Indian Journal of Agricultural Sciences **95** (8): 897–903, August 2025/Article
https://doi.org/10.56093/ijas.v95i8.161380

Soil organic carbon variability assessment using satellite imagery and artificial neural network

SHYAMAL MUNDADA[1]* and POOJA JAIN[1]

Indian Institute of Information Technology, Nagpur, Maharashtra 441 108, India

Received: 28 November 2024; Accepted: 14 July 2025

ABSTRACT

The present study was carried out during 2022–2024 at Indian Institute of Information Technology, Nagpur, Maharashtra to evaluate SOC stocks in the Dhamtari district of Chhattisgarh, India. Two machine learning models and their variants-Boosted Regression Tree, Boosted Regression Tree with Early Stopping, Multilayer Perceptron, and Multilayer Perceptron with Early Stopping were used for predicting Soil Organic Carbon (SOC). The findings of the research indicated that Multilayer Perceptron produced better results in both scenarios that is, without and with Early Stopping technique applied. Multilayer Perceptron with Early Stopping model recorded nearly the same RMSE for both calibration and validation datasets as 0.1618 and 0.1601, respectively. Produced soil maps will assist farmers in adopting accurate information for decisions which will boost farm output and offer security for food through the balanced use of nutrients.

Keywords: Environmental covariates, Neural network soil organic carbon, Spatial modelling

Earth's topsoil provides a variety of ecosystem services that allow life to exist. The world's soil is under strain because of the fast transformations in land use and cover, particularly the transformation of natural ecosystems into agroecosystems. Various characteristics of soil are affected by agricultural land uses, consequences of which is soil degradation, especially the demise of soil organic matter (SOM). The primary component of SOM, i.e. SOC, controls soil properties. It preserves quality of soil by providing nutrients and increasing water-holding ability (Bationo *et al.* 2007). Frequent ploughing and other disorders degenerate the aggregates and change the soil's aeration, water holding capacity, and temperature conditions, which leads to the depletion of SOC which influences fertility of soil and, consequently, potential for agriculture (Batjes 1996, Zhenxing Bian and Jia 2020). Furthermore, there is a correlation between SOC stock and soil water penetration, water holding, and soil structural stability (Lefevre *et al.* 2017). Thus, information on the various aspects of SOC is needed worldwide for different purposes. To forecast the SOC of different types of soil or locations, it is therefore essential to develop a system that is more accurate and reliable. To use soil for agricultural and other ecological purposes, it is essential to know the spatial spread of these important nutritional elements in soil (Brady and Weil 2008, El-Ramady *et al.* 2014). Also, proper nutrient management in soil enhances the crop productivity (Sarkar *et al.* 2025). With advancements in data analysis, remote sensing, and geographic information systems, a variety of mapping techniques have been used and developed to increase the precision of the approach and the produced spatial maps. Digital soil mapping (DSM) is based on this concept where soil property's variability is described by how it relates to soil-forming elements including terrain, climate, vegetation, and soil nature. Prediction of SOC using this technique made use of a wide range of statistical techniques, such as kriging (Cambule *et al.* 2014), regression-kriging (Hengl and Heuvelink 2004, Hengl and Heuvelink 2007, Kumar *et al.* 2012), multiple linear regression (Meersmans *et al.* 2008), generalized linear models (Yuanhe *et al.* 2008), linear mixed models (Doetterl *et al.* 2013, Karunaratne *et al.* 2014). Recently, a few studies have also used cutting-edge techniques from the field of machine learning, such as artificial neural networks (Malone *et al.* 2009, Jaber and Al-Qinna 2011, Li *et al.* 2013) , support vector machines (Rossel and Behrens 2010), boosted regression trees (Martin *et al.* 2011), Cubist (Kumaraperumal *et al.* 2022, Kumar *et al.* 2023, Meliho *et al.* 2023) and random forests (Grimm *et al.* 2008, Wiesmeier *et al.* 2011, Vagen and Winowiecki 2013) to prepare spatial maps of SOC. Machine learning techniques address the limitations of parametric and non-parametric statistical techniques (Drake *et al.* 2006). Thus, goal of this research is to create and assess machine learning

[1]Indian Institute of Information Technology, Nagpur, Maharashtra. *Corresponding author email: mundadasg30@gmail.com

models that predict and map variations of SOC stocks in the Dhamtari District of Chhattisgarh State.

