OPEN ACCESS
PAPER

Mongolian Journal of Geography and Geoecology - MJGG

Received: 16 May 2025; Accepted: 27 June 2025; Published online: 1 September 2025
https://doi.org/10.5564/mjgg.v62i46.4161

Mapping of soil organic carbon content using machine learning algorithms in Bayanzurkh soum

***Corresponding author**
Maralmaa Ariunbold
maralmaa_a@mas.ac.mn

CITATION
Maralmaa A, Saruulzaya A, Purevdulam Y, Dawaadorj D (2025) Mapping of Soil Organic Carbon Content Using Machine Learning Algorithms in Bayanzurkh Soum. *Mongolian Journal of Geograhy and Geoecology,* 62(46), 1–7.
https://doi.org/10.5564/mjgg.v62i46.4161

Maralmaa Ariunbold[1,2*], Saruulzaya Adiya[1], Purevdulam Yondonrentsen[1], Dawaadorj Dawaasuren[1,2]

[1]*Institute of Geography and Geoecology, Mongolian Academy of Sciences, Ulaanbaatar 15170, Mongolia*
[2]*School of Arts and Sciences, National University of Mongolia, Ulaanbaatar 210646, Mongolia*

ABSTRACT

Soil organic carbon (SOC) is the largest carbon reservoir in the terrestrial ecosystem and plays an important role in the global carbon cycle. Consequently, even a slight change in SOC content due to land use, soil management, or rates of soil erosion can considerably increase atmospheric CO2 concentrations. The main purpose of this study is to predict and map SOC content in small area by applying machine learning (ML) algorithms using field measurements and remote sensing data. We used to three different algorithms such as Random Forest (RF), Extreme Gradient Boosting (eXGB), and Gradient Boosted Regression (GBR) of ML. According to field work, 123 soil samples (0–30 cm) were collected from Bayanzurkh soum in Khuvsgul, and 26 variables were used to predict SOC content. As shown the prediction results, the GBR algorithm demonstrated the highest performance, explaining 78% of the variation in soil SOC content, with an RMSE of 42.9 g/kg and an MAE of 33.1 g/kg. The ranking of model performance in terms of prediction accuracy was GBR > eXGB > RF. Therefore, we found a strong relationship (R^2 = 0.94) between the predicted and measured values based on linear regression analysis. The most influential predictor variables were SILT (13.6%), CLAY (7.8%), NDVI (7.3%), and SOLAR RADIATION (6.3%). These results demonstrate that SOC content can be effectively predicted using machine learning algorithms. However, it is advisable to compare the performance of multiple algorithms and select the most suitable approach for the small area.

KEYWORDS

Soil organic carbon content, Machine learning

1. INTRODUCTION

Soils contain approximately 1,550 petagrams (Pg) of carbon globally and represent the largest terrestrial reservoir of carbon [1]. This amount is 2 to 3 times greater than the carbon stored in the atmosphere and vegetation [2]. On average, 33% of total soil organic carbon (SOC) stocks are stored below 30 cm depth [3]. Even minor changes in terrestrial carbon stock could have significant impact on global warming [4]. SOC stock is a critical function of soil, influencing climate regulation and other soil functions [5] An increase in SOC content can enhance soil water retention, fertility, and structure; promote plant growth; and improve the diversity and activity of soil microorganisms. [6].

Determining the spatial distribution of SOC provides crucial information for mitigating global climate change. [7]. Numerous methods have been employed to model and map SOC content and stocks in recent years at various scales. However, a universally optimal method for predicting the spatial distribution of SOC has yet to be established, as it depends heavily on the spatial variability of soil properties, climatic conditions, and land-use management practices [8], [9], [10]. SOC mapping methodologies typically integrate three fundamental approaches: conventional soil survey techniques, geostatistical interpolation methods, and machine learning-based predictive modeling [11], [12], [13]. Machine learning algorithms exhibit superior capabilities in data processing and predictive model optimization [14]. Comparative studies indicate that tree-based ensemble methods, including random forest (RF), extreme gradient boosting (eXGB), and gradient boosting regression (GBR), consistently outperform alternative approaches in SOC stock prediction accuracy [11], [15], [16].

