Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Prediction of Soil Organic Carbon Using Machine Learning Algorithm in Long-Term Ecological Monitoring Plots of Southern Dry Mixed Deciduous Forests, Telangana, India

Ruby Patel

patelr@icfre.org

Institute of Forest Biodiversity https://orcid.org/0009-0007-0628-2554

Bharati Patel

Sravanthi Palle

Jaison M

Research Article

Keywords: Soil organic carbon. Southern dry mixed deciduous forest. Prediction. Machine learning algorithm. Soil moisture. Soil nutrients

Posted Date: August 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10665919/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Prediction of Soil Organic Carbon Using Machine Learning Algorithm in Long-Term Ecological Monitoring Plots of Southern Dry Mixed Deciduous Forests, Telangana, India

Ruby Patel[a,b*], Bharati Patel[b], Sravanthi Palle[b], and Jaison M[c]

[a*] Silviculture and Forest Management Division, ICFRE-Institute of Wood Science and Technology, Malleswaram, Bengaluru- 560080, Karnataka, India.

[b] Forest Ecology and Climate Change Division, ICFRE-Institute of Forest Biodiversity, Dulapally, Komaplly(SPO), Hyderabad-500100, Telangana, India.

[c] Soil Science and Agriculture Chemistry, Bidhan Chandra Krishi Viswavidyalaya, Mohanpur, Nadia- 741252, West Bengal-India.

*Corresponding author: patelr@icfre.org; rubypatelssac@gmail.com

ORCiD: https://orcid.org/0009-0007-0628-2554

Abstract:

Background and aims The present investigation aims to identify the best-fit machine learning algorithm and highly important soil variable for prediction of SOC content in forest soils in Telangana, India.

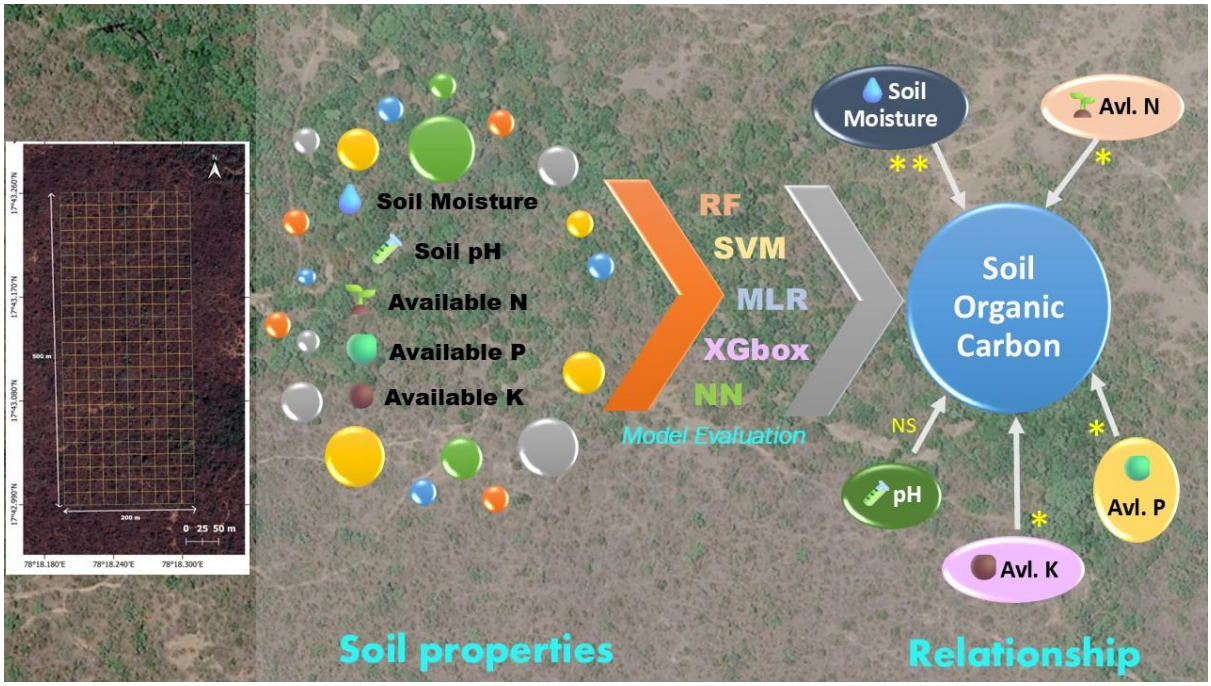
Methods The air-dried and processed samples were analysed for soil pH, soil organic carbon (SOC) content, available nitrogen (N), available phosphorus (P), available potassium (K), and soil moisture using standard methods. Five machine learning algorithms such as multiple linear regression, random forest, neural networks, extreme gradient boosting, and support vector machines were employed to model SOC variability in the southern dry mixed deciduous forest soils of Telangana. Model performance was assessed using cross-validation statistics based on three evaluation criteria: the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE).

Results Soil pH was near neutral. Available nitrogen and available potassium indicated moderate fertility, whereas mean available phosphorus was comparatively low. Soil organic carbon content fell within the medium range. Based on average feature importance, soil moisture was identified as the most influential variable for predicting SOC content in forest soils. Among the five algorithms tested, the random forest model performed best. In contrast, the support vector machine model exhibited the poorest predictive performance.

Conclusion The findings highlight the importance of soil moisture and available nutrients (N, P and K) as key covariates for accurate SOC prediction in dry deciduous forest soils, with the random forest model performing best, while the support vector machine model requires further optimisation.

Keywords: Soil organic carbon. Southern dry mixed deciduous forest. Prediction. Machine learning algorithm. Soil moisture. Soil nutrients

Graphical Abstract:



1. Introduction

Soil organic matter serves as a powerhouse for storing organic carbon and other essential nutrients required for biomass production. Organic carbon sequestered in forest soils is two to three times higher than that stored in terrestrial vegetation and the atmosphere (Schmidt et al. 2011). The carbon dynamics between soil and atmosphere largely depend on soil organic carbon (SOC) stocks. To mitigate the escalating threats of climate change, the release of carbon into the atmosphere should be minimized (Zhang et al. 2016).

The decomposition of organic matter is primarily influenced by soil physical properties (such as bulk density, temperature, and moisture) and soil fertility status (pH, available nitrogen, available phosphorus, and available potassium). In organic matter, carbon, nitrogen, phosphorus, and sulphur are closely associated. Understanding SOC dynamics and the key driving factors including soil moisture, soil temperature, and essential nutrients is critical for predicting carbon sequestration potential, emissions to the atmosphere, and developing sustainable forest land management strategies (Wiesmeier et al. 2012).

Forest soils are crucial natural resources that play a vital role in carbon dynamics, including carbon sequestration, the global carbon cycle, and climate change regulation (Chmura et al. 2003). The degradation of organic matter is highly sensitive to microbial activity, climate, and anthropogenic disturbances, making it a key driver of climate change (Smith et al. 2015). The sequestration of SOC depends on several key factors such as vegetation type, climate (rainfall and temperature), soil physical properties, and soil fertility status (Garcia-Pausas et al. 2007;

Lindner et al. 2010). Furthermore, tree species diversity and the richness of flora and fauna significantly influence the capacity of soils to sequester carbon (Jones et al. 2009).

Soil carbon and nutrients including N, P, and K play an important role in the formation of plant communities, their species and structural diversity, thus soil conservation has fundamental significance for biodiversity conservation and sustainable land use in forest ecosystems (Tiessen, et al. 1994; Ryals, et al. 2014).

According to the Forest Survey of India (FSI) report (2023), there are 12 various Forest Types in Telangana and out of 12 Forest Types, southern dry mixed deciduous forest dominant with 59.31 percent area of total forest cover. In the study area or a long-term ecological monitoring plot was established in the southern dry mixed deciduous forest to capture the SOC variability and soil fertility status. It aimed to improve the biomass and biodiversity, monitoring the deficiency or sufficiency of the nutrients.

Several predictive models have been developed to map the spatial distribution of SOC using key environmental factors and other related covariates. For several decades, multiple linear regression (MLR) has been the most widely used approach for SOC prediction due to its simplicity and ease of interpretation (Thompson and Kolka, 2005; Thompson et al. 2006; Guo et al. 2015). However, MLR is limited to capturing only linear interactions between SOC and its predictors. Given the highly heterogeneous nature of forest soils, the interactions between SOC and environmental variables are often complex and non-linear (Lark 1999).

In response, later studies employed generalized linear models (GLMs) and classification and regression trees (CART) to detect non-linear interactions between SOC and its predictors (Alvarez et al. 2011; Guo et al. 2013; Hobley et al. 2015; Ru et al. 2016). CART techniques gained wide popularity due to their superior ability to handle complex, non-linear relationships (Myles et al. 2004).

In recent years, there has been significant growth in the development and application of advanced machine learning models such as Random Forest (RF), Neural Networks (NN), Support Vector Machines (SVM), and Extreme Gradient Boosting (XGBoost) to further improve prediction accuracy in complex systems (Mutale et al. 2024; Uc Castillo et al. 2025). These models offer several advantages over traditional prediction methods, including greater resistance to over fitting, insensitivity to noisy features, and the provision of an unbiased estimate of error rates (Breiman 2001).

In several studies, to predict soil properties, auxiliary environmental data (covariates) were extracted using remote sensing and GIS techniques (Guo et al. 2013). A wide range of datasets derived from Digital Elevation Models (DEM) such as elevation, slope, and topographic indices along with vegetation indices like NDVI and EVI, have been commonly used as predictors for mapping SOC content.

The Cubist model outperformed RF model for improved SOC prediction mapping using multi-temporal satellite data in Iran (Fathololoumi et al. 2020). In another study conducted in southwestern Burkina Faso, Forkuor et al. (2017) reported that RF performed better for SOC prediction using environmental variables derived from remote sensing. Deep Neural Network (DNN) techniques performed superior with lower error, bias, and uncertainty for SOC prediction in northern Iran (Emadi et al. 2020). Boosted Regression Trees (BRT) and multiple stepwise regression have also been used to map SOC in the coastal regions of China (Bian et al. 2020).

Recently, several studies have employed machine learning algorithms including Random Forest (RF), Support Vector Regression (SVR), Neural Networks (NN), Regression Tree (RT), k-Nearest Neighbour (kNN), and Genetic Programming (GP) for SOC prediction using covariates extracted through remote sensing and GIS technologies (Chen et al. 2024; Zhang et al. 2024; Adeel et al. 2026). However, to the best of our knowledge, very few or no studies have incorporated soil nutrients and soil physical properties as covariates for predicting and modelling SOC content in dry deciduous forests, both in India and at the global level.