MATERIALS AND METHODS

Research terrain: The present study was carried out during 2022–2024 at Indian Institute of Information Technology, Nagpur, Maharashtra. To anticipate SOC, a set of attributes representing terrain, climate and spectral indices was chosen.

Data processing: Numerous biological and environmental factors, as well as the associations between them, influence the amount of nutrients in the soil. A set of factors encompassing topography, climate, and remote sensing were chosen to forecast soil properties. For this research, soil health card data were used. Block-by-block matching of the locations was done, and any missing or incorrectly valued data was eliminated. Multitemporal information was retrieved from SRTM DEM and Landsat-8 (Roy *et al.* 2014) images collected from the USGS/NASA. Climate data at a resolution of 21 km^2 was acquired from WorldClim 2.1 spanning more than 20 years. 17 terrain variables, 19 bioclimate variables and 7 soil-related spectral indices, were retrieved using set of pre-processed raster images (Table 1). The SRTM DEM was used to extract topographical data with a spatial resolution of 30 m. To match the digital elevation model's (DEM) resolution, climate data were interpolated to 30 m resolution. The SAGA GIS tool was utilized to calculate the bioclimate variables and terrain variables.

Modelling techniques: It has been observed that most ML models used for experimentation for such research problems suffer from overfitting and need high computation time. Two machine learning algorithms have been used, namely Multi-Layer Perceptron (MLP) used and second one, Boosted Regression Tree (BRT). To overcome the overfitting problem, in this study a technique called early stopping has been used. It is an optimization strategy which reduces overfitting without affecting model accuracy. It is primarily about terminating training before a model becomes overfit. Summary for the used ML models is mentioned below.

Table 1 List of predictor variables for modelling

Category	Predictor Variables
Topography	Plan Curvature, Flow Accumulation, Topographic Position Index, Aspect, Channel Network Base Level, Total Catchment Area, Elevation, Multi-Resolution Ridge Top Flatness, Channel Network Distance, Slope, Terrain Ruggedness Index, Valley Depth, Convergence Index, Terrain Wetness Index, Profile Curvature, Multi Resolution Valley Bottom Flatness, Relative Slope Position
Climate	Bio-Climate1 to Bio-Climate19
Spectral Indices	Saturation Index (SI) (Raya *et al.* 2004), Atmospherically Resistant Vegetation Index (ARVI) (Kaufman and Tanre 1992), Normalized Difference Vegetation Index (NDVI) (Huete *et al.* 2002), Coloration Index (CI) (Raya *et al.* 2004), Brightness Index (BI) (Raya *et al.* 2004), Crust Index (CrI) (Karnieli 1997), Soil Adjusted Vegetation Index (SAVI)

Two techniques are combined, namely Boosting and Decision Trees algorithms to create Boosted Regression Trees (BRT) models. BRT, a tree-based algorithm was designed by (Friedman *et al.* 2000) and uses boosting to enhance accuracy. Instead of obtaining a single, highly accurate model, boosting relies on merging multiple approximation prediction models (Schapire 2003). As a result, the decision trees grow successively so that each one forecasts the residual of the one before it; as a result, the algorithm's performance is affected by the number of trees and needs to be adjusted. However, the trees are developed on a randomly chosen data subset with no replacement to introduce randomness into the model and hence boost the robustness of performance (Friedman 2002). The learning rate also referred to as shrinkage regulates each new tree's contribution to the final model (Hastie *et al.* 2009).

A specific branch of artificial intelligence that is frequently utilized for modelling is artificial neural networks. One kind of neural network comprised of Multilayer Perceptron (Gardner and Dorling 1998). It is a model made up of a network of fundamentally connected neurons, or nodes, that shows a non-linear relation between an input vector and an output vector. Every node in the layer was linked to every other node in the layer preceding it. Strengths and weights may be symmetrical or non-divergent for each node in a network, data enters the input layer and progresses progressively through each layer to the last layer, i.e. output layer. The architecture of a MLP can vary, although it usually has multiple layers of neurons. Just the input vector is sent to the network by the input layer; no computations are performed there. Fig. 1 depicts architecture for MLP considered in this study, here n1=200, n2=150, n3=100, n4=50, n5=10 tells us about size of hidden layers, respectively. The input layer has 7 nodes which are components derived from SVD, and output layer consists of one node as SOC prediction value.

