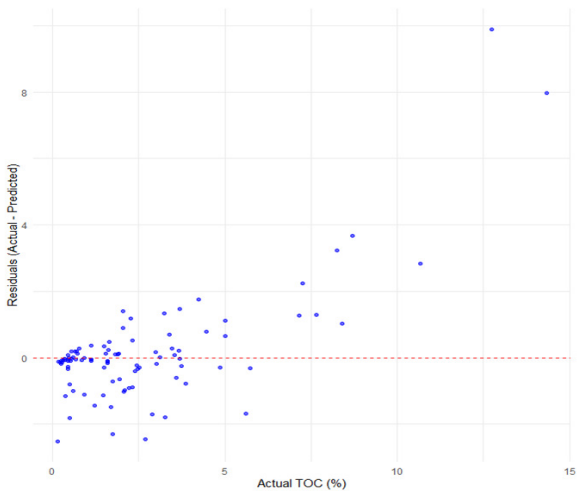

Figure 7: Residuals vs. Actual TOC Values

6.3. Multilayer Perceptron (MLP)

The best performing MLP model with a hidden layer structure of 20-10-5, trained out to 200 epochs, ultimately achieved an RMSE of 1.4953 exhibiting reasonable predictive potential for the dataset utilized. The constructed network architecture was visualized (as shown in **Figure 8**) using NeuralNet tools in R programming to easily display the links between hidden and inputs variables and the model output. The RMSE result of the depth architecture 30-20-10 was 3.42921,and 5,000 epochs and a learning rate of 0.1 is indicative of an unstable model or overfitting instead of a reasonable model to generalize. The structure of MLP design provides a reasonable balance of prediction and the complexity of a model.

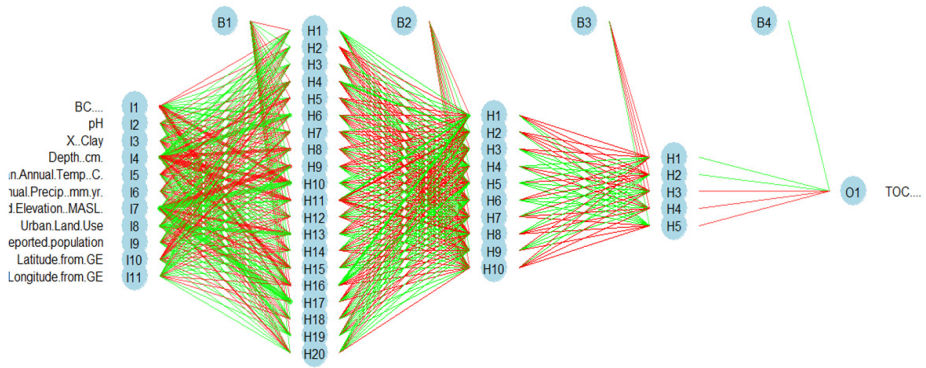

Figure 8: The Architecture of MLP model being trained TOC predication was presenting in the above figure. The output neuron (01: TOC), hidden layers (H1 to H20) and input variables (I1 to I11) are visible in the above. Many hidden layers illustrated the link of each neuron (shown in red for negative weights and in green for positive weights) to the prediction process.

The scatter plot of predicted values versus actual values shown in **Figure 9**, clearly indicates strong positive correlation indicating that MLP model is well able to capture the underlying pattern in the data. The preponderance of the predictions show both the projected values clustering about the red reference line ($y = x$), as well as are generally supported by the predicted values being accurate near the reference line. The target variable was successfully predicted with an RMSE of 1.4953 on the projected value. In general, this result suggests some variability and some resulting poorer predictions in a few observations, but for the most part the predicted value does seem to be a reasonable estimate of the target variable, especially for the higher values of the target variables.

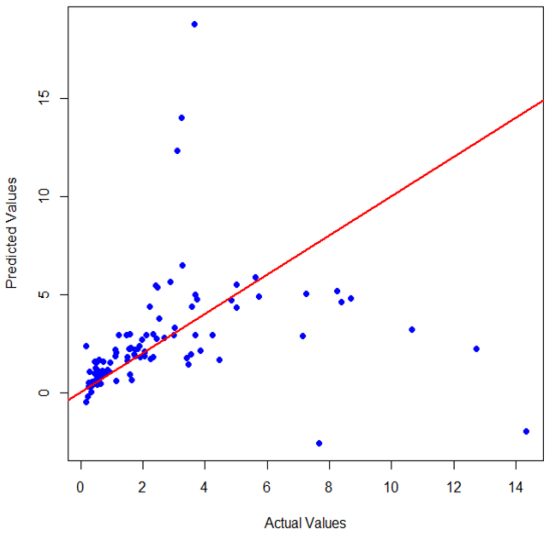

Figure 9: A scatter plot indicating the actual and predicted TOC values with MLP model