A significant research gap remains regarding how other soil nutrients and soil moisture, when used as covariates, can improve the prediction accuracy of SOC and contribute to better soil health management in forest ecosystems. The present study aims to test the hypothesis that soil physical properties and soil nutrient status play important roles in explaining SOC variability in forest soils when used as auxiliary environmental data. In this study, five machine learning algorithms Multiple Linear Regression (MLR), Random Forest (RF), Neural Networks (NN), Extreme Gradient Boosting (XGBoost), and Support Vector Machines (SVM) were used to capture SOC variability in the southern dry mixed deciduous forest soils of Telangana, India, with soil nutrients and soil moisture as key environmental variables.

The present study focuses on identifying the best-fit model for predicting SOC content and examining the effects of soil moisture and soil nutrient status on SOC in forest soils. Accordingly, the following were research objectives:

(i) to estimate SOC content in southern dry mixed deciduous forest soils of Telangana, India;

(ii) to assess the influence of soil moisture and soil nutrients on SOC content; and

(iii) to evaluate machine learning algorithms and identify the best-performing model for SOC prediction in southern dry mixed deciduous forest soils.

2. Material and methods

2.1. Study Area

A permanent long-term ecological monitoring research plot of 10 hectares (ha) area was established in the southern dry mixed deciduous forest of Narsapur, Telangana, following the forest census protocol of Condit (1998). The study plot is rectangular in shape, measuring 500 m in length and 200 m in width, and is oriented in a north-south direction (Fig. 1). The 10-ha plot is divided into ten one-ha subplots (100 m × 100 m each), and each subplot is further subdivided into twenty-five 20 m × 20 m quadrats (0.04 ha each). All quadrats and subplots are demarcated with stone pillars installed at the corners of each 20 m × 20 m quadrat.

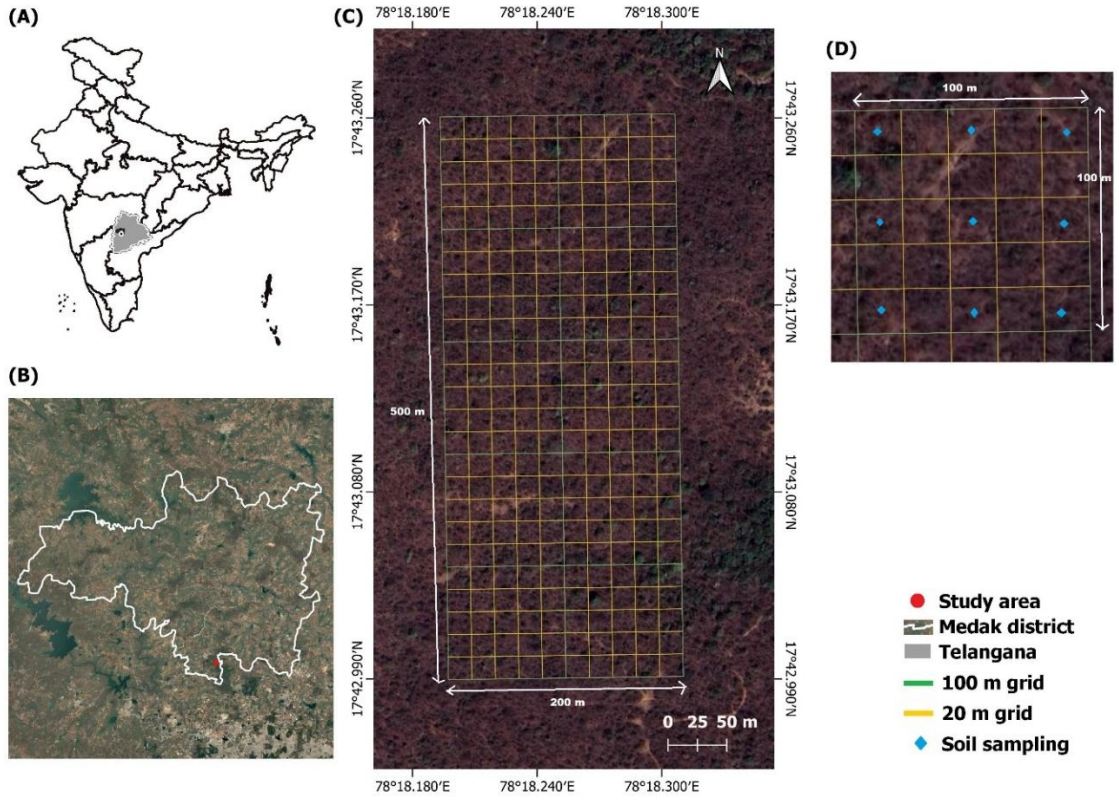


Fig. 1 Study area map of the permanent long-term ecological monitoring plot in the Southern Dry Mixed Forests of Narsapur, Telangana.

2.2. Soil Samples Collection

In each hectare, nine soil sampling points (one A point, four B points, and four C points) were established using a systematic grid approach (Fig. 2). From each sampling point, soil samples were collected at five depth intervals: 0-10 cm, 10-20 cm, 20-30 cm, 30-50 cm, and 50-100 cm. A total of 450 samples were collected from the entire 10 ha study plot (9 sampling points per hectare × 5 soil depths × 10 hectares). Prior to sampling, surface vegetation such as grasses and shrubs, along with litter including leaves and twigs, was removed at soil sampling points.

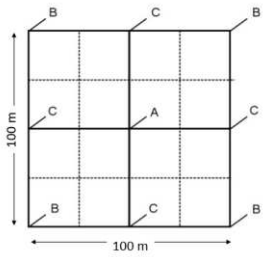


Fig. 2 Layout of sampling points for each hectare.

2.3. Soil analysis

The collected soil samples were shade-dried and then hand-crushed to remove gravels and plant residues (roots, stems, and leaves). The air-dried samples, processed and passed through a 2.0 mm sieve, were further analyzed for soil pH, SOC content, available N, available P, available K, and soil moisture using standard methods (Walkley and Black, 1934; Subbiah and Asija, 1956; Watanabe and Olsen, 1965; Jackson, 1973).

2.4. Data analysis

2.4.1. Descriptive statistics

Descriptive statistical parameters were analysed for all the parameter including SOC, pH, available form of N, P and K, and soil moisture including the mean, standard error of mean, standard deviation (SD), Median, Range and coefficient of variation (CV). The coefficient of variation was utilized to explain the variability in soil properties.

2.4.2. Multi linear Regression (MLR)

MLR is a classical algorithm widely applied to predict the values of a dependent variable using covariates (predictor variables). In the present study, MLR was employed to predict soil organic carbon (SOC) content based on five covariates: pH, available nitrogen (N), available phosphorus (P), available potassium (K), and soil moisture. A linear relationship was established between the target variable (SOC content) and these soil properties. The MLR equation is presented below.

$$y = a + \sum_i^n b_i \times x_i \pm \epsilon_i \quad (1)$$

Where, n → number of predictors; y → SOC content (target variable); x_i → predictors (soil properties); a →intercept; b_i → partial regression coefficients; ϵ_i → the residuals.

2.4.3. Random Forest (RF)

RF is a user-friendly, non-parametric ensemble method that builds multiple decision trees from bootstrap samples of the training dataset (Breiman, 2001). In addition to generating predictions, it provides valuable insights into feature importance. Unlike many other machine-learning approaches, RF relies on only two main parameters: (i) ntree, the number of regression trees grown in the forest, and (ii) mtry, the number of predictor variables randomly chosen at each node split. Randomness is incorporated into the model by selecting the optimal split from a randomly drawn subset of predictors. RF enhances model accuracy and reduces over fitting by averaging the outputs of multiple decision trees Taghizadeh-Mehrjardi et al. 2016). Approximately 60% of the dataset is used to train the regression trees, while the remaining 20% is allocated for validation and another 20% for testing. The relative error is then calculated by comparing the model predictions with the observed data**.** For fitting the Random Forest (RF) model, we employed the RandomForestRegressor function from the sklearn.ensemble in python language.

Equation (2) employs the averaged outputs of multiple decision trees to predict SOC. The performance of the RF model is evaluated, and the optimal split at each node is determined by minimizing error using the Mean Squared Error (MSE), as shown in Equation (2).

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i\,(x) \tag{2}$$

where:

$T_i(x) \rightarrow$ Predicted SOC from the i^{th} tree; $N \rightarrow$ Total number of regression trees

2.4.4. Support Vector Machines (SVMs)

SVMs are versatile supervised learning algorithms capable of handling both linear and non-linear interactions in large datasets. The core principle of SVM is to separate data points into various classes using optimal hyperplanes. During SVM modelling of the large datasets, two key parameters must be tuned: (i) the penalty parameter (C), which balances maximizing the margin against minimizing training errors, and (ii) the kernel width (σ), which defines the degree of nonlinearity in the model. SVMs have been successfully applied in agriculture, forestry, and ecological studies, particularly for predicting soil organic carbon (SOC) using other explanatory variables. Several researchers have reported that SVM often outperforms alternative models due to its strong learning capability and lower prediction errors (Pisner and Schnyer, 2020). Since soil organic carbon (SOC) content and other soil parameters exhibit non-linear relationships in the present study, we applied a Radial Basis Function (RBF) kernel as a transformation function to project the data into a higher-dimensional space. We implemented the SVM model

using the SVM class from the sklearn.svm module in Python. The following function f(x) is used to reduce the SOC prediction error within margin:

$$f(x) = w^T x + b \quad (3)$$

Where, $w \rightarrow$ weight vector; $x \rightarrow$ input feature vector; $b \rightarrow$ bias term.

2.4.5. Extreme Gradient Boosting (XGBoost)

XGBoost is an algorithm composed of multiple decision tree iterations, first introduced by Chen and Guestrin (2016). XGBoost enhances the performance of regression trees and classification methods and improve weak learners to a strong learner by employing boosting strategies. Moreover, XGBoost reduces over-estimation during prediction and allows simultaneous computations of functions during dataset training. The algorithm is refined through an iterative process, where each successive tree is trained to capture the residuals of the previous one. With the addition of more trees, the ensemble model gradually increases in complexity, eventually approximating the complexity of the data itself. This progressive refinement enables the training process to achieve optimal predictive performance.