Hyperparameter setting: The potential of hyperparameters to directly regulate the training algorithm's behaviour contributes to their significance. The choice of appropriate hyperparameters has a significant effect on the training model's performance. In this research work a set of hyperparameters have been considered which are derived using optimization technique called GridSearch. It is perhaps the most straightforward traditional approach to carrying out hyperparameter optimisation (Shekar and Dagnew 2019). It generates a Cartesian product of every possible combination of hyperparameters. Grid Search trains the machine learning algorithm for every feasible combination of hyperparameters and tests performance using the "cross-validation" technique on the training set.

Model evaluation: We utilized root mean square error, RMSE (Willmott and Matsuura 2005) and coefficient of determination, R^2 (Wright 1921) as the performance indicator. Also, computation time (CT), the total amount

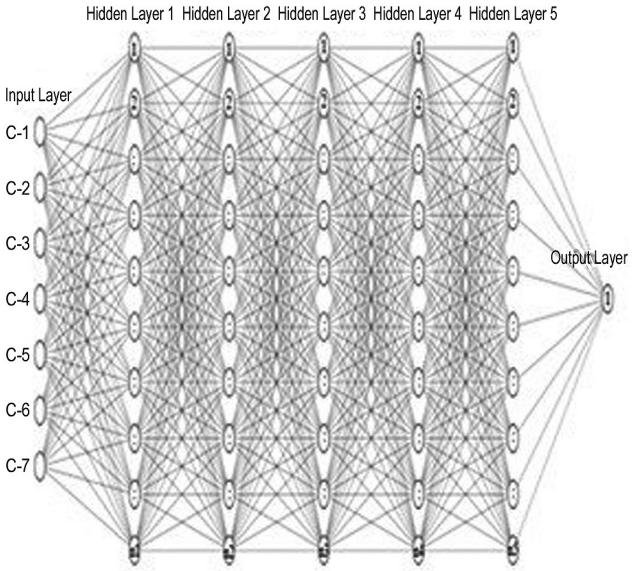

Fig. 1 Architecture of MLP.

of time needed to finish the training process, was measured as a metric for model efficiency.

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(P_i - O_i)^2}{\sum_{i=1}^{m}(\bar{O} - O_i)^2} d$$

$$RMSE = \sqrt{\frac{1}{m}\left(\sum_{i=1}^{m}(P_i - O_i)^2\right)}$$

RESULTS AND DISCUSSION

Prior to prediction, the dimensions of satellite data to be reduced using a feature reduction technique. As a large quantity of features for any model can increase prediction error, dimensionality reduction technique has the potential to improve prediction accuracies, shorten processing times needed to complete a prediction, and enable the removal of noise. A statistically more stable approach is to use the technique known as singular value decomposition (SVD) demonstrated in (Golub and Reinsch 1970, Danaher and O'Mongain 1992). The SVD is a strong contender for feature reduction because of the inherent correlations that exist in nature. Furthermore, the scales of the singular values that the SVD reveals will demonstrate if there is no conceivable reduction. Using the SVD method, any matrix can be factored into three new matrices with unique properties that can be used further. SVD of a linear transformation S is written as:

$$S = U\Sigma V^T \quad (1)$$

Where U, Orthogonal matrix of size M × M and referred as left singular vectors of S, will M × N diagonal matrix in which diagonal elements are termed as singular values of S and V^T is orthogonal right singular vectors of S having size of N × N elements. Expanded version of linear transformation S, in SVD is as shown below, considering M<N:

$$\begin{bmatrix} s_{11} & \cdots & s_{1N} \\ \vdots & \ddots & \vdots \\ s_{M1} & \cdots & s_{MN} \end{bmatrix} = \begin{bmatrix} u_{11} & \cdots & u_{1M} \\ \vdots & \ddots & \vdots \\ u_{M1} & \cdots & u_{MM} \end{bmatrix} \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1N} \\ \vdots & \ddots & \vdots \\ \sigma_{M1} & \cdots & \sigma_{MN} \end{bmatrix} \begin{bmatrix} v_{11} & \cdots & v_{1N} \\ \vdots & \ddots & \vdots \\ v_{N1} & \cdots & v_{NN} \end{bmatrix} \quad (2)$$

Principal Components were considered with cumulative variance in the range of 90–99%. Correlation coefficients are as depicted in Fig. 2. It was identified that all the vegetation indices except Crust Index have good correlation coefficient in all components. Also, Bio-climate variables from Bio14 to Bio19 and terrain attributes namely Aspect,