SOC is strongly influenced by multiple environmental factors, including vegetation characteristics, climate, topography, and physico-chemical soil properties [11]. The accurate identification and selection of these key factors critically determine the reliability of SOC mapping results. In recent years, international research on SOC mapping has increasingly utilized remote sensing data, integrated with machine learning algorithms. However, previous researchers have evaluated SOC mapping by using traditional, geostatistical methods in Mongolia. Thus, we proposed to evaluate and compare predictive machine learning algorithms, such as RF, eXGB, and GBR, for modeling and forecasting SOC.

2. RESEARCH METHODS

2.1. Field Sampling and Laboratory Analysis

Field sampling was conducted in Bayanzurkh soum, Khuvsgul Province, during the 2022–2023 period. A total of 123 soil samples were randomly collected from the topsoil layer (0–30 cm) across the study area. This depth was selected because it represents the primary zone of organic matter accumulation due to inputs from plant residues and biological activity, making it the most suitable depth for assessing SOC content.

SOC was determined using the loss-on-ignition (LOI) method. For analysis, samples were first oven-dried at 105°C for 6 hours to remove moisture, then combusted at 360°C for 2 hours to estimate organic matter loss, which was used to calculate SOC content.

2.2. Environmental variables

We selected 26 SOC content variables based on their impact as soil-forming factors. Soil characteristics variables, including bulk density (BULK), clay content (CLAY), and silt content (SILT), were obtained from the SoilGrids 2.0 database developed by ISRIC–World Soil Information. These variables were processed and aggregated to represent the 0-30 cm soil depth. For climatic variables, we utilized spatial datasets from the WorldClim 2.1 database, specifically the long-term mean annual temperature (MAAT) and mean annual precipitation (MAP) for the period 1970–2000.

To characterize vegetation and moisture conditions, Landsat 8 OLI/TIRS Surface Reflectance were masked with a 20 percent cloud cover image from June to August (2015–2023) were used to calculate the average values of the following spectral bands: RED, NIR, SWIR1, SWIR2, and TIRS1. Based on these bands, a set of spectral indices was derived, including the Normalized Difference Vegetation Index (NDVI), Normalized Difference Moisture Index (NDMI), Normalized Difference Water Index (NDWI), Bare Soil Index (BSI), Soil Brightness Index (BI), and Redness Index (RI). All satellite data processing and index calculations were conducted using the Google Earth Engine (GEE) platform. In addition, spatial information on forest cover and surface water was incorporated: the 2022 global forest cover data at 30 m resolution was sourced from the Hansen Global Forest Change database, while the 2020 surface water occurrence data at 30 m resolution was obtained from the JRC Global Surface Water Explorer v1.4. Topographic analysis was carried out using Shuttle Radar Topography Mission (STRM) digital elevation data with a spatial resolution of 90 m.

We used digital elevation data ArcGIS 10.8 software to calculate 8 terrain indices: the digital elevation model (DEM), slope, aspect, solar radiation, topographic wetness index (TWI), topographic position index (TPI), and multi-resolution valley bottom flatness (MRVBF) and ridge top flatness (MRRTF) indices. In this study, we utilized data with various spatial resolutions ranging from 30 m to 1000m. Due to this variability, we resampled spatial resolutions to 250 m with the ArcGIS 10.8 software.

2.3. Modeling Techniques

2.3.1. Random Forest (RF)

The RF algorithm is a widely used machine learning algorithm for regression analysis [17]. This ensemble method aggregates predictions from multiple decision trees to produce robust regression outputs. Random forests can handle not only nonlinear relationships between independent variables and dependent variables but also a large number of covariates [18].

2.3.2. Gradient Boosting Regression (GBR)

The GBR algorithm is a type of boosting technique in which multiple weak learners are generated sequentially. Each weak learner is trained to approximate the negative gradient of the loss function concerning the current ensemble model. By adding each new learner in the direction of the negative gradient, the overall loss of the ensemble model is iteratively reduced, thereby improving predictive performance [19].