2.4.6. Neural Networks (NN)

The NN approach was developed as a signal-processing technique for pattern recognition, inspired by nervous systems of human brain (Haykin 1994). The neural network architecture and training rules for model refinement are the key characteristics of NN. De Wilde (1996) suggested that training data can be introduced to the neural network that is form of nonlinear relationships, which can be addressed using simple algorithms. One of the major strengths of NN applications in biological and ecological sciences is its ability to process complex datasets containing intricate correlations. Moreover, NN has the capacity to identify statistically significant factors. In this study, NN-based models were employed to predict SOC using the five set of soil properties (pH, available N, available P, available K, and soil moisture) as input variables (Fig. 3). Over recent years, NN have emerged as powerful modelling tools in agriculture and forestry sciences due to their ease of application, strong extrapolative capabilities, and robustness of outcomes compared with conventional process-based models (Khoshnevisan et al.

2013). NN models do not require *a priori* assumptions about data structure and are highly effective in handling complex, non-linear interactions, making them superior to traditional mathematical approaches. The architecture of an NN generally comprises three fully connected neuronal layers: the input layer, consisting of neurons or nodes corresponding to input variables; the hidden layer, which determines the level of complexity during model development; and the output layer, containing neurons or nodes dedicated to each output variable. The output layer generates predictions for unknown values, while the input layer represents past observations. The hidden layer receives information from the input layer and applies suitable non-linear transfer functions. Learning is achieved through the back-propagation (BP) algorithm, which assigns weights (w) from the output layer and propagates them backward through the network. To minimize error in the model for particular training patterns, the BP algorithm employs gradient descent methods (Molga 2003). NNs capture non-linear interactions among independent variables by applying transformation functions (e.g. sigmoid) within a cross-linked neural network architecture. Target/output variables are derived by applying the sigmoid function as a non-linear transformation in the hidden layer and a linear function in the output layer.

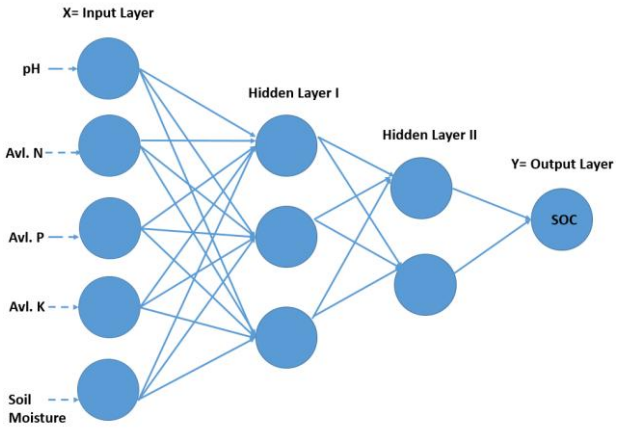


Fig. 3 Neural Network Structure for SOC prediction in forest soils.

2.5. *Evaluation of Model Performance*

In the present study, five algorithms were employed to predict SOC using other soil parameters as covariates. The performance of each algorithms was evaluated through cross-validation approach for prediction of SOC content in the southern dry mixed deciduous forest. For this purpose, the dataset of 450 samples was randomly split into 270 samples for training, 90 samples for validation, and 90 samples for testing each algorithm. Model performance was assessed using cross-validation statistics on the basis of three evaluation criteria: the coefficient of

determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). The best performing prediction algorithms was selected on the basis of the highest R^2, lowest mean absolute error (MAE) and RMSE values as the best for SOC prediction using covariates. The formulas for R^2, RMSE, and MAE are expressed as (Kambhammettu et al. 2011):

$$\text{Mean Absolute Error (MAE)} = \frac{1}{n}\sum_{i=1}^{n} | P_i - O_i | \quad (4)$$

$$\text{Root Mean Square Error (RMSE)} = \sqrt[2]{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \quad (5)$$

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{n}(P_i - P)^2}{\sum_{i=1}^{n}(O_i - O)^2}\right) \quad (6)$$

where, $n \rightarrow$ Total number of samples; $O_i \rightarrow$ observed SOC contents; $P_i \rightarrow$ predicted SOC contents.

3. Result

3.1. Descriptive statistics of soil parameters

The descriptive statistics of six soil parameters pH, available nitrogen (N), available phosphorus (P), available potassium (K), soil moisture, and SOC content are presented in Table 1. In the permanent study plot, soil pH varied from 6.1 to 9.07, SOC content from 0.17% to 2.45%, and soil moisture from 2% to 21% (Table 1). Available nitrogen, available phosphorus, and available potassium varied between 90 and 324.66 kg ha^{-1}, 1.05 and 26.33 kg ha^{-1}, and 49.04 and 344.4 kg ha^{-1}, respectively. Soil pH is stable (7.03 ± 0.02), indicating near neutral soils. Available N (173.3 ± 2.18 kg ha^{-1}) and available K (166.1 ± 2.34 kg ha^{-1}) show moderate fertility. However the mean available P was lower 10.2±0.21. SOC content was in medium range (0.78 ± 0.02%). Median SOC (0.68%) is slightly lower than the mean, suggesting a right skewed distribution with some higher values. Similar skewness is observed for Available P and K. SOC has a high CV (51.06%), showing strong variability across samples. The coefficient of variation (CV) was used to assess the degree of dispersion in the values of variables. CV < 10% indicates weak variability; 10-90% reflects moderate variability and > 90% denotes strong variability in the dataset (Guan et al. 2017). Soil moisture also varies widely (CV = 49.82%), reflecting heterogeneous site conditions. pH is the most stable parameter (CV = 5.46%). SOC ranges from 0.17% to 2.45%, highlighting variability in organic matter content. Soil moisture varies between 2.04% and 21.20%, indicating differences in micro site hydrology. Nutrient availability (N, P, and K) shows broad ranges, reflecting spatial heterogeneity in fertility. In this study, CV values for the six soil parameters ranged from 5.46% to 51.06%. Soil pH exhibited very low variability, while

parameters such as available form of nitrogen, phosphorus and potassium, soil moisture, and SOC content showed moderate variability in the datasets.

The descriptive statistics of six key soil parameters including pH, available nitrogen (N), available phosphorus (P), available potassium (K), soil moisture, and soil organic carbon (SOC) content are presented in Table 1. In the 10-ha permanent study plot, soil pH ranged from 6.1 to 9.07, SOC content from 0.17% to 2.45%, and soil moisture from 2.04% to 21.20%. Available nitrogen, phosphorus, and potassium ranged from 90 to 324.66 kg ha^{-1}, 1.05 to 26.33 kg ha^{-1}, and 49.04 to 344.4 kg ha^{-1}, respectively. The soils were near neutral in reaction (mean pH 7.03 ± 0.02). The mean available N (173.3 ± 2.18 kg ha^{-1}) and available K (166.1 ± 2.34 kg ha^{-1}) indicated moderate fertility levels, while mean available P was relatively low (10.2 ± 0.21 kg ha^{-1}). The mean SOC content was 0.78 ± 0.02%, which falls in the medium range for tropical forest soils.

The distribution of SOC was slightly right-skewed, as indicated by the lower median value (0.68%) compared to the mean. Similar skewness was observed for available P and K. SOC showed high variability across the plot (CV = 51.06%), followed by soil moisture (CV = 49.82%). In contrast, soil pH was the most stable parameter (CV = 5.46%). According to the classification proposed by Guan et al. (2017), CV values <10% indicate low variability, 10-90% indicate moderate variability, and >90% indicate high variability. In the present study, CV values for the six soil parameters ranged from 5.46% to 51.06%, reflecting moderate to high spatial heterogeneity in the dry deciduous forest ecosystem.

Table 1 Descriptive analysis of soil parameters in 10 Ha permanent ecological monitoring plot in southern dry mixed deciduous forest.

Parameters	**pH**	**Available N (Kg ha^{-1})**	**Available P (Kg ha^{-1})**	**Available K (Kg ha^{-1})**	**Soil Moisture (%)**	**SOC (%)**
Mean±SEm	7.03±0.02	173.3±2.18	10.2±0.21	166.1±2.34	10.94±0.26	0.78±0.02
Median	7.03	175.4361	11.13	161.112	10.95	0.68
SD	0.38	46.31	4.46	49.54	5.45	0.40
CV	5.46	26.73	43.76	29.82	49.82	51.06
Range	2.97	234.4	25.28	295.36	19.16	2.28
Minimum	6.1	90.25	1.05	49.04	2.04	0.17
Maximum	9.07	324.66	26.33	344.4	21.20	2.45

SOC- Soil Organic Carbon; SD - Standard Deviation; CV- Coefficient of Variation

3.2. *Correlation between SOC content and others soil parameters*

SOC showed a very weak negative correlation with soil pH (r = -0.101). Fig. 4 presents the correlation matrix between SOC and other soil parameters. In contrast, SOC was moderately correlated with soil moisture (r = 0.323) and strongly correlated with available nutrients such as nitrogen (r = 0.582), phosphorus (r = 0.585), and potassium (r = 0.517). The correlations between soil nutrients and soil moisture with SOC were positive, whereas the relationship with soil pH was negative.

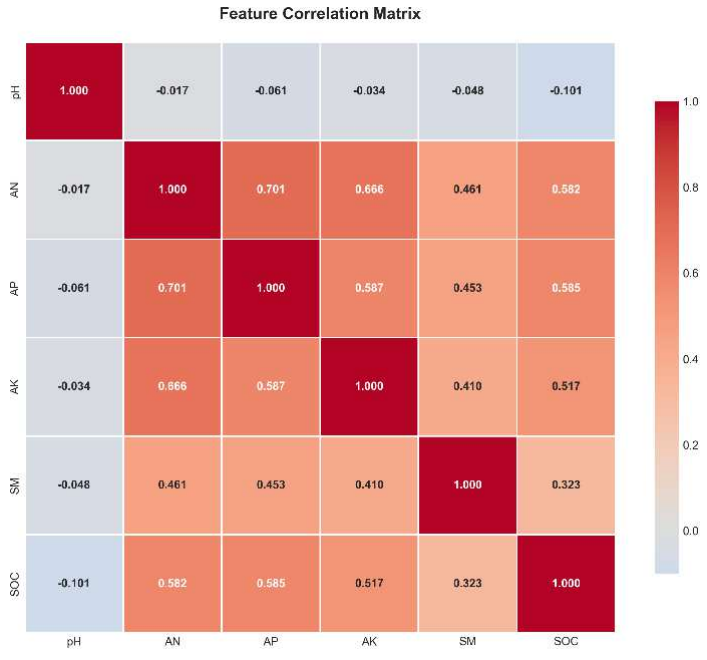


Fig. 4 Correlation matrix between SOC (%) content and other soil parameters. (AN → Available Nitrogen; AP → Available Phosphorus; AK → Available Potassium; SM → Soil Moisture; SOC → Soil Organic Carbon)

3.3. Comparison of MLR and other machine learning models

In this study, a total of 450 datasets for each soil variables were used for SOC content prediction in southern mixed dry deciduous forest. The data were split into a 60:20:20 ratio for the training (n = 270), validation (n = 90), and test (n = 90) sets for each covariates. Figure 6 illustrates the predicted SOC values plotted against their corresponding actual SOC values for the test set (n = 90) using five machine learning algorithms. The prediction accuracy of each algorithm was evaluated using standard validation indices such as MSE, MAE, R^2, RMSE, and uncertainty. The machine learning model RF had highest coefficient of determination (R^2) 0.5974 and the lowest MSE 0.350, lowest RMSE 0.2576, and uncertainty values -0.0410 was considered the best-fit prediction model. MSE and RMSE values close to zero indicate better performance of the machine learning models. The Fig. 6 present scatter plots of MLR, RF, SVM, and NN predictions versus measured SOC values.