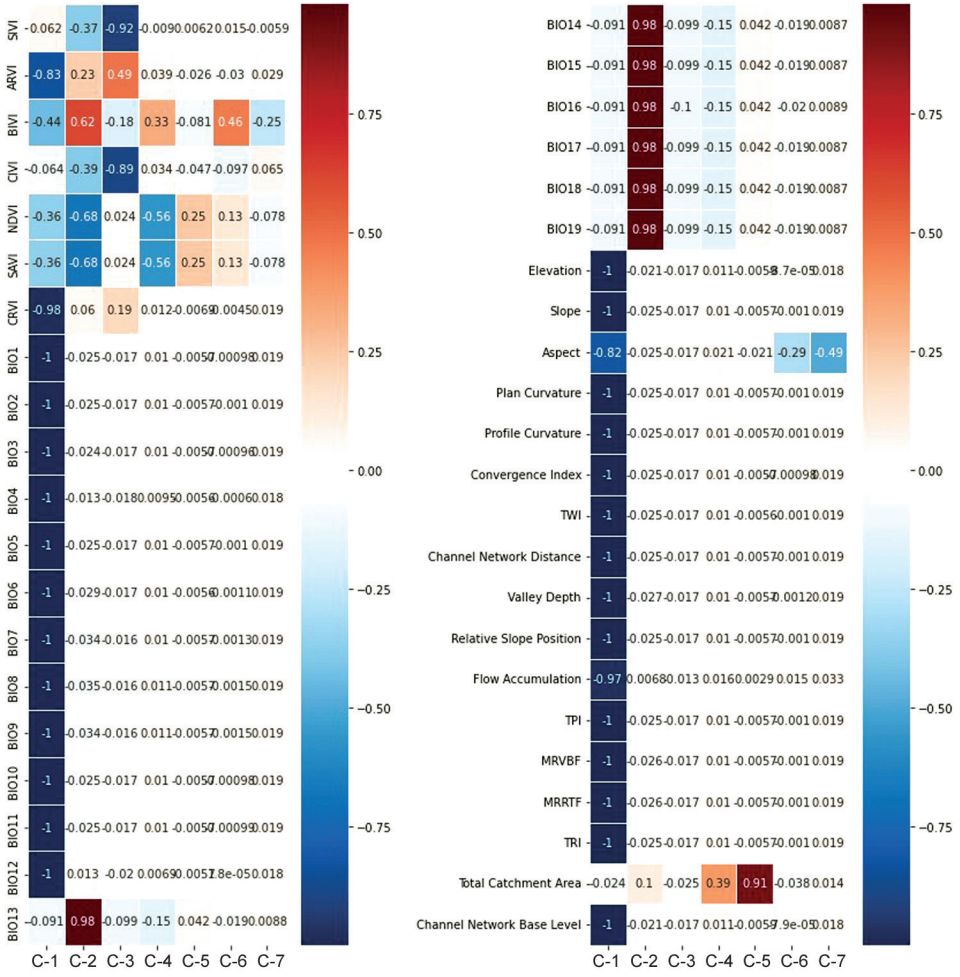

Fig. 2 Feature correlation analysis between principal components derived and environmental covariates.

Total Catchment Area and Channel Network Base Level have good contribution in prediction of SOC for the study area. On statistical analysis, it was observed that SOC of study location ranged from 0.06–1.76, having mean of 0.44 and a standard deviation of 0.24, presenting a positive-skewed distribution with value as 0.8365. Skewness in the values of a specific independent variable (feature) degrades model performance. That's why skewness of data has been reduced by applying logarithmic transformation. After the transformation, skewness of 0.4290 was noted.