2.3.3. eXtreme Gradient Boosting (eXGB)

The eXGB is an advanced and highly efficient machine learning algorithm within the gradient boosting framework [20]. It provides substantial improvements in both computational performance and predictive accuracy compared to traditional gradient boosting methods. The core principle involves training a series of weak learners typically decision trees in an iterative manner. In each iteration, eXGB updates the sample weights based on the errors of the preceding model, thereby directing greater attention to previously misclassified or poorly predicted samples. This iterative reweighting enables the model to progressively reduce error and enhance overall performance [21]. Figure 1 shows the conceptual framework of this study.

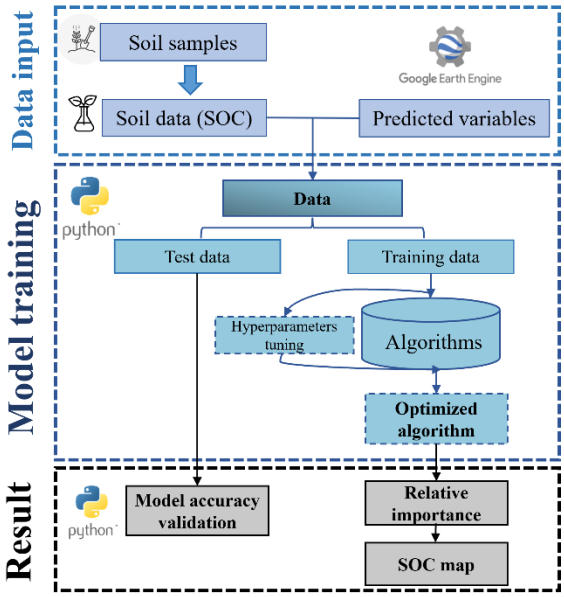

Figure 1. Methodological flowchart for modeling and mapping SOC in the study area

The issue of correlation between the variables used in SOC mapping is important. Consequently, we employed variogram analysis to evaluate the spatial relationships of these variables. Also, we employed the Recursive Feature Elimination based on Random Forest (RFE-RF) method to reduce model complexity and computational costs by eliminating redundant and irrelevant variables. RFE is a widely used feature selection technique that identifies the most informative predictors by recursively ranking feature importance and removing those with the least contribution to model performance [22]. In this process, we repeatedly selected all possible combinations of variables, considering subsets of 10 to 26. In this study, Python v3.11.7 programming language was used for machine learning modeling, prediction, and feature importance analysis. A range of accompanying libraries and packages were employed, including NumPy, pandas, Matplotlib, scikit-learn, SciPy, Rasterio, EarthPy, and GDAL.

2.3.4. Model performance

The original dataset was randomly divided into 85% of points and 15% for testing. After the division, we trained the final model using the training dataset and the selected features from the RFE-RF.

The final model evaluation was carried out on the prediction results using the test dataset expressed in the coefficient of determination (R^2), the root mean square error (RMSE), and the mean absolute error (MAE).

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \underline{y})^2} \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (3)$$

where y_i and $\hat{y}_i$ are the observed and predicted values for test sample i, $\overline{y}$ is the mean of the observed values, and n is the sample size of the test set.

3. RESULT AND DISCUSSION

3.1. Descriptive statistics

The descriptive statistics of the SOC at the 0–30 cm soil depth at the study site are presented in (Table 1). SOC values ranged from 17.2 g/kg to 612.2 g/kg with a mean of 143.3 g/kg, a standard deviation of 123.7 g/kg, and a coefficient of variation of 86.3%. The high coefficient of variation indicates substantial spatial variability in SOC content across the study site. This variability can be attributed to heterogeneous environmental and soil conditions. In particular, differences in land cover types, vegetation productivity, and topographic position contribute to this variation.

