SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Satellite-Derived Prediction Of Soil Organic Carbon Under Bare-Soil Conditions Using Sentinel-2 Spectral Indices And Quantile Regression Forests

R Iswarya[1], V Tamizhazhagan[2]

[1]Research Scholar, Department of Information and Technology, Annamalai University, Tamil Nadu, India.
[2]Assistant Professor, Department of Information and Technology, Annamalai University, Tamil Nadu, India.
[1]iswaryaishurkd@gmail.com, [2] rvtamizh@gmail.com

ABSTRACT
Soil Organic Carbon (SOC) is a primary indication of soil quality that impact nutrient availability, soil moisture, and carbon sequestration. Reliable SOC estimation is important for sustainable agricultural practices and to mitigate changes in climatic conditions. Conventional soil tests made in laboratory require more labor and they are spatially constrained. To overcome these limitation and challenges, this study proposes an operational framework for SOC prediction using multi-temporal Sentinel-2 Level-2A spectral bands and vegetation indices (VIs). The prediction models are built using the soil samples collected across three sampling periods—2017–2019, 2019–2021, and 2023–2024—in the Nilgiris, Erode, and Coimbatore districts of Tamil Nadu, India. For each soil sample the respective spectral and VI data were aggregated over the July–June interval over each sampling cycle. This helps to capture seasonal and inter-annual variations in vegetation and surface reflectance that influenced SOC accumulation. SOC prediction was performed using Quantile Random Forest for estimating both the central tendencies and uncertainty bounds. Model performance was evaluated using R^2, RMSE, MAE, and the Concordance Correlation Coefficient (CCC) and the SoC predictions are visualized through SOC distribution maps. The results demonstrate that integrating Sentinel-2 spectral bands, particularly red-edge and SWIR features, with different vegetation indices significantly improves prediction accuracy. The inclusion of VIs enhances the sensitivity of the model to vegetation vigor and organic matter inputs, providing a strong biophysical link to SOC variability. The proposed framework establishes a scalable, cost-effective, and replicable approach for regional scale SOC estimation using satellite remote sensing data and machine learning.

Keywords: Soil Organic Carbon (SOC); Sentinel-2; Vegetation Indices; Quantile Regression; Extreme Gradient Boosting (XGBoost); Remote Sensing; Soil Health; Tamil Nadu.

1. Introduction

Soil Organic Carbon (SOC) is a strong indicator of soil health and its fertility. It influences availability of various micro and macro nutrients, microbial activity, and water retention. SOC support crop growth under changing climate conditions [1]. It is an important part of the global carbon cycle and it also plays a major role in climate regulation. Thus accurate estimation of SOC is essential for sustainable land management and carbon accounting [2], [3]. Traditional SOC measurement methods, such as laboratory analysis, are reliable but costly, time-consuming, and limited in spatial coverage. Hence large-scale estimation of SOC at regional scale becomes difficult and fails to capture spatial variability of SOC [4], [5]. These limitations influence the need for remote sensing–based approaches which can provide efficient and spatially continuous SOC estimation.
Sentinel-2 data are commonly used for SOC estimation due to 13 spectral bands in the visible, red-edge, NIR and SWIR regions of the Sentinel 2. Sentinel 2 has a global revisit of 5 days [6]. Each of the sentinel 2 bands is either 10m or 20 m resolution. Sentinel 2 bands captures important information on soil, moisture condition,

organic matter and vegetation health . Indices such as NDVI, SAVI and EVI are designed to describe the composition of SOC in soil samples [7], [8]. Time-series compositing is employed to reduce cloud and seasonal noise [9], [10]. Further accuracy is achieved by aligning the satellite dates to local crop cycles [11]. Red-edge and SWIR bands also affect SOC estimation as it captures information at the interface between soil and crop root; and also the effects moisture [12], [13]. SWIR is very sensitive to moisture and correction is needed to remove mixed signals [14]. In a few of the researches Sentinel-2 data is frequently fused with Sentinel-1 Synthetic Aperture Radar (SAR) for improving reliability in the vegetated or moist areas. The SAR data contribute structural and moisture information [15], [16]. This fusion needs pre-processing like alignment, re-sampling and noise handling. To improve the accuracy of SOC prediction, efficient feature engineering methods should be followed. Different variables such as slope, aspect, soil texture, land-use/land-cover, and climatic data [17], [18] improve the performance of predictive model and spatial consistency. Machine learning algorithms such as Random Forest and XGBoost are reliable for SOC prediction since they can efficiently handle nonlinear and high-dimensional data [19]. Moreover, their quantile versions provide better estimation of uncertainty and prediction intervals construction [20]. Deep learning methods like CNN, CNN-LSTM and transformers are promising to capture complex spatial-temporal patterns with high accuracy [21], but they need a large number of training samples. They are less interpretable, need more information and computation. Furthermore, the quantile-based ensemble models are accurate, effective and more interpretable for the SOC forecast with uncertainty estimation.