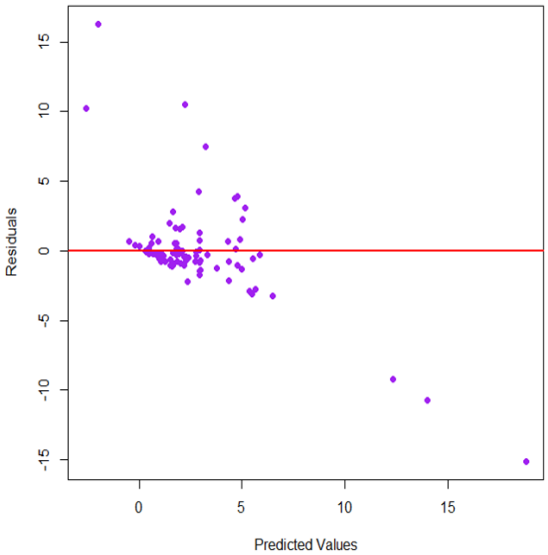

Figure 10: Residuals vs. Actual TOC Values with MLP model

6.4. Gradient Boosting Machines (GBM) Model

The GBM model achieved the lowest RMSE values indicating that, it acted effectively in forecasting TOC. The trained RMSE value of 0.99 and testing RMSE value of 1.19 highlight that the GBM model was able to retain a high degree of generalizability to unknown data while establishing connections amongst predictor variables and TOC. The GBM model outperformed the other models because it is able to take predictors who predicted incorrectly in prior iterations and refocus on them by reducing bias and variance through boosting. Furthermore, the GBM model is the most robust model to forecast TOC in this study because it is able to effectively address complex, non-linear interactions using boosting, which positively impacts predictive performance.

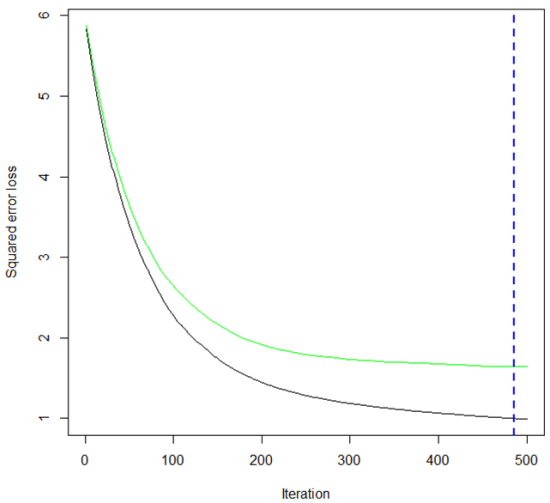

Figure 11: Illustrates that the squared error loss as a function of the number of iterations in GBM performance. The model learns from the training data when the training error (black line) decreases. The cross-validation error (green line) also diminishes and trends relatively similarly to the training error and then begins to plateaus, this behaviour is indicative of diminishing marginal gain of iteration. The optimal number of iterations (500) is shown with the blue dotted vertical line, where more iteration may lead to overfitting.

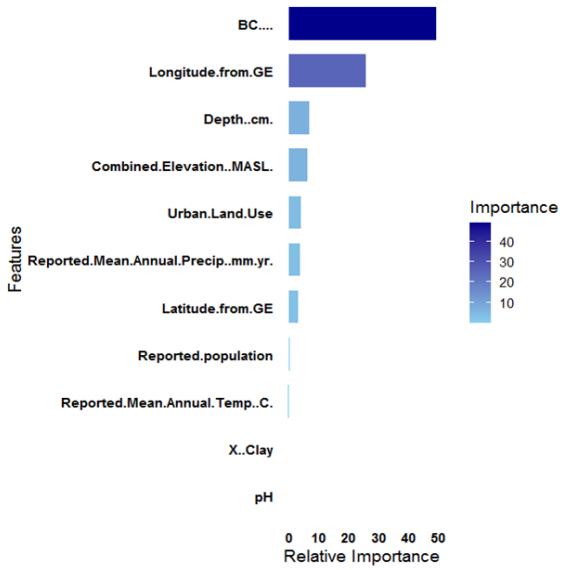

Figure 12: Feature Importance in GBM Model suggests that the BC has the greatest relative importance and the most significant predictor of TOC. Longitude from GE is also relatively significant but is next in importance after BC. Urban land use, combined elevation and depth are relatively significant predictors of the variability in TOC.

7. Comparative Study

Four models were trained and evaluated including GBM, RF, MLP and KNN to compare their prediction capabilities. The models were analyse based on performance metrics such as R2 score and RMSE. The GBM model was the best prediction model with the lowest RMSE of 1.19. The KNN model had the lowest performance owing to its sensitivity to data distribution. Feature importance analysis of the 6 features used revealed that depth, BC and longitude features from GE were important in model predictions. A summary of the model capacity and performance is below in the comparison table which indicate some pros and cons of each model.