Table 1. The descriptive statistics of the SOC content at the 0–30 cm depth

Descriptive statistics	Value
Maximum value	612.2
Minimum value	17.2
Mean value	143.3
Standard deviation	123.7
Coefficient of variation (%)	86.3

In addition, we evaluated the spatial autocorrelation of the sampling points using Moran's I analysis. As a result, the Moran's Index was 0.42 with a p-value of >0.0001. These results indicate that the sampling points exhibit a clustered distribution (Moran's index > 0.1) and strong positive autocorrelation.

3.2. Analysis of the Importance of Variables

The results of the variogram analysis shows the level of spatial autocorrelation between pixel values (Figure 2). This confirms the presence of spatial structure in the data. The variogram plot shows that the values of neighboring pixels are more similar to each other, while the difference in values (semivariance) increases with distance.

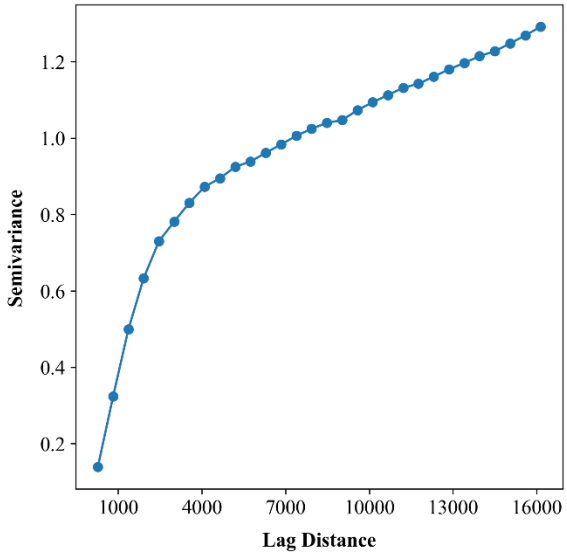

Figure 2. Variogram analysis of variable raster data

The result of the feature selection depends on the geographic location and distance of each measured point; the contribution and significance of the variables to the model varied. Based on a 5-fold cross-validation prediction error, multiple RF models were fitted repeatedly using all possible combinations of higher ranked variables considering the subsets: 1 to the 25 (all variables). Consequently, our model showed the peak predictive power with 19 variables. The ranking of variable importance based on the results of the RFE-RF expressed as a percentage (Figure 3). The results of the predictor variable importance analysis showed that silt (13%) and clay (7.6%) had the highest importance in predicting SOC content. These are fundamental components of soil physical structure, directly and indirectly influencing water retention, nutrient accumulation, and the habitat of soil microorganisms. Soils with higher silt and clay content have greater water-holding capacity compared to sandy soils, which in turn positively affects microbial activity, the decomposition of organic matter, and thus carbon sequestration.

In addition, BSI (8.9%) and RI (8.8%) were also identified as important predictors of SOC content. These indices are related to soil color and cover, and have a direct relationship with soil organic content. For example, areas with high BSI values are often associated with lower organic matter accumulation, while RI can reflect soil mineral composition—particularly iron content—potentially influencing the balance between organic and mineral materials. NDVI (8.1%) also had a high importance in predicting SOC content, reflecting the relationship between high plant biomass and the accumulation of root residues and humus.

Lastly, the importance of solar radiation (7.1%) can be explained by its role in regulating microbial activity and plant photosynthesis, ultimately affecting SOC accumulation.