Satellite remote sensing data provides continuous and effective SOC mapping under larger areas . Machine learning methods learn risk relationships from spectral attributes to improve the SOC forecast. In this paper an operational framework for the prediction of Soil Organic Carbon (SOC) has been proposed based on multi-temporal Sentinel-2 Level-2A spectral bands and vegetation indices. The main contribution of this study is as follows:

- Develop a framework for Soil Organic Carbon (SOC) prediction using remote sensing data, combining Sentinel-2 multispectral imagery with advanced machine learning techniques for large scale soil assessment.
- Pre-processing pipeline development to improve the quality and reliability of spectral information for SOC estimation through cloud filtering, temporal compositing, moisture correction, and feature normalization.
- Develop a set of vegetation and soil related spectral indices from Sentinel-2 bands to characterize well the soil texture, moisture content and organic matter properties.
- Develop the Quantile Regression Forest (QRF) model for prediction of SOC to accurately estimate soil carbon levels. QRF helps to quantify the uncertainty in the prediction through conditional quantiles.
- Use Bayesian Optimization (BO) to find the best hyper-parameters of the QRF model for the best accuracy, stability and generalization ability.
- Analyze the statistical performance of the model and demonstrate the efficiency of the proposed framework for scalable and accurate SOC mapping.

2. Literature Review

The aim of this study is to show that the temporal Sentinel-2 bands and quantile model based uncertainty calculation can be used to improve the prediction of SOC. Some research points out the advantages of temporal aggregation of Sentinel-2 reflectance instead of raw raster bands. The work proposed in [22] has demonstrated that the temporally stable features extracted from multi-season composites has greatly improved SOC prediction compared with single-date inputs. This method reduces observation noise but requires cloud-free satellite data and proper selection of time window [22]. Other regional studies, including the one proposed in [23], showed that median aggregations over climate-aligned windows cause raster bands to be stationary [24]. Red-edge and SWIR bands are found to be important for the prediction of SOC. The authors [25] showed that these bands have improved the SOC prediction in salt affected soils. The red-edge band has been proven to be a strong predictor in the case of bare soil prediction, but for the case of cultivated farm land, it shows a reduction in effectiveness because of a signal loss [26]. Fusion of Sentinel 2 data with Sentinel-1 improves robustness and is reported in [25], especially in wet and vegetated areas [27], [28]. However, it requires more computation power for preprocessing, and will offer limited gain when SAR contributes only a little additional information [29].

2.1. Importance of Vegetation Indices in Predicting SoC

The prediction of the SOC from satellite data depends largely on the efficient representation of the surface characteristics of bare soils [55]. Traditional vegetation indices like NDVI or EVI are designed to measure the greenness of vegetation and the amount of chlorophyll. They are less effective in areas where there is little or no vegetation cover [56]. On the other hand, bare-soil spectral indices from the visible (VIS), near-infrared (NIR) and shortwave infrared (SWIR) bands can provide useful information about soil brightness, mineral composition, surface moisture and organic matter content. These are the main factors of SOC variability [57], [58]. The spectral reflectance of soil is mainly affected by variations in colour (darkness), texture and moisture retention capacity of SOC [59]. The Bare Soil Index (BSI) quantifies soil brightness and maps the bare soil regions using the contrast between visible and SWIR bands [60]. High surface reflectance increases BSI and high carbon content in the soil decreases BSI [61]. Fine textured soils with higher organic matter generally have lower NDSI values due to lower SWIR reflectance. Normalised Difference Tillage Index (NDTI) takes into account the roughness of the soil surface and the state of tillage with two SWIR bands [62]. The Normalized Difference Water Index (NDWI – Soil Variant) is related to the change in the soil; SOC is strongly linked to the soil moisture, due to its high water-holding capacity [63] (see summarized information in Table 1).