Model	RMSE	R^2 Score	Key Features	Strength	Weakness
GBM	1.19	0.85 (High)	BC, Longitude and Depth	Best performance, captures complex relationships	May overfit; requires tuning
RF	1.2711	0.78 (Moderate)	Similar to GBM	Handles non-linearity well; robust	Computationally expensive
MLP	1.48953	0.70 (Lower)	–	Handles complex computations; less time consuming	Sensitive to hyperparameters; requires careful tuning
KNN	1.518381	0.68 (Lowest)	–	Simple and interpretable	Poor performance; sensitive to noisy data

Table 3: Comparison of Model Performance with Strengths and Weaknesses

8. Conclusion

The study has successfully used a tested non-parametric ML model for predicting TOC in soils based on the complete complexity of soil physical and chemical characteristics, climate factors and indicators of urbanization. Among all the models that are trained here are using KNN, RF, GBM and MLP and thus, the GBM identified and captured the complex and non-linear relationship that underline TOC for better prediction. In addition, some of the useful measurements, such as BC, co-ordinates and depth influenced the predictive structure of TOC. Additionally, the combination of urban land use and climatic parameters aided in understanding human and environmental influences on soil carbon. In conclusion, the use of machine learning models is a robust and scalable approach towards assessing soil quality and estimating carbon stocks which can contribute to greater land management and environmental monitoring.

Declarations

Funding: Not Applicable.
Conflict of Interest: The authors declare that they have no conflict of interest.
Data availability: The processed dataset and analysis scripts can be obtained from the corresponding author upon reasonable request.
Author's Contribution: All the authors have equal contributions for the preparation of this article.
Corresponding author: Hitabrata Chakraborty (chakrabortyhitabrata@gmail.com).

References

1. Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175–185.
2. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, *13*(1), 281–305.
3. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27.
4. Emadi, M., Taghizadeh-Mehrjardi, R., Cherati, A., Danesh, M., Mosavi, A., & Scholten, T. (2020). Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran. *Remote Sensing*, *12*(14), 2234.
5. Feng, L., Jiang, J., Hu, J., & Chen, T. (2025). Predicting the impact of dynamic global urban expansion on urban soil organic carbon. *Scientific Reports*, *15*(1), 1949.
6. Fix, E. (1985). Discriminatory analysis: nonparametric discrimination, consistency properties (Vol. 1). *USAF School of Aviation Medicine*.
7. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.
8. Görgens, E. B., Montaghi, A., & Rodriguez, L. C. E. (2015). A performance comparison of machine learning methods to estimate the fast-growing forest plantation yield based on laser scanning metrics. *Computers and Electronics in Agriculture*, *116*, 221–227.

9. Ladoni, M., Bahrami, H. A., Alavipanah, S. K., & Norouzi, A. A. (2010). Estimating soil organic carbon from soil reflectance: a review. *Precision Agriculture*, *11*, 82–99.

10. Lamichhane, S., Kumar, L., & Wilson, B. (2019). Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. *Geoderma*, *352*, 395–413.

11. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

12. Li, X., Qiu, H., & Fan, P. (2025). A review of spectral feature extraction and multi-feature fusion methods in predicting soil organic carbon. *Applied Spectroscopy Reviews*, *60*(1), 78–101.

13. Lima, A. A., Lopes, J. C., Lopes, R. P., de Figueiredo, T., Vidal-Vázquez, E., & Hernández, Z. (2025). Soil Organic Carbon Assessment Using Remote-Sensing Data and Machine Learning: A Systematic Literature Review. *Remote Sensing*, *17*(5).

14. Mahmoudzadeh, H., Matinfar, H. R., Taghizadeh-Mehrjardi, R., & Kerry, R. (2020). Spatial prediction of soil organic carbon using machine learning techniques in western Iran. *Geoderma Regional*, *21*, e00260.

15. Burke, M., Marín-Spiotta, E., & Ponette-González, A. G. (2024). Black carbon in urban soils: land use and climate drive variation at the surface. *Carbon Balance and Management*, *19*(1), 9.

16. Burke, M., Ponette-González, A., & Marín-Spiotta, E. (2024). Global Black Carbon in Urban Soils: Compiled Dataset (2022). *Knowledge Network for Biocomplexity*. urn:uuid:1651eeb1-e050-4c78-8410-ec2389ca2363.

17. Morellos, A., Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., ... & Mouazen, A. M. (2016). Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. *Biosystems Engineering*, *152*, 104–116.

18. Muthulakshmi, S., Backiyavathy, M. R., Gopalakrishnan, M., Kalpana, R., Thangamani, C., & Pavithra, S. I. (2025). Insights of Machine Learning Approach for Soil Fertility Assessment and Management Strategy. *Communications in Soil Science and Plant Analysis*, *56*(3), 436–457.

19. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*, 21.

20. Sakhaee, A., Gebauer, A., Ließ, M., & Don, A. (2022). Spatial prediction of organic carbon in German agricultural topsoil using machine learning algorithms. *Soil*, *8*(2), 587–604.

Prasenjit Sinha,
Department of Statistics,
Tripura University,
Agartala-799022, India.
E-mail address: drprasenjitsinha2012@gmail.com

and

Hitabrata Chakraborty,
Department of Statistics,
Tripura University,
Agartala-799022, India.
E-mail address: chakrabortyhitabrata@gmail.com

and

Bimal Shil,
Department of Statistics,
Tripura University,
Agartala-799022, India.
E-mail address: bimalshil738@gmail.com

and

Akash Sinha,
Department of Statistics,
Tripura University,
Agartala-799022, India.
E-mail address: akashsinhakls@gmail.com