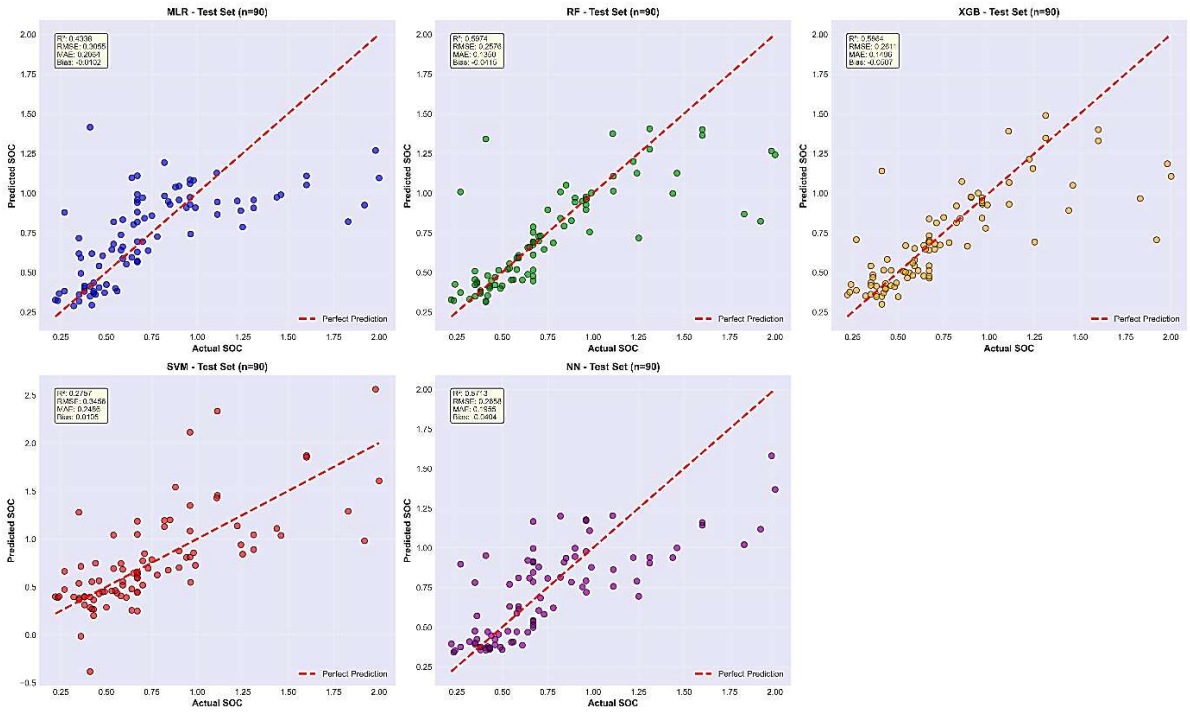


Fig. 5 Actual SOC vs. predicted SOC content using five machine learning algorithms: (**A**) MLR, (**B**) RF, (**C**) XGB, (**D**) SVM, and (**E**) NN. (MLR: Multiple linear regression; RF: Random Forest; XGB: Extreme Gradient Boosting; SVM: Support Vector Machine; NN: neural network)

Table 2 Cross-validation results of the five machine learning models for SOC prediction.

Model	MSE	RMSE	MAE	R^2	Uncertainty
Multiple Linear Regression (MLR)	0.093	0.305	0.206	43.3	0.305
Random Forest (RF)	0.066	0.257	0.135	59.7	0.254
Extreme Gradient Boosting (XGB)	0.068	0.261	0.149	58.6	0.256
Support Vector Machine (SVM)	0.119	0.345	0.248	27.5	0.345
Neural Network (NN)	0.070	0.265	0.195	57.1	0.262

MSE → Mean Square Error; RMSE → Root Mean Square Error; MAE → Mean Absolute Error; R^2 → Coefficient of Determination

In Fig. 5, the red 1:1 line represents the ideal case (perfect prediction) where predicted values equal measured values. Fig. 5 shows that the scatter plots of RF, followed by XGB and NN, were closer to the 1:1 line compared to the other algorithms, indicating superior predictive accuracy. The RF model recorded the highest R^2 value (59.7%), followed by XGB and NN, while MLR showed the lowest R^2 value (43.3%) (Table 2). Similarly, RF achieved the lowest RMSE (0.257), MSE (0.066), and MAE (0.135), whereas MLR exhibited the highest error values. Overall, the RF model, with the lowest RMSE, MSE, and MAE and the highest R^2, performed best among all the algorithms.

3.4. *Residual analysis of the algorithms*

Fig. 6 presents the scatter plots of residuals versus predicted SOC content. The red dashed line represents the ideal case of perfect prediction, where residuals or error equal zero. For MLR and NN algorithms, residuals were more widely dispersed from zero, indicating moderate prediction errors. In contrast, residuals from RF and XGB were more closely clustered around the zero line, reflecting greater stability and consistency in predictions. The SVM algorithm exhibited the widest spread of residuals, suggesting substantial variability in prediction errors.

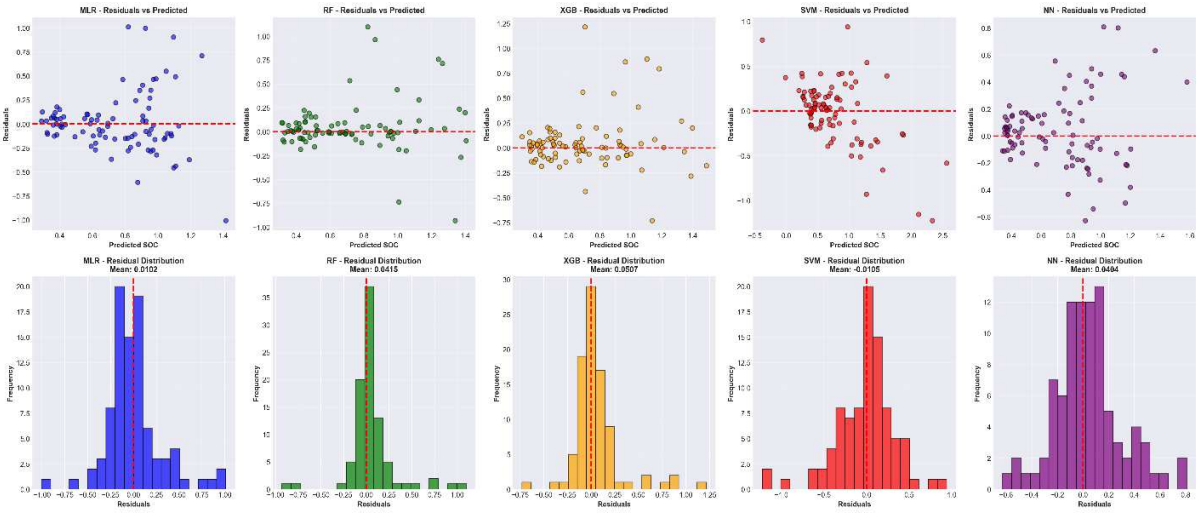


Fig. 6 Residuals vs. predicted scatter plot of five machine learning algorithm.

The accompanying histograms illustrate the distribution and symmetry of residuals for each algorithm. RF and XGB displayed narrow, symmetric distributions, indicative of stable predictive performance. SVM showed a negatively biased distribution with high variability, while NN exhibited a positively biased distribution with moderate spread, and more dispersion compared to the ensemble models. Overall, ensemble methods (RF and XGB) demonstrated superior predictive reliability compared to MLR, SVM, and NN.

3.5. *Feature importance of the variables in each algorithm for prediction*

In the MLR model, all soil parameters exhibited relatively small coefficient magnitudes for SOC prediction. However, soil pH showed a comparatively higher coefficient magnitude (0.076) than the other parameters. In the Random Forest and XGBoost models, soil moisture content emerged as the most important predictor, with feature importance values of 0.61 and 0.48, respectively, in explaining SOC variability in the southern mixed dry deciduous forest (Fig. 7). In the NN model, available phosphorus (0.14) had the highest feature importance, followed by soil pH and available nitrogen. For the SVM model, available nitrogen (0.94) and soil moisture (0.92)

were the strongest predictors. On average, across all algorithms, soil moisture (0.78) was identified as the most influential predictor of SOC content in forest soils (Fig. 7).