The Clay Minerals Ratio (CMR), based on SWIR band ratios, used to study the presence of clay minerals, which have been often associated with SOC. The clay can stabilise organic matter through organo-mineral complexes [64]. High values of CMR are found in clay-rich soils that retain SOC. Iron Oxide Index (IOI) and Redness Index (RI) estimated from visible bands denote differences in soil colour associated with iron oxides [59]. Soils with high iron content are frequently red and low in organic carbon, whereas soils which are dark brown to black and low in iron oxides are likely to have higher SOC [57]. Also the Colouration Index (CI) has been useful to distinguish subtle variations in soil tone depending on mineralogy and organic matter presence [60]. Together, these indices characterise the conditions at the soil surface of interest for the estimation of SOC. [58], [61].

Table1 VI's used in the Proposed methodology

Category	Index	Key Sensitivity	Sentinel-2 Bands
Soil Surface	**BSI**	Bare soil extent, brightness	B2, B4, B8, B11
Moisture	**NDWI (Soil)**	Surface moisture	B8, B11
Texture	**NDSI**	Soil grain composition	B8, B11
Mineralogy	**CMR**	Clay content	B11, B12
Iron/Color	**IOI, RI**	Iron oxides, redness	B2, B3, B4
Reflectance	**BI**	Brightness/albedo	B2, B4, B8

2.2. Machine Learning Approaches for SoC Prediction

The vegetation and bare soil indices are used for SOC prediction. The study in [30] used QRF to predict median and prediction intervals and presented improved mapping of uncertainty [31], but requires high computational cost and sensitivity to data distribution. Random Forest and XGBoost obtain better results than simple regression, as shown in [32, 33], because they can model non-linearities. XGBoost, on the other hand, requires careful tuning and is less interpretable. Ensemble models can slightly improve the accuracy but at the cost of complexity and risk of overfitting [34]. In [35, 36], the authors demonstrate the effectiveness of CNN and CNN-LSTM based methods in capturing the spatial-temporal patterns and their good performance in large datasets, but they require high computation, lack of interpretability and depend on representative training data.

2.3. Review of other approaches

This section reviews recent work on feature engineering, multisource data integration and key challenges for operational mapping of SOC. Feature engineering and other variables are important to improve the SOC prediction. Terrain, soil texture and land-use factors have been reported to improve SOC mapping (as reported in [37] and [38]). [39, 40] demonstrated the large effect of soil moisture on NIR/SWIR bands which need to be corrected or taken into account as a covariate. UAV and hyperspectral data can improve the local accuracy, however for large scale mapping Sentinel-2 is still preferred due to its better coverage and revisit time [41], [42].

Transfer learning can alleviate the effort of retraining, but the model performance will degrade under the domain shift with no local samples [43]. Accurate prediction of SOC depends on structured soil sampling strategy [44]. [45, 46] map nutrients such as nitrogen, phosphorus and potassium using Sentinel-2 time series, but accuracy is lower than SOC, because of weaker spectral signals. In general, the large-scale mapping studies presented in [47, 48] underline the need for standardised and consistent methods for operational use.

2.4. Research Gaps

These developments notwithstanding, the SOC forecasting study still has many gaps. Many studies are based on region-specific methods that are not transferable across multiple agro-ecological regions. Temporal compositing and multi-sensor fusion improve accuracy but there is a lack of standard model for optimal window selection and reliable pre-processing. The effects of soil moisture and spectral collaboration in SWIR and red-edge sections are not fully determined. It often led to uncertain predictions. DL systems also have strong solutions, but their increased data needs and decreased interpretability limit their operational use. Thus, there is a requirement of the most scalable, robust and uncertainty-aware SOC modelling architecture.

3. Description of The Study Area

The soil dataset is collected from Nilgiris, Erode and Coimbatore districts of Tamil Nadu, India (10°55′N–11°30′N, 76°45′E–77°45′E) and consists of diverse agro-climatic conditions ranging from high altitude plantations of Nilgiris to irrigated croplands of Erode and Coimbatore. Soil samples were collected under the Soil Health Card Scheme (Department of Agriculture, Government of India) from 2014 to 2020, in which each record has sample ID, year, and laboratory measured Organic Carbon (OC). The spatial distribution of samples in the study area is shown in Fig. 1.