The performance of BRT, BRT_ES, MLP, MLP_ES was evaluated with the help of testing datasets and validation datasets. Table 2 shows the implementation outcomes of all ML models. Using both training and testing datasets, the effectiveness of four models for predicting the SOC in the Dhamtari District Indian state of Chattisgarh was assessed. During the study, experimentation was carried out with multiple sets of principal components as input, which were generated using SVD algorithm. Early stopping technique (Stankewitz 2024) used to overcome the problem of overfitting. It proved its significance prominently in BRT than MLP. In line with the observations depicted in Table 2, modelling technique with input as 7 components proved more accurate as compared to others. On comparing ML models based on computation time, it is found that both algorithms using early stopping technique acquired very little time. For the training dataset, it was observed that Boosted Regression Tree algorithm showed the highest R^2 as 0.9884 while MLP recorded lowest R_2 as 0.0670. Based on RMSE, Boosted Regression Tree (BRT) recorded the lowest RMSE as 0.0175 only for Training Dataset while Multilayer Perceptron (MLP) showed good result for both datasets without suffering from overfitting i.e. can be considered as a good fit model. Low R^2 may result from high irregularity in climate variables and spectral indices caused by the geological conditions in the study. Soil nutrient status is currently determined via laboratory-based chemical analysis. This soil evaluation method is based on routine soil sampling design, sample collection, sample preparation, and subsequent laboratory chemical analysis. However, evaluating soil across a wide region using this method is expensive, time-consuming, and labour-intensive. Furthermore, if handled incorrectly, the laboratory's acid-base waste liquid may result in secondary environmental contamination. Thus, a quick, on-site, ongoing, and non-polluting detection technique mentioned in this work for soil composition detection is very much useful and desperately needed.

Spatial predictions of SOC, generated by each of the machine learning models used in this study are displayed in Fig. 2 for the whole Dhamtari district. All the prediction model maps displayed both rapid and regular fluctuations throughout the research region. For the BRT, BRT_ES, MLP, and MLP_ES models, estimated SOC ranged from -0.002 to 0.71, 0.09 to 0.61, -0.36 to 1.72, and -0.037 to 0.85, in that order. It is difficult to choose the most accurate model in the absence of individualistically validating these predictions, we selected the Multi-Layer Perceptron with Early Stopping (MLP_ES) model as the "best" due to accuracy metrics and the fact that the spatial projections visually matched our perception of the scene. Multi-Layer Perceptron with Early Stopping (MLP_ES) model predictions more closely matched the geo-graphical distribution of the SOC we expected in the research area. The study area's northern and some part of eastern region have minor values due to intensive farming, which causes significant wearing away and crop cultivation, while the western and southern regions, which are covered with good vegetation and dominated by forests, have high values of SOC in all models. The various management strategies used in regions with higher concentrations of intensive farming can be used to clarify it.

Table 2 Model assessment results based on training data and testing data

ML model	Principal components	CT (sec)	Training		Testing	
			R^2	RMSE	R^2	RMSE
BRT	3	21.49	0.9884	0.0175	-0.3676	0.1958
	4	28.32	0.9884	0.0175	-0.4089	0.1886
	7	42.76	0.9884	0.0175	-0.2418	0.1771
BRT_ES	3	0.91	0.6684	0.0941	-0.1937	0.1746
	4	1.53	0.6981	0.0898	-0.2009	0.1741
	7	2.06	0.7480	0.0820	-0.1269	0.1687
MLP	3	14.01	0.0741	0.1567	-0.0476	0.1648
	4	14.37	0.0743	0.1567	-0.4078	0.1911
	7	14.45	0.2391	0.1421	-0.0898	0.1681
MLP_ES	3	1.17	-0.0009	0.1630	0.0078	0.1608
	4	1.09	0.0171	0.1615	0.0098	0.1604
	7	0.84	0.0032	0.1618	-0.1887	0.1601

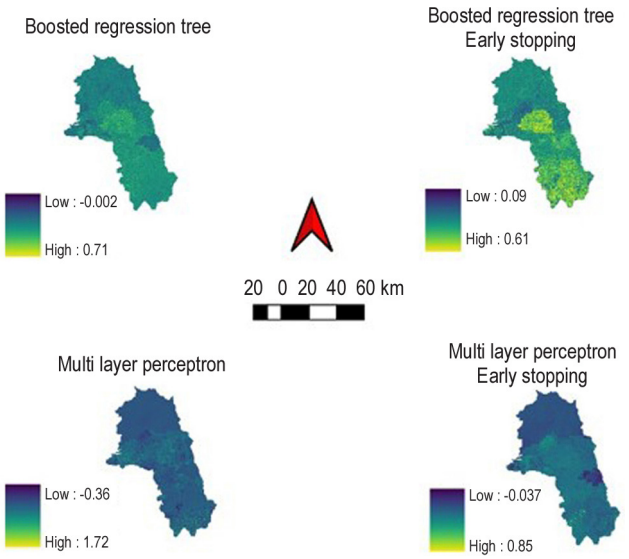

Fig. 2 Spatial prediction of SOC using different models.