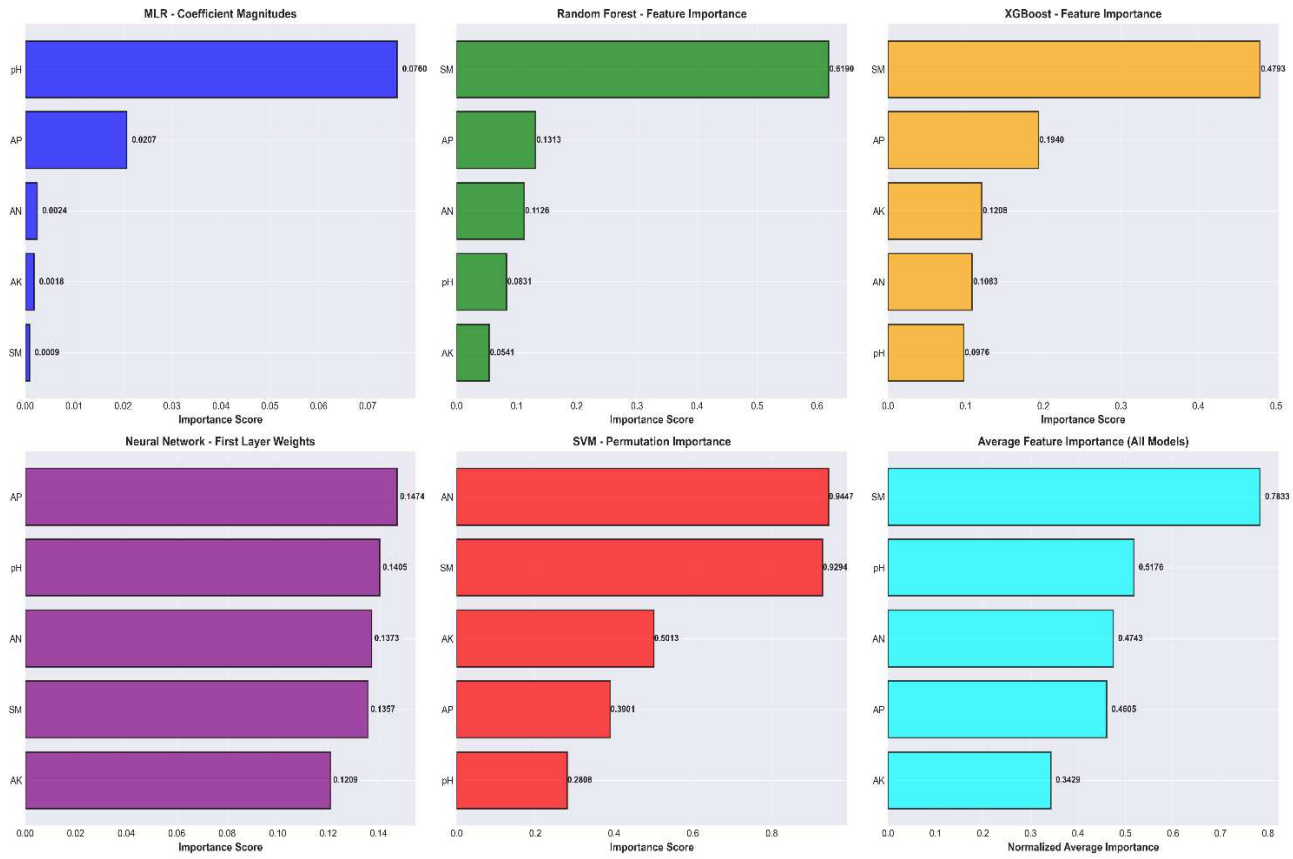


Fig. 7 Feature importance of the variables for prediction of SOC content in each machine learning model. (AN → Available Nitrogen; AP → Available Phosphorus; AK → Available Potassium; SM → Soil Moisture)

3.6. Uncertainty analysis of the models for the prediction

Fig. 8 illustrates the model comparison, highlighting the uncertainty among five predictive models MLR, RF, XGB, SVM, and NN. For uncertainty analysis, MLR (0.3053) was taken as the baseline, indicated by the red dashed line. Random Forest, XGB, and Neural Network models, with standard deviations of error values of 0.2543, 0.256, and 0.2628 respectively, exhibited lower uncertainty than the baseline, indicating better prediction consistency and stability for SOC content in forest soils (Table 4 and Fig. 8). In contrast, the SVM model showed the highest uncertainty (standard error = 0.3454), suggesting greater variability in predictions. Overall, Random Forest demonstrated the least uncertainty, followed closely by XGBoost, while SVM displayed the highest fluctuation in prediction errors.

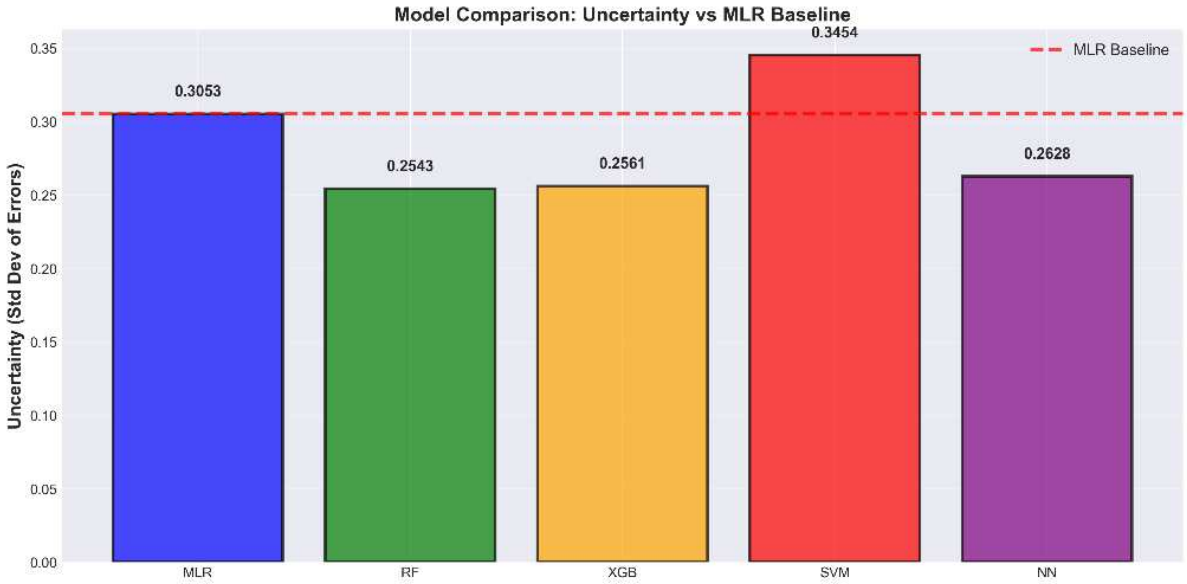


Fig. 8 Uncertainties of all five machine learning models for the SOC content prediction.

4. Discussion

4.1. Soil physical properties and nutrient status of southern dry mixed deciduous forest

The present investigation reports the soil fertility status of southern dry mixed forests. The forest soils were heterogeneous in nature and exhibited moderate variability in primary nutrient content. The average soil pH was recorded as 7.03 ± 0.02, indicating near-neutral conditions, with values ranging from 6.1 to 9.07. Similarly, Solanki et al. (2024) reported a mean soil pH of 7.20 ± 0.14 during the summer season in dry deciduous tropical forests. Likewise, Patel and Panwar (2024) reported soil pH values ranging from 5.12 to 8.14 in the forest soils of Telangana. Our findings are comparable to those of Jana et al. (2010), who reported soil pH values ranging from 5.90 to 7.01 in natural forests of West Bengal. In contrast, Joshi (2012) recorded acidic pH values (4.3-4.94) in tropical dry deciduous forests of West Bengal. Several other studies have also documented more acidic soil pH ranges compared to the present study (Gautam and Mandal 2013; Kuruppuarachchi et al. 2016). This may be attributed to the semi-arid climate of the study area, which favours neutral to alkaline soil pH. Under such conditions, lower soil moisture limits the decomposition of organic matter, resulting in reduced release of organic acids, which in turn often raises soil pH (Patel and Panwar 2024; Solanki et al. 2024).

The variation in SOC in forest soils at the same sites is highly uncertain, and identifying its primary cause is very difficult, as soil properties are governed by both intrinsic and extrinsic factors. The present study showed the average SOC content was 0.78 ± 0.02 % in southern dry mixed deciduous forest. Similarly, Patel et al. (2024) reported SOC ranging from 0.57 to 1.22% in southern dry deciduous forest soils. In contrast, Salas et al. (2017)

reported a higher SOC content of 3.3% in dry deciduous forest soils. The variation in SOC content may be attributed to differences in vegetation type, litter fall and its decomposition, and root activity in southern dry mixed deciduous forests (Jandl et al. 2007; Dawud et al. 2016; Li et al. 2019; Raha 2023).

The available nitrogen (N), phosphorus (P), and potassium (K) showed medium variability across the study area. The mean available N (173.3 ± 2.18 kg ha^{-1}) and available K (166.1 ± 2.34 kg ha^{-1}) indicated moderate fertility status, whereas the mean available P (10.2 ± 0.21 kg ha^{-1}) reflected low fertility. Rawat et al. (2022) reported available N ranging from 134.13 to 179.00 kg ha^{-1} in *Quercus leucotrichophora* forests of the Himalaya, which is comparable to the present findings. However, the available P values reported in their study (82.91-87.31 kg ha^{-1}) were substantially higher than those observed in the current investigation. In contrast, Thakrey et al. (2022) recorded available N, P, and K concentrations of 43.90 ± 0.33, 4.70 ± 0.79, and 291.85 ± 4.09 kg ha^{-1}, respectively, in undisturbed tropical dry deciduous forests of Chhattisgarh. Their values for available N and P were lower, while the available K was higher compared to the present study.

The mean soil moisture content in the study plots was 10.94 ± 0.26 %. This value is comparable to the mean soil moisture of 13.38 % reported by Rawat et al. (2022) in *Quercus leucotrichophora forests*. Similarly, Manral et al. (2022) recorded soil moisture ranging from 10.79 % to 22.48 % in Himalayan forest soils. In contrast, Solanki et al. (2024) reported considerably higher seasonal moisture contents in tropical dry deciduous forests, with values of 39-45 % in summer and 25-31 % in winter. Luo et al. (2023) documented soil moisture values between 8.22 % and 10.75 % in deciduous forests, which closely aligns with the present findings. Soil moisture content is primarily governed by climatic factors (temperature and rainfall), evaporation rates, and vegetation characteristics. The southern dry mixed deciduous forest experiences a semi-arid climate with moderate annual rainfall and a prolonged dry season (Tomar and Baishya 2020). The vegetation diversity in these forests is higher than that of dry teak or scrub forests but lower than in evergreen and semi-evergreen forests (Hasnat and Hossain 2022). This moderate tree species diversity, combined with the prevailing semi-arid conditions, results in medium-range soil moisture content in the southern dry mixed deciduous forest.