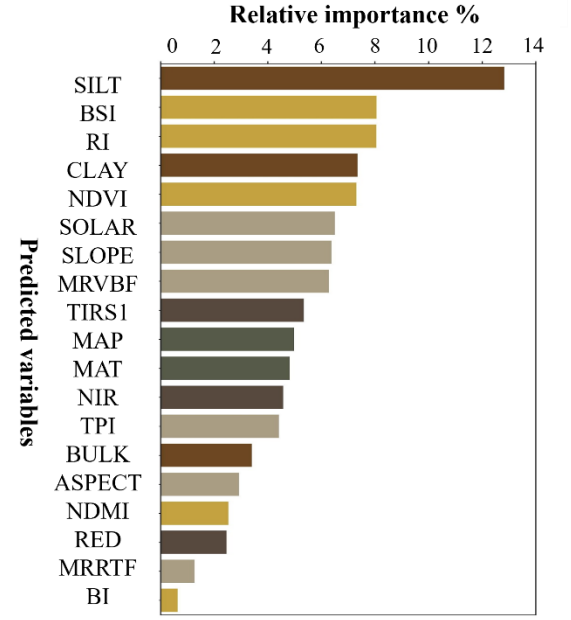

Figure 3. Relative importance of each variable

3.3. Comparison of the Performances of Different Models

The spatial distribution of SOC at the study site was modeled using the RF, GBR, and eXGB algorithms, and their performances were evaluated (Table 2). Subsequently, the hyperparameters of each algorithm were optimized using gridsearch as follows. The GridSearch best results for RF, n_estimators=15; for GBR, n_estimators=33 with a learning rate=0.14; and eXGB, n_estimators=100, learning rate=0.03, reg_alpha=0.75, reg_lambda=0.5, and colsample_bytree=0.9, while other parameters were kept at their default values. Model performance was assessed using the test dataset, where the RF model explained variance in SOC measurements (R^2 = 0.72, the eXGB model explained (R^2 = 0.76), and the GBR model explained (R^2 = 0.78). Among the three algorithms, GBR showed the lowest prediction error with an MAE of 33.1 g/kg and an RMSE of 42.9 g/kg. The overall ranking of model performance was GBR > eXGB > RF.

Following this, a simple linear regression analysis of the model results was carried out, with the test dataset achieving an R^2 of 0.8 and the training dataset reaching an R^2 of 0.94. Overall, all data points yielded an R^2 of 0.94 (Figure 4), indicating that our model reliably predicts SOC content. The mean difference between the measured and predicted values of SOC content is 0.1 g/kg.

Table 2. Model performance to predict SOC content based on the test dataset evaluation

Algorithm	RF	eXGB	GBR
R^2	0.72	0.76	0.78
MAE (g/kg)	37.3	35.1	33.1
RMSE (g/kg)	48.6	45.0	42.9
P value	<0.00001	<0.00001	<0.00001

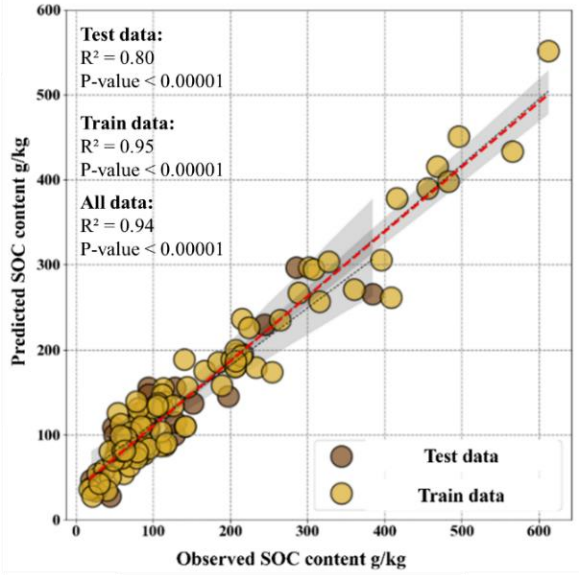

Figure 4. Correlation between predicted and observed SOC content values.

3.4. SOC spatial predictions

We used the GBR algorithm with the best prediction results to generate a model of the spatial distribution of SOC in the study area (Figure 5). According to the spatial distribution map, its content ranges between 16.4 g/kg- 483.3 g/kg and the average value is 175.0 g/kg. SOC content was found to be lower in steep areas and regions heavily impacted by grazing, whereas the highest SOC concentrations were observed in river valleys and areas with dense vegetation cover.

The findings of this study have practical significance for sustainable agriculture. Therefore, priority should be given to protecting river valleys and areas with dense vegetation cover to preserve SOC resources. In addition, it is necessary to prevent soil degradation and implement restoration measures in grazing lands with low SOC content.