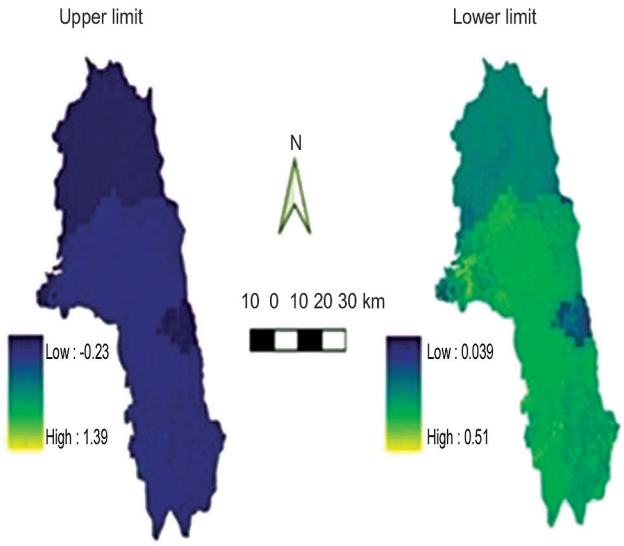

Fig. 3 Dispersion of the 90% prediction interval's low and high bounds for estimated SOC.

A certain amount of uncertainty is unavoidable in machine learning models. There are various methods for measuring this uncertainty. Quantile Regression technique was utilized to quantify this prediction uncertainty. The uncertainty analysis for our best model, Multi-Layer Perceptron with Early Stopping (MLP_ES), is shown in Fig. 3. The tendency of the machine learning algorithms' propensity to forecast the SOC was validated by the uncertainty analysis. A confidence interval of 90% with both lower and higher prediction bounds have been constructed to show the level of uncertainty. The maximum predicted SOC values were found to lie within the 90% confidence interval. In Supplementary Fig. 1, just a sample of the dataset covering the full research area has been presented for presentation because of its vast size. The equilibrium between carbon inputs and outputs in soils determines SOC concentrations; this equilibrium is influenced by several factors, including regional features like vegetation, topography, and environmental circumstances (Sahoo *et al.* 2019). The uncertainty map revealed that the southern portion, which is mostly covered in healthy vegetation and is dominated by forests, had stronger SOC estimates than the northern parts.

In conclusion, the results of the study revealed how successful digital soil-mapping techniques are at generating accurate soil-related data, such as details on soil nutrients. Before planting in each cycle, farmers evaluate the quality of the soil (nutrient contents) to choose the best agricultural management strategy for the soil's current state. However, fertilization is usually done mechanically or involuntarily to produce a high yield, which leads to an unequal distribution of chemical fertilizers. However, scientific fertilization done in line with the amount and quality of nutrients in the soil produces high-yield crops of superior quality. Therefore, determining the composition of the soil has become essential for fertilization in precision agriculture. High-resolution digital maps may be used to assess crop compatibility, soil and land management strategies, site-specific fertilizer recommendations, and irrigation scheduling all of which can save operational costs. Fewer input costs and increased farm outputs can be achieved by strategic crop selection, effective landscape and soil management practices, and balanced fertilization that use DSM. Two machine-learning methods were investigated in this work. Patterns of SOC spatial distribution in Dhamtari District of Indian state Chattisgarh. This study used data associated with SOC observations, environmental parameters, and optimal models to determine the spatial map of the SOC. BRT produced good results for training dataset, with highest R^2 as 0.9884 and lowest RMSE as 0.0175 but for testing dataset R^2 was noted very low. A similar pattern was observed with BRT_ES variant. MLP and MLP_ES showed better results for both training and testing datasets. MLP_ES is confirmed as our best model because of its computing time and lowest RMSE for testing dataset as 0.1601.

REFERENCES

Bationo Andre, Job Kihara, Bernard Vanlauwe, Boaz Waswa and Joseph Kimetu. 2007. Soil organic carbon dynamics, functions and management in West African agro-ecosystems. *Agricultural Systems* **94**: 13–25.

Batjes N H. 1996. Total carbon and nitrogen in the soils of the world. *European Journal of Soil Science* **47**: 151–63.

Brady, Nyle C and Ray R Weil. 2008. *The Nature and Properties of Soils*, Vol. 13. Prentice Hall Upper Saddle River, New Jersey.