The present study revealed that soil organic carbon (SOC) was positively correlated with major soil nutrients and moderately correlated with soil moisture content. Similarly, Solanki et al. (2024) observed positive correlations between soil moisture and both SOC and primary nutrients in tropical dry deciduous forests. Kerr and Ochsner (2020) reported moderate to high positive correlations between soil moisture and SOC content across soil depths of 5-60 cm in temperate grasslands. Several other studies have also documented strong positive correlations

399 between SOC and the available forms of nitrogen (N), phosphorus (P), and potassium (K) in dry deciduous forest
400 soils (Gan et al. 2020; Thakrey et al. 2022; Spohn and Stendahl 2022). The strong association between SOC and
401 available nutrients can be attributed to the binding of nutrients within organic matter. During the decomposition
402 of organic matter, nutrients such as N, P, K, and S are mineralized and released into the soil along with the gaseous
403 release of CO_2 (Qiao et al. 2020; Chen et al. 2022).

404 *4.2. Evaluation of Machine Learning models for SOC prediction*

405 The present study showed that the Random Forest (RF) algorithm performed best for predicting soil organic
406 carbon (SOC) in southern dry mixed deciduous forests, achieving the highest R^2 value along with the lowest MSE,
407 RMSE, and MAE values. It was followed by Extreme Gradient Boosting (XGBoost) and Neural Network (NN),
408 while Multiple Linear Regression (MLR) showed moderate performance (R^2 = 0.43). Overall, the machine
409 learning algorithms performed well for SOC prediction, except for Support Vector Machine (SVM), which
410 recorded a low R^2 value of 0.27.These findings are consistent with Li et al. (2016), who reported that RF and
411 Cubist models performed best (R^2 value = 0.50-0.75), whereas NN, MLR, and SVM showed relatively poor
412 performance (R^2 value < 0.50) under flat terrain conditions. The RF model in the present study achieved a notably
413 higher R^2 value (0.59) compared to the mean R^2 of 0.48 reported by Wang et al. (2020) for RF-based spatial SOC
414 prediction in semiarid pastures in eastern Australia. Similarly, Nath (2006) reported an R^2 of 0.31 for MLR,
415 whereas the present study recorded a higher performance (R^2 = 0.43). Furthermore, Emadi et al. (2020) found that
416 RF, NN, and Cubist models exhibited comparable predictive capability for SOC in Iran; likewise, in the present
417 study, only RF and NN demonstrated strong predictive performance.

418 In contrast, some studies have reported Neural Networks as the best-performing model for SOC prediction
419 (Taghizadeh-Mehrjardi et al. 2016; Zhang et al. 2019). However, Mosleh et al. (2016) and Padarian et al. (2019)
420 observed that NN exhibited low predictive capacity with poor R^2 values, suggesting that model performance could
421 be improved by incorporating larger datasets of environmental variables and soil parameters. Bou Kheir et al.
422 (2010) successfully captured SOC variability in floodplain soils using a combination of soil properties and
423 environmental covariates derived from remote sensing. Similarly, Wiesmeier et al. (2011) emphasized that soil
424 type, land use, and parent material are critical environmental variables influencing SOC distribution. In flat terrain,
425 SOC spatial distribution has been well explained by both environmental factors and soil properties (Adhikari et
426 al. 2014). The present study explored the inclusion of key soil nutrients (N, P, and K) and soil moisture as

covariates for SOC prediction. These variables play a crucial role in carbon dynamics and are important for maintaining the balance between carbon sequestration in soils and gaseous carbon release into the atmosphere.

4.3. Soil parameters importance for SOC prediction in each algorithm

The analysis of feature importance revealed that soil moisture content played a crucial role in prediction of SOC in southern dry mixed deciduous forests. The RF and XGBoost models identified soil moisture as the most important variable, whereas the NN model assigned the highest importance to available phosphorus, followed by soil pH. In contrast, the SVM model highlighted available nitrogen as the most influential factor, followed by soil moisture. These findings align with John et al. (2020), who reported that the wetness index was the best predictor among 22 environmental variables for explaining SOC variability in flat terrain. They also emphasized that soil nutrient status explained SOC variability more effectively than other environmental variables in flat topographic landscapes. Similarly, several studies have shown that soil physical and chemical properties serve as stronger predictors of SOC in flat or relatively homogeneous terrains compared to topographic and climatic variables (Song et al. 2017; Solly et al. 2020). Li et al. (2020) further highlighted that environmental covariates derived from remote sensing perform better for large-scale heterogeneous areas, whereas their inclusion in small-scale studies under flat terrain conditions can reduce prediction accuracy and increase model complexity. The RF model identified soil moisture as the most robust and important predictor for accurately estimating SOC and total nitrogen contents in soils with considerable variation in soil moisture (Qu et al. 2025).

Each machine learning algorithm is designed to address specific challenges, and its predictive accuracy depends on several factors, including data availability, data quality, and dataset size. Moreover, the suitability of certain features or parameters for soil organic carbon (SOC) prediction varies depending on the ML algorithm used. This highlights the importance of careful feature selection to enhance the predictive power of the models. In the present study, the SVM algorithm exhibited relatively poor predictive performance ($R^2 = 0.27$) compared to MLR and other machine learning models. This poor performance may be attributed to the small sample size, complex non-linear interactions among covariates, and high sensitivity of SVM to the choice of kernel function, kernel coefficient, and penalty parameter (Zhao et al. 2022).

Several studies have reported that SVM often performs poorly in ecological datasets with limited sample sizes (Zhao et al. 2022; Emadi et al. 2020). Suboptimal hyperparameter tuning frequently leads to poor generalization, whereas tree-based ensemble models such as RF and XGBoost are more robust for small-to-medium datasets.

These models effectively capture complex non-linear relationships and feature interactions with the target variable SOC through their ensemble decision tree structure, requiring comparatively less fine-tuning.

4.4. Algorithm uncertainty for SOC prediction in forest soils

Machine learning algorithms primarily rely on the quality and quantity of input dataset. The dataset is typically divided into three subsets: training, validation, and testing, to develop the model and evaluate its performance. Data uncertainty can significantly reduce model accuracy by adversely affecting algorithm performance. The major limitations of any machine learning algorithm are the quantity and quality of the dataset, as well as the presence of uncertainty within the data sets (Saha and Srivastava 2014). Factors such as data veracity (data quality and uncertainty), data variety, incomplete data (missing values and data scarcity), and diverse data sources in large datasets are the primary contributors to uncertainty and reduced prediction accuracy (Wang and He 2016; Liu et al. 2018; Reimer and Madigan 2019).The present investigation revealed that Random Forest (RF), XGBoost, and Neural Network (NN) models exhibited lower uncertainty compared to the baseline Multiple Linear Regression (MLR) model, indicating superior prediction accuracy for SOC content in forest soils. In contrast, the Support Vector Machine model showed the highest uncertainty, resulting in greater variability in SOC predictions. The SVM model assumes that soil parameters such as pH, soil moisture, SOC, and available forms of N, P, and K in the dataset are precise and error-free. However, this assumption does not hold true for forest soils due to their high intrinsic (e.g. spatial heterogeneity) and extrinsic (e.g. climatic and biotic) variability (Wang and Pardalos, 2014). In future studies, SVM models can be further optimized by defining a hyper-rectangle for each parameter, assigning Sobol total effect indices to identify the key soil parameters influencing model robustness, and incorporating new feature weights based on Sobol indices. Such improvements are expected to enhance the performance of SVM models for SOC prediction in forest soils (Raghavendra et al. 2014; Utkin and Zhuk 2017; Singla et al. 2019).

5. Conclusion

Forest soils function as both sources and sinks for soil organic carbon (SOC). The decomposition of organic matter is strongly influenced by soil physical properties and nutrient status. Among these, soil physical properties particularly soil moisture play a critical role in SOC sequestration, followed by soil fertility parameters such as pH and the availability of nitrogen (N), phosphorus (P), and potassium (K).

The present study demonstrates that soil moisture is a highly important predictor of SOC content in southern dry mixed deciduous forests. Since carbon, nitrogen, phosphorus, and potassium are strongly bound within organic

matter, this study was conducted to predict SOC content in forest soils using key soil physico-chemical properties as environmental covariates. Five machine learning algorithms were evaluated for SOC prediction in forest soils. The correlation analysis revealed that SOC was weakly and negatively correlated with soil pH, moderately correlated with soil moisture, and strongly correlated with available nutrients (N, P, and K). Feature importance analysis further indicated that soil moisture was the most influential predictor, followed by soil pH and available nutrients. Among the models tested, the Random Forest (RF) algorithm achieved the best predictive performance with the highest R^2 value (0.59) and the lowest MSE, RMSE, and MAE values. It was followed by Extreme Gradient Boosting (XGBoost), Neural Network (NN), and Multiple Linear Regression (MLR). In contrast, the Support Vector Machine (SVM) model performed poorly, recording a low R^2 value of 0.27. Uncertainty analysis showed that RF and XGBoost had the lowest uncertainty compared to the baseline MLR model, while SVM exhibited very high uncertainty. These findings underscore the importance of incorporating soil moisture and available forms of nutrients (N, P, and K) as key covariates for accurate SOC prediction in dry mixed deciduous forest soils. The study also suggests that the performance of the SVM model can be significantly enhanced through targeted hyperparameter optimization for applications in forest ecosystems.

CRediT authorship contribution statement

Ruby Patel: Conceptualization, Methodology, Investigation, Formal analysis, Project administration; Resources; Supervision; Validation; Visualization; Writing – review & editing, Writing – original draft. **Bharati Patel**: Methodology, Investigation, Supervision, Writing – review & editing. **Sravanthi Palle**: Laboratory analysis. **Jaison M**: Formal analysis, Visualization, Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

Financial assistance from the project ‘Study the climate driven effects on Indian forests through climate change through long-term monitoring’ was funded by Compensatory Afforestation Fund Management and Planning Authority (CAMPA), Ministry of Environment, Forest and Climate Change, Government of India under the scheme ‘Strengthening Forestry Research for Ecological Sustainability and Productivity Enhancement’ as All India Coordinated Research Project to the Indian Council of Forestry Research and Education (ICFRE).