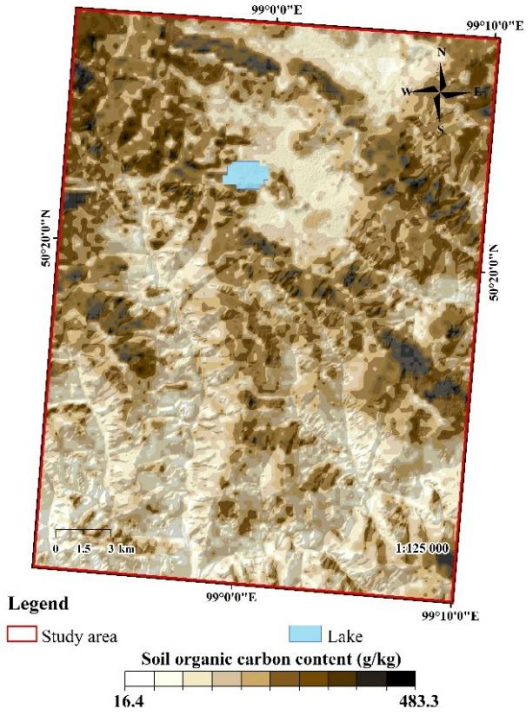

Figure 5. Spatial distribution of the SOC content across study area

4. CONCLUSION

This study evaluated machine learning algorithms for predicting the spatial distribution of SOC content in the study area, identifying the most optimal approach through comparative analysis. Among the machine learning algorithms tested, GBR demonstrated superior predictive performance, achieving 78% accuracy on test data, outperforming both RF (72%) and XGB (76%) in terms of both stability and overall performance. The relative importance of predictive factors was assessed using RFE, revealing (SILT, 13.6%), (CLAY, 7.8%), (NDVI, 7.3%), and (Solar radiation, 6.3%) as the most significant predictors of SOC variability. Overall, the findings confirm that machine learning algorithms can effectively predict and map SOC content.

Furthermore, comparing multiple algorithms and selecting the best-performing model for the specific study area is essential for enhancing the reliability and precision of spatial SOC estimations.

ACKNOWLEDGMENTS

This work was supported by the project "Modelling future trends in GHG emissions due to permafrost degradation". The project number is 2022/154. The authors thank the Mongolian Foundation for Science and Technology for financial support.

REFERENCES

[1] R. La et al., "The carbon sequestration potential of terrestrial ecosystems," *J. Soil Water Conserv.*, vol. 73, no. 6, pp. 1, Nov. 2018. Available: doi: 10.2489/jswc.73.6.145A.

[2] J. P. Scharlemann, E. V. Tanner, R. Hiederer, and V. Kapos, "Global soil carbon: understanding and managing the largest terrestrial carbon pool," *Carbon Manag.*, vol. 5, no. 1, pp. 81–91, Feb. 2014. Available: doi: 10.4155/cmt.13.77.

[3] M. Lacoste et al., "High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape," *Geoderma*, vol. 213, pp. 296–311, Jan. 2014. Available: doi: 10.1016/j.geoderma.2013.07.002.

[4] J. Lin, D. Hui, A. Kumar, Z. Yu, and Y. Huang, "Editorial: Climate change and/or pollution on the carbon cycle in terrestrial ecosystems," *Front. Environ. Sci.*, vol. 11, p. 1253172, Jul. 2023. Available: doi: 10.3389/fenvs.2023.1253172.

[5] M. Wiesmeier et al., "Soil organic carbon storage as a key function of soils - A review of drivers and indicators at various scales," *Geoderma*, vol. 333, pp. 149–162, Jan. 2019. Available: doi: 10.1016/j.geoderma.2018.07.026.

[6] J. Lehmann and M. Kleber, "The contentious nature of soil organic matter," *Nature*, vol. 528, no. 7580, pp. 60–68, Dec. 2015. Available: doi: 10.1038/nature16069.