Cambule A H, Rossiter D G, Stoorvogel J J and Smaling E M A. 2014. Soil organic carbon stocks in the Limpopo National Park, Mozambique: Amount spatial distribution and uncertainty. *Geoderma* **213**: 46–56.

Danaher S and E O'Mongain. 1992. Singular value decomposition in multispectral radiometry. *International Journal of Remote Sensing* **13**: 1771–77.

Doetterl Sebastian, Antoine Stevens, Kristof van Oost, Timothy A Quine and Bas van Wesemael. 2013. Spatially-explicit regional-

scale prediction of soil organic carbon stocks in cropland using environmental variables and mixed model approaches. *Geoderma* **204–05**: 31–42.

Drake, John M, Randin, Christophe, Guisan and Antoine. 2006. Modelling ecological niches with support vector machines. *Journal of Applied Ecology* **43**: 424–32.

El-Ramady, Hassan R, T A Alshaal, M Amer, E Domokos-Szabolcsy, N Elhawat, J Prokisch, and M Fári. 2014. Soil quality and plant nutrition. *Sustainable Agriculture Reviews Agroecology and Global Change,* pp. 345–447. Springer.

Friedman Jerome H. 2002. Stochastic gradient boosting. *Computational Statistics and Data Analysis* **38**: 367–78.

Friedman, Jerome, Trevor Hastie and Robert Tibshirani. 2000. Additive logistic regression: A statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics (Institute of Mathematical Statistics)* **28**: 337–407.

Gardner M W and Dorling S R. 1998. Artificial neural networks (the multilayer perceptron)–A review of applications in the atmospheric sciences. *Atmospheric Environment* **32**: 2627–36.

Golub G H and C Reinsch. 1970. Singular value decomposition and least squares solutions. *Numerische Mathematik* **14**: 403–20.

Grimm R, T Behrens, M Märker and H Elsenbeer. 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island-Digital soil mapping using Random Forests analysis. *Geoderma* **146**: 102–13.

Hastie Trevor, Robert Tibshirani and Jerome Friedman. 2009. *Basis Expansions and Regularization in the Elements of Statistical Learning: Data Mining, Inference, and Prediction*, pp. 139–89. Springer New York.

Hengl Tomislav, Gerard B M, Heuvelink G B and Alfred Stein. 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma* **120**: 75–93.

Hengl, Heuvelink G B and David G Rossiter. 2007. About regression-kriging: From equations to case studies. *Computers and Geosciences* **33**: 1301–15.

Huete, Alfredo, Kamel Didan, Tomoaki Miura, E Patricia Rodriguez, Xiang Gao and Laerte G Ferreira. 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment* **83**: 195–213.

Jaber Salahuddin M and Mohammed I Al-Qinna. 2011. Soil organic carbon modeling and mapping in a semi-arid environment using thematic mapper data. *Photogrammetric Engineering and Remote Sensing* **77**: 709–19.

Karnieli Arnon. 1997. Development and implementation of spectral crust index over dune sands. *International Journal of Remote Sensing* **18**: 1207–20.

Karunaratne S B, T F A Bishop, J A Baldock and I O A Odeh. 2014. Catchment scale mapping of measureable soil organic carbon fractions. *Geoderma* **219**: 14–23.

Kaufman J and Didier Tanre. 1992. Atmospherically resistant vegetation index (ARVI) for EOS-MODIS. *IEEE Transactions on Geoscience and Remote Sensing* **30**: 261–70.

Kumar A, Moharana P C, Jena R K, Malyan S K, Sharma G K, Fagodiya R K, Shabnam A A, Jigyasu D K, Kumari K M V and Doss S G. 2023. Digital mapping of soil organic carbon using machine learning algorithms in the upper Brahmaputra valley of north-eastern India. *Land* **12**.

Kumar Sandeep, Lal Rattan and Desheng Liu. 2012. A geographically weighted regression kriging approach for mapping soil organic carbon stock. *Geoderma* **189**: 627–34.

Kumaraperumal R, Pazhanivelan S, Geethalakshmi V, Nivas Raj M, Muthumanickam D, Kaliaperumal R, Shankar V, Nair A M, Yadav M K and Tarun Kshatriya T V. 2022. Comparison of machine learning-based prediction of qualitative and quantitative digital soil-mapping approaches for eastern districts of Tamil Nadu, India. *Land* **11**(12): 2279.