References

514 Adeel A, Hasani M, Jadhav AS (2026) Soil organic carbon modeling in cropland under several climatic scenarios
515 using machine learning in western India. Sci Rep 16:5485. https://doi.org/10.1038/s41598-026-35191-4
516
517 Adhikari K, Hartemink AE, Minasny B, Bou Kheir R, Greve MB, Greve MH (2014) Digital mapping of soil
518 organic carbon contents and stocks in Denmark. PLoS One 9:e105519.
519 https://doi.org/10.1371/journal.pone.0105519
520
521 Alvarez R, Steinbach H, Bono A (2011) An artificial neural network approach for predicting soil carbon budget
522 in agroecosystems. Soil Sci Soc Am J 75:965–975. https://doi.org/10.2136/sssaj2009.0427
523
524 Bian Z, Guo X, Wang S, Zhuang Q, Jin X, Wang Q, Jia S (2020) Applying statistical methods to map soil organic
525 carbon of agricultural lands in northeastern coastal areas of China. Arch Agron Soil Sci 66:532–544.
526 https://doi.org/10.1080/03650340.2019.1626983
527
528 Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
529
530 Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD
531 International Conference on Knowledge Discovery and Data Mining. ACM, New York, pp 785–794.
532 https://doi.org/10.1145/2939672.2939785
533
534 Chen X, Feng J, Ding Z, Tang M, Zhu B (2022) Changes in soil total, microbial and enzymatic C:N:P contents
535 and stoichiometry with depth and latitude in forest ecosystems. Sci Total Environ 816:151583.
536 https://doi.org/10.1016/j.scitotenv.2021.151583
537
538 Uc Castillo JL, Marín Celestino AE, Martínez Cruz DA, Tuxpan Vargas J, Ramos Leal JA, Morán Ramírez J
539 (2025) A systematic review of Machine Learning and Deep Learning approaches in Mexico: challenges
540 and opportunities. Front Artif Intell 7:1479855. https://doi.org/10.3389/frai.2024.1479855
541
542 Chen Q, Wang Y, Zhu X (2024) Soil organic carbon estimation using remote sensing data-driven machine
543 learning. PeerJ 12:e17836. https://doi.org/10.7717/peerj.17836
544
545 Chmura GL, Anisfeld SC, Cahoon DR, Lynch JC (2003) Global carbon sequestration in tidal, saline wetland soils.
546 Glob Biogeochem Cycles 17:1111. https://doi.org/10.1029/2002GB001917
547
548 Condit R (1998) Tropical forest census plots: methods and results from Barro Colorado Island, Panama and a
549 comparison with other plots. Springer Science & Business Media, Berlin
550
551 Dawud SM, Raulund-Rasmussen K, Domisch T, Finér L, Jaroszewicz B, Vesterdal L (2016) Is tree species
552 diversity or species identity the more important driver of soil carbon stocks, C/N ratio, and pH? Ecosystems
553 19:645–660. https://doi.org/10.1007/s10021-016-9958-1

De Wilde P (1996) Neural Networks Models: An Analysis. Springer, Berlin Heidelberg. https://doi.org/10.1007/BFb0034479

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 Sens 12:2234. https://doi.org/10.3390/rs12142234

Fathololoumi S, Vaezi AR, Alavipanah SK, Ghorbani A, Saurette D, Biswas A (2020) Improved digital soil mapping with multitemporal remotely sensed satellite data fusion: A case study in Iran. Sci Total Environ 721:137703. https://doi.org/10.1016/j.scitotenv.2020.137703

Forkuor G, Hounkpatin OK, Welp G, Thiel M (2017) High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: a comparison of machine learning and multiple linear regression models. PLoS One 12:e0170478. https://doi.org/10.1371/journal.pone.0170478

Gan HY, Schöning I, Schall P, Ammer C, Schrumpf M (2020) Soil organic matter mineralization as driven by nutrient stoichiometry in soils under differently managed forest stands. Front For Glob Chang 3:99. https://doi.org/10.3389/ffgc.2020.00099

Garcia-Pausas J, Casals P, Camarero L, Huguet C, Sebastia MT, Thompson R, Romanya J (2007) Soil organic carbon storage in mountain grasslands of the Pyrenees: effects of climate and topography. Biogeochemistry 82:279–289. https://doi.org/10.1007/s10533-007-9071-9

Gautam TP, Mandal TN (2013) Soil characteristics in moist tropical forest of Sunsari district, Nepal. Nepal J Sci Technol 14:35–40. https://doi.org/10.3126/njst.v14i1.8876

Guan F, Xia M, Tang X, Fan S (2017) Spatial variability of soil nitrogen, phosphorus and potassium contents in Moso bamboo forests in Yong'an City, China. Catena 150:161–172. https://doi.org/10.1016/j.catena.2016.11.017

Guo PT, Li MF, Luo W, Tang QF, Liu ZW, Lin ZM (2015) Digital mapping of soil organic matter for rubber plantation at regional scale: an application of random forest plus residuals kriging approach. Geoderma 237–238:49–59. https://doi.org/10.1016/j.geoderma.2014.08.009

Guo PT, Wu W, Sheng QK, Li MF, Liu HB, Wang ZY (2013) Prediction of soil organic matter using artificial neural network and topographic indicators in hilly areas. Nutr Cycl Agroecosyst 95:333–344. https://doi.org/10.1007/s10705-013-9566-9

Hasnat GT, Hossain MK (2022) Global overview of tropical dry forests. In: Research Anthology on Ecosystem Conservation and Preserving Biodiversity. IGI Global, pp 69–91. https://doi.org/10.4018/978-1-6684-5678-1.ch005

Haykin S (1994) Neural networks: a comprehensive foundation. Prentice Hall PTR, Upper Saddle River

Hobley E, Wilson B, Wilkie A, Gray J, Koen T (2015) Drivers of soil organic carbon storage and vertical distribution in Eastern Australia. Plant Soil 390:111–127. https://doi.org/10.1007/s11104-015-2380-1

Jackson ML (1973) Soil chemical analysis-advanced course: A manual of methods useful for instruction and research in soil chemistry, physical chemistry of soils, soil fertility, and soil genesis. University of Wisconsin, Madison

Jana BK, Biswas S, Sonkar S, Majumder M, Roy P, Mazumdar A (2010) Estimation of soil carbon stock and soil respiration rate of recreational and natural forests in India. In: Impact of climate change on natural resource management. Springer, Dordrecht, pp 329–343. https://doi.org/10.1007/978-90-481-3581-3_18

Jandl R, Lindner M, Vesterdal L, Bauwens B, Baritz R, Hagedorn F, Johnson DW, Minkkinen K, Byrne KA (2007) How strongly can forest management influence soil carbon sequestration? Geoderma 137:253–268. https://doi.org/10.1016/j.geoderma.2006.09.003

John K, Abraham Isong I, Michael Kebonye N, Okon Ayito E, Chapman Agyeman P, Marcus Afu S (2020) Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil. Land 9:487. https://doi.org/10.3390/land9120487

Jones A, Stolbovoy V, Rusco E, Gentile AR, Gardi C, Marechal B, Montanarella L (2009) Climate change in Europe. 2. Impact on soil. A review. Agron Sustain Dev 29:423–432. https://doi.org/10.1051/agro:2008067

Joshi HG (2012) Vegetation structure, floristic composition and soil nutrient status in three sites of tropical dry deciduous forest of West Bengal, India. Indian J Fundam Appl Life Sci 2:355–364

Kambhammettu BVNP, Allena P, King JP (2011) Application and evaluation of universal kriging for optimal contouring of groundwater levels. J Earth Syst Sci 120:413–422. https://doi.org/10.1007/s12040-011-0075-4

Kerr DD, Ochsner TE (2020) Soil organic carbon more strongly related to soil moisture than soil temperature in temperate grasslands. Soil Sci Soc Am J 84:587–596. https://doi.org/10.1002/saj2.20018

Kheir RB, Greve MH, Bøcher PK, Greve MB, Larsen R, McCloy K (2010) Predictive mapping of soil organic carbon in wet cultivated lands using classification-tree based models: The case study of Denmark. J Environ Manage 91:1150–1160. https://doi.org/10.1016/j.jenvman.2010.01.001

Khoshnevisan B, Rafiee S, Omid M, Mousazadeh H, Sefeedpari P (2013) Prognostication of environmental indices in potato production using artificial neural networks. J Clean Prod 52:402–409. https://doi.org/10.1016/j.jclepro.2013.03.028

Kuruppuarachchi KAJM, Seneviratne G, Madurapperuma D (2016) Carbon sequestration in tropical forest stands: its control by plant, soil and climatic factors. Open J For 6:59–71. https://doi.org/10.4236/ojf.2016.62006

Lark RM (1999) Soil–landform relationships at within-field scales: an investigation using continuous classification. Geoderma 92:141–165. https://doi.org/10.1016/S0016-7061(99)00028-2

Li L, Lu J, Wang S, Ma Y, Wei Q, Li X, Cong R, Ren T (2016) Methods for estimating leaf nitrogen concentration of winter oilseed rape (*Brassica napus* L.) using in situ leaf spectroscopy. Ind Crops Prod 91:194–204. https://doi.org/10.1016/j.indcrop.2016.07.008

Li X, McCarty GW, Du L, Lee S (2020) Use of topographic models for mapping soil properties and processes. Soil Syst 4:32. https://doi.org/10.3390/soilsystems4020032

Li Y, Bruelheide H, Scholten T, Schmid B, Sun Z, Zhang N, Bu W, Liu X, Ma K (2019) Early positive effects of tree species richness on soil organic carbon accumulation in a large-scale forest biodiversity experiment. J Plant Ecol 12:882–893. https://doi.org/10.1093/jpe/rtz026

Lindner M, Maroschek M, Netherer S, Kremer A, Barbati A, Garcia-Gonzalo J, Seidl R, Delzon S, Corona P, Kolström M, Lexer MJ (2010) Climate change impacts, adaptive capacity, and vulnerability of European forest ecosystems. For Ecol Manage 259:698–709. https://doi.org/10.1016/j.foreco.2009.09.023

Liu X, Tamminen S, Su X, Siirtola P, Röning J, Riekki J, Kiljander J, Soininen JP (2018) Enhancing veracity of IoT generated big data in decision making. In: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops. IEEE, pp 149–154. https://doi.org/10.1109/PERCOMW.2018.8480371

Manral V, Bargali K, Bargali SS, Jhariya MK, Padalia K (2022) Relationships between soil and microbial biomass properties and annual flux of nutrients in Central Himalaya forests, India. Land Degrad Dev 33:2014–2025. https://doi.org/10.1002/ldr.4283

Molga EJ (2003) Neural network approach to support modelling of chemical reactors: problems, resolutions, criteria of application. Chem Eng Process Process Intensif 42:675–695. https://doi.org/10.1016/S0255-2701(02)00205-2

Mosleh Z, Salehi M, Jafari A, Esfandiarpour I, Mehnatkesh A (2016) The effectiveness of digital soil mapping to predict soil properties over low-relief areas. Environ Monit Assess 188:195. https://doi.org/10.1007/s10661-016-5204-8