[7] B. Ren et al., "Comparison of machine learning for predicting and mapping soil organic carbon in cultivated land in a subtropical complex geomorphic region," *Chin. J. Eco-Agric.*, vol. 29, pp. 1042–1050, 2021. Available: doi: 10.13930/j.cnki.cjea.200939

[8] Y. Liu, L. Guo, Q. Jiang, H. Zhang, and Y. Chen, "Comparing geospatial techniques to predict SOC stocks," *Soil Tillage Res.*, vol. 148, pp. 46–58, May 2015. Available: doi: 10.1016/j.still.2014.12.002.

[9] A. Mondal, D. Khare, S. Kundu, S. Mondal, S. Mukherjee, and A. Mukhopadhyay, "Spatial soil organic carbon (SOC) prediction by regression kriging using remote sensing data,"

Egypt. J. Remote Sens. Space Sci., vol. 20, no. 1, pp. 61–70, Jun. 2017. Available: doi: 10.1016/j.ejrs.2016.06.004.

[10] S. Kumar and R. Lal, "Mapping the organic carbon stocks of surface soils using local spatial interpolator," *J. Environ. Monit.*, vol. 13, no. 11, pp. 3128, 2011. Available: doi: 10.1039/c1em10520e.

[11] C. Sothe, A. Gonsamo, J. Arabian, and J. Snider, "Large scale mapping of soil organic carbon concentration with 3D machine learning and satellite observations," *Geoderma*, vol. 405, p. 115402, Jan. 2022. Available: doi: 10.1016/j.geoderma.2021.115402.

[12] S. M and P. Ts, "Geospatial modeling approaches for mapping topsoil organic carbon stock in northern part of Mongolia," *Proc. Mong. Acad. Sci.*, pp. 4–17, Oct. 2019. Available: doi: 10.5564/pmas.v59i2.1215.

[13] M. Emadi, R. Taghizadeh-Mehrjardi, A. Cherati, M. Danesh, A. Mosavi, and T. Scholten, "Predicting and Mapping of Soil Organic Carbon Using Machine Learning Algorithms in Northern Iran," *Remote Sens.*, vol. 12, no. 14, p. 2234, Jul. 2020. Available: doi: 10.3390/rs12142234.

[14] Q. Chen, Y. Wang, and X. Zhu, "Soil organic carbon estimation using remote sensing data-driven machine learning," *PeerJ*, vol. 12, p. e17836, Aug. 2024, doi: 10.7717/peerj.17836.

[15] T. Hengl et al., "SoilGrids250m: Global gridded soil information based on machine learning," *PLOS ONE*, vol. 12, no. 2, pp. e0169748, Feb. 2017. Available: doi: 10.1371/journal.pone.0169748.

[16] H. Keskin, S. Grunwald, and W. G. Harris, "Digital mapping of soil carbon fractions with machine learning," *Geoderma*, vol. 339, pp. 40–58, Apr. 2019. Available: doi: 10.1016/j.geoderma.2018.12.037.

[17] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Syst. Appl.*, vol. 134, pp. 93–101, Nov. 2019. Available: doi: 10.1016/j.eswa.2019.05.028.

[18] Z. Zhang et al., "Exploring the inter-decadal variability of soil organic carbon in China," *CATENA*, vol. 230, p. 107242, Sep. 2023. Available: doi: 10.1016/j.catena.2023.107242.

[19] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *Ann. Stat.*, vol. 29, no. 5, Oct. 2001. Available: doi: 10.1214/aos/1013203451.

[20] S. C. Wang, Y. Huo, and X. Mu, "Estimation of surface NO2 concentration in China based on extreme gradient boosted tree and deep learning methods," *Acta Sci Circumstantiae*, vol. 43, pp. 298–308, 2023.

[21] T. Chen, J. Jio, and Z. Zhang, "Soil quality evaluation of the alluvial fan in the Lhasa River Basin, Qinghai-Tibet Plateau," *Catena*, vol. 209, 2022.

[22] M. Kuhn, M. Campillos, P. González, L. J. Jensen, and P. Bork, "Large-scale prediction of drug–target relationships," *FEBS Lett.*, vol. 582, no. 8, pp. 1283–1290, Apr. 2008. Available: doi: 10.1016/j.febslet.2008.02.024.