Lefevre C, Rekik F, Alcantara V and Wiese L. 2017. *Soil Organic Carbon: The Hidden Potential*. Food and Agriculture Organization, United Nations.

Li Qi-quan, Tian-xiang Yue, Chang-quan Wang, Wen-jiang Zhang, Yong Yu, Bing Li, Juan Yang and Gen-chuan Bai. 2013. Spatially distributed modeling of soil organic matter across China: An application of artificial neural network approach. *Catena* **104**: 210–18.

Malone B P, A B McBratney, B Minasny and G M Laslett. 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma* **154**: 138–52.

Martin M P, M Wattenbach, P Smith, J Meersmans, C Jolivet, L Boulonne and D Arrouays. 2011. Spatial distribution of soil organic carbon stocks in France. *Biogeosciences* **8**: 1053–65.

Meersmans J, F De Ridder, F Canters, S De Baets and M Van Molle. 2008. A multiple regression approach to assess the spatial distribution of Soil Organic Carbon (SOC) at the regional scale (Flanders, Belgium). *Geoderma* **143**: 1–13.

Meliho M, Boulmane M, Khattabi A, Dansou C E, Orlando C A, Mhammdi N and Noumonvi K D. 2023. Spatial prediction of soil organic carbon stock in the Moroccan high atlas using machine learning. *Remote Sensing* **15**.

Raya S S, J P Singhb, Gargi Dasa and Sushma Panigrahyb. 2004. Use of high resolution remote sensing data for generating site-specific soil management plan. *Red* **550**: 727.

Rossel R A Viscarra and T Behrens. 2010. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* **158**: 46–54.

Roy, David P, Michael A Wulder, Thomas R Loveland, Curtis E Woodcock, Richard G Allen, Martha C Anderson and Dennis Helder. 2014. Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment* **145**: 154–72.

Sahoo Uttam Kumar, Soibam Lanabir Singh, Anudip Gogoi, Alice Kenye and Snehasudha S Sahoo. 2019. Active and passive soil organic carbon pools as affected by different land use types in Mizoram north-east India. *PLOS one* **14:** 1–16.

Sarkar S, De B, Das P, Saha D, Awasthi D P, Paul N and Sen D. 2025. Nutrient management approach to improve productivity and profitability of pigeonpea (*Cajanus cajan*) in north-east hill zones of India. *The Indian Journal of Agricultural Sciences* **95**(4): 400–05.

Schapire Robert E. 2003. *The Boosting Approach to Machine Learning: An Overview. In Nonlinear Estimation and Classification*, pp. 149–71. David D Denison, Mark H Hansen, Christopher C Holmes, Bani Mallick and Bin Yu (Eds). Springer, New York.

Shekar B H and Guesh Dagnew. 2019. Grid search-based hyperparameter tuning and classification of microarray cancer data. *Second International Conference on Advanced Computational and Communication Paradigms,* pp. 1–8.

Stankewitz Bernhard. 2024. Early stopping for L^2- boosting in high-dimensional linear models. *The Annals of Statistics* **52**(2): 491–518.

Vagen Tor-Gunnar and Leigh A Winowiecki. 2013. Mapping of soil organic carbon stocks for spatially explicit assessments of climate change mitigation potential. *Environmental Research*

Letters (IOP Publishing) **8**: 015–011.

Wiesmeier Martin, Frauke Barthold, Benjamin Blank and Ingrid Kogel-Knabner. 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. *Plant and Soil* **340**: 7–24.

Willmott Cort J and Kenji Matsuura. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* **30**: 79–82.

Wright Sewall. 1921. Correlation and causation. *Journal of Agricultural Research* **20**(7): 557.

Yang Yuanhe, Jingyun Fang, Yanhong Tang, Chengjun Ji, Chengyang Zheng, Jinsheng He and Biao Zhu. 2008. Storage, patterns and controls of soil organic carbon in the Tibetan grasslands. *Global Change Biology* **14**(7): 1592–99.

Zhenxing Bian, Shuai Wang, Qianlai Zhuang Xinxin Jin Qiubing Wang and Shuhai Jia. 2020. Applying statistical methods to map soil organic carbon of agricultural lands in northeastern coastal areas of China. *Archives of Agronomy and Soil Science* **66**: 532–44.