Mutale B, Withanage NC, Mishra PK, Shen J, Abdelrahman K, Fnais MS (2024) A performance evaluation of random forest, artificial neural network, and support vector machine learning algorithms to predict spatio-temporal land use-land cover dynamics: a case from Lusaka and Colombo. Front Environ Sci 12:1431645. https://doi.org/10.3389/fenvs.2024.1431645

Myles AJ, Feudale RN, Liu Y, Woody NA, Brown SD (2004) An introduction to decision tree modeling. J Chemom 18:275–285. https://doi.org/10.1002/cem.873

Nath DA (2006) Soil landscape modeling in the Northwest Iowa Plains region of O'Brien County, Iowa. Iowa State University, Ames

Padarian J, Minasny B, McBratney AB (2019) Using deep learning for digital soil mapping. Soil 5:79–89. https://doi.org/10.5194/soil-5-79-2019

Patel R, Panwar VP (2024) Spatial distribution of soil pH status in forest soils of Telangana using GIS-based geo-statistical models. Natl Acad Sci Lett 49:27–35. https://doi.org/10.1007/s40009-024-01559-1

Patel R, Sharief S, Sherline, Kodchire A, Panwar VP (2024) Predicting soil organic carbon in forest soils of Telangana, India: A comparative analysis of deterministic and geostatistical interpolation models. J Soil Sci Plant Nutr 24:7853–7868. https://doi.org/10.1007/s42729-024-02082-4

Pisner DA, Schnyer DM (2020) Support vector machine. In: Machine learning. Academic Press, pp 101–121. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

Qiao Y, Wang J, Liu H, Huang K, Yang Q, Lu R, Yan L, Wang X, Xia J (2020) Depth-dependent soil C:N:P stoichiometry in a mature subtropical broadleaf forest. Geoderma 370:114357. https://doi.org/10.1016/j.geoderma.2020.114357

Qu K, Nie L, Cui L, Li H, Xiong M, Zhai X, Zhao X, Wang J, Lei Y, Li W (2025) Vis–NIR spectroscopy characteristics of wetland soils with different water contents and machine learning models for carbon and nitrogen content. Ecologies 6:75. https://doi.org/10.3390/ecologies6040075

Raghavendra S, Deka PC (2014) Support vector machine applications in the field of hydrology: a review. Appl Soft Comput 19:372–386. https://doi.org/10.1016/j.asoc.2014.02.002

Raha D (2023) Tree diversity, stand structure and species composition in three tropical dry deciduous forests of Madhya Pradesh, Central India. Proc Int Acad Ecol Environ Sci 13:158

Rawat S, Khanduri VP, Singh B, Riyal MK, Thakur TK, Kumar M, Cabral-Pinto MM (2022) Variation in carbon stock and soil properties in different *Quercus leucotrichophora* forests of Garhwal Himalaya. Catena 213:106210. https://doi.org/10.1016/j.catena.2022.106210

Reimer AP, Madigan EA (2019) Veracity in big data: How good is good enough. Health Informatics J 25:1290–1298. https://doi.org/10.1177/1460458217744369

Ru F, Yin A, Jin J, Zhang X, Yang X, Zhang M, Gao C (2016) Prediction of cadmium enrichment in reclaimed coastal soils by classification and regression tree. Estuar Coast Shelf Sci 177:1–7. https://doi.org/10.1016/j.ecss.2016.04.018

Ryals R, Kaiser M, Torn MS, Berhe AA, Silver WL (2014) Impacts of organic matter amendments on carbon and nitrogen dynamics in grassland soils. Soil Biol Biochem 68:52–61. https://doi.org/10.1016/j.soilbio.2013.09.011

Saha B, Srivastava D (2014) Data quality: The other face of big data. In: 2014 IEEE 30th International Conference on Data Engineering. IEEE, pp 1294–1297. https://doi.org/10.1109/ICDE.2014.6816764

Salas Macias CA, Alegre Orihuela JC, Iglesias Abad S (2017) Estimation of above-ground live biomass and carbon stocks in different plant formations and in the soil of dry forests of the Ecuadorian coast. Food Energy Secur 6:e00115. https://doi.org/10.1002/fes3.115

Schmidt MW, Torn MS, Abiven S, Dittmar T, Guggenberger G, Janssens IA, Kleber M, Kögel-Knabner I, Lehmann J, Manning DA (2011) Persistence of soil organic matter as an ecosystem property. Nature 478:49–56. https://doi.org/10.1038/nature10386

Singla M, Ghosh D, Shukla KK (2020) A survey of robust optimization based machine learning with special reference to support vector machines. Int J Mach Learn Cybern 11:1359–1385. https://doi.org/10.1007/s13042-019-01044-y

Smith P, House JI, Bustamante M, Sobocká J, Harper R, Pan G, West P, Clark J, Adhya T, Rumpel C (2015) Global change pressures on soils from land use and management. Glob Chang Biol 22:1008–1028. https://doi.org/10.1111/gcb.13068

Solanki AC, Gurjar NS, Sharma S, Wang Z, Kumar A, Solanki MK, Kashyap BK (2024) Decoding seasonal changes: soil parameters and microbial communities in tropical dry deciduous forests. Front Microbiol 15:1258934. https://doi.org/10.3389/fmicb.2024.1258934

Solly EF, Weber V, Zimmermann S, Walthert L, Hagedorn F, Schmidt MW (2020) A critical evaluation of the relationship between the effective cation exchange capacity and soil organic carbon content in Swiss forest soils. Front For Glob Chang 3:98. https://doi.org/10.3389/ffgc.2020.00098

Spohn M, Stendahl J (2022) Carbon, nitrogen, and phosphorus stoichiometry of organic matter in Swedish forest soils and its relationship with climate, tree species, and soil texture. Biogeosciences 19:2171–2186. https://doi.org/10.5194/bg-19-2171-2022

Subbiah BV, Asija GL (1956) A rapid procedure for the estimation of available nitrogen in soils. Curr Sci 25:259–260

Taghizadeh-Mehrjardi R, Nabiollahi K, Kerry R (2016) Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. Geoderma 266:98–110. https://doi.org/10.1016/j.geoderma.2015.12.003

Thakrey M, Singh L, Jhariya MK, Tomar A, Singh AK, Toppo S (2022) Impact of disturbance on biomass, carbon, and nitrogen storage in vegetation and on soil properties of tropical dry deciduous forest in Chhattisgarh, India. Land Degrad Dev 33:1810–1820. https://doi.org/10.1002/ldr.4263

Thompson JA, Kolka RK (2005) Soil carbon storage estimation in a forested watershed using quantitative soil-landscape modeling. Soil Sci Soc Am J 69:1086–1093. https://doi.org/10.2136/sssaj2004.0322

Thompson JA, Pena-Yewtukhiw EM, Grove JH (2006) Soil–landscape modeling across a physiographic region: topographic patterns and model transportability. Geoderma 133:57–70. https://doi.org/10.1016/j.geoderma.2006.03.037

Tiessen H, Cuevas E, Chacon P (1994) The role of soil organic matter in sustaining soil fertility. Nature 371:783–785. https://doi.org/10.1038/371783a0

Tomar U, Baishya R (2020) Moisture regime influence on soil carbon stock and carbon sequestration rates in semi-arid forests of the National Capital Region, India. J For Res 31:2323–2332. https://doi.org/10.1007/s11676-019-01032-6

Utkin LV, Zhuk YA (2017) Interval SVM-based classification algorithm using the uncertainty trick. Int J Artif Intell Tools 26:1750014. https://doi.org/10.1142/S0218213017500142

790 Walkley A, Black IA (1934) An examination of the Degtjareff method for determining soil organic matter, and a
791 proposed modification of the chromic acid titration method. Soil Sci 37:29–38
792
793 Wang X, He Y (2016) Learning from uncertainty for big data: future analytical challenges and strategies. IEEE
794 Syst Man Cybern Mag 2:26–31. https://doi.org/10.1109/MSMC.2016.2557479
795
796 Wang X, Pardalos PM (2014) A survey of support vector machines with uncertainties. Ann Data Sci 1:293–309.
797 https://doi.org/10.1007/s40745-014-0022-8
798
799 Wang X, Zhang Y, Atkinson PM, Yao H (2020) Predicting soil organic carbon content in Spain by combining
800 Landsat TM and ALOS PALSAR images. Int J Appl Earth Obs Geoinf 92:102182.
801 https://doi.org/10.1016/j.jag.2020.102182
802
803 Watanabe FS, Olsen SR (1965) Test of an ascorbic acid method for determining phosphorus in water and $NaHCO_3$
804 extracts from soil. Soil Sci Soc Am J 29:677–678
805
806 Wiesmeier M, Barthold F, Blank B, Kögel-Knabner I (2011) Digital mapping of soil organic matter stocks using
807 Random Forest modeling in a semi-arid steppe ecosystem. Plant Soil 340:7–24.
808 https://doi.org/10.1007/s11104-010-0425-z
809
810 Wiesmeier M, Spörlein P, Geuß U, Hangen E, Haug S, Reischl A, Schilling B, von Lützow M, Kögel-Knabner I
811 (2012) Soil organic carbon stocks in southeast Germany (Bavaria) as affected by land use, soil type and
812 sampling depth. Glob Chang Biol 18:2233–2245. https://doi.org/10.1111/j.1365-2486.2012.02699.x
813
814 Zhang T, Li Y, Wang M (2024) Remote sensing-based prediction of organic carbon in agricultural and natural
815 soils influenced by salt and sand mining using machine learning. J Environ Manage 352:120107.
816 https://doi.org/10.1016/j.jenvman.2024.120107
817
818 Zhang X, Xu X, Liu Y, Wang J, Xiong Z (2016) Global warming potential and greenhouse gas intensity in rice
819 agriculture driven by high yields and nitrogen use efficiency. Biogeosciences 13:2701–2714.
820 https://doi.org/10.5194/bg-13-2701-2016
821
822 Zhang Y, Sui B, Shen H, Ouyang L (2019) Mapping stocks of soil total nitrogen using remote sensing data: A
823 comparison of random forest models with different predictors. Comput Electron Agric 160:23–30.
824 https://doi.org/10.1016/j.compag.2019.03.015
825
826 Zhao X, Zhao D, Wang J, Triantafilis J (2022) Soil organic carbon (SOC) prediction in Australian sugarcane
827 fields using Vis–NIR spectroscopy with different model setting approaches. Geoderma Reg 30:e00566.
828 https://doi.org/10.1016/j.geodrs.2022.e00566

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementrymaterials.docx