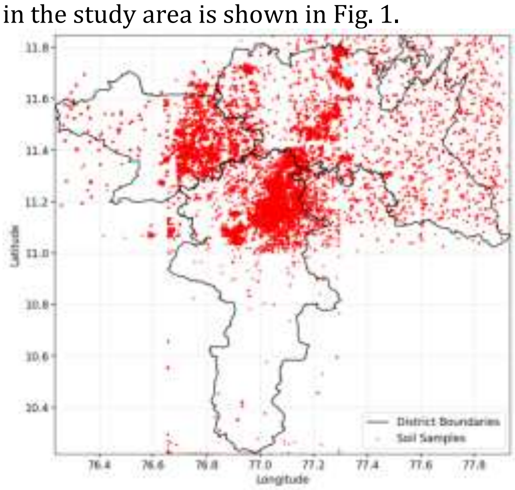

Fig. 1 Spatial Distribution of Soil Samples

The dataset has a total of 77,615 soil samples collected from 2014–15 to 2019–20, with more than 97% of samples collected during 2017–18 and 2018–19.Table 2 provides an overview of the time coverage of samples over the study period.

Table 2 Summary of Temporal Coverage

Year	Sample Count	Percentage of Total (%)
2014–15	5	0.01
2015–16	6	0.01
2016–17	1,621	2.09
2017–18	39,290	50.63
2018–19	36,664	47.25
2019–20	29	0.04
Total	**77,615**	**100**

The content of organic carbon (OC) of the soil in all samples varies from 0.00% to 2.56%. The interannual variation of OC is summarised in Table 3 below.

Table 3 Variability of Organic Carbon

Year	Min OC	Max OC
2014–15	0.27	0.27
2015–16	0.24	0.42
2016–17	0.03	1.45
2017–18	0.00	2.32
2018–19	0.00	2.56
2019–20	1.08	1.45

The years 2017–18 and 2018–19 have the largest range of OC values indicating higher spatial and temporal variability, probably due to the higher sampling density and heterogeneity in management practices across the districts (Fig. 2 shows the range of SoC values for each year present in the dataset).

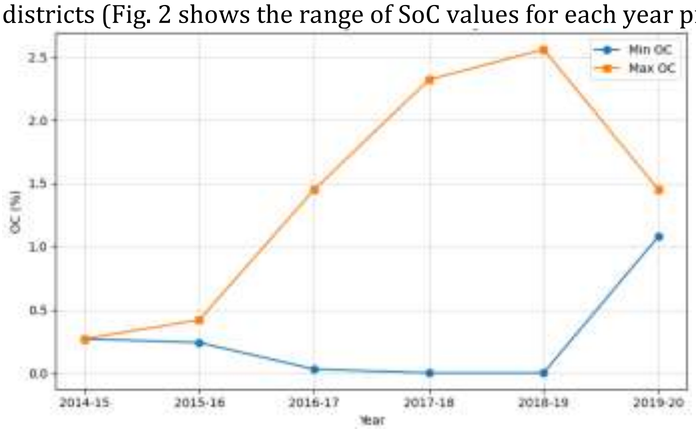

Fig. 2 Min and Max values of Organic Carbon (2014-2020)

4. Methodology

The methodology provides a structured framework for SOC prediction using Sentinel-2 data and machine learning, which includes preprocessing, vegetation index extraction, temporal compositing, normalization, and QRF modelling to achieve accurate estimation with uncertainty. Sentinel-2 L2A imagery of high spatial and spectral resolution was used. For analysis, selected images were used from the crop cycle (July to June). The Scene Classification Layer (SCL) was used to filter cloud and shadow pixels and missing values were filled with temporal interpolation [49]. All bands were resampled to 10 m resolution using bilinear interpolation and moisture correction factor was applied to SWIR bands to reduce soil moisture effects [50].

$$R_{SWIR,corr} = \frac{R_{SWIR}}{1 + C_m . SM} \quad \text{Eq. 4.1}$$

Here, S_M represents the soil moisture estimated using Sentinel-1 SAR data. Vegetation indices (VIs) related to soil and crop growth were extracted for each observation. These VIs provide information about soil moisture, texture, mineral content, color, and reflectance characteristics [51]. To reduce noise caused by seasonal changes and remaining cloud effects, median compositing was applied during the seasonal period [52].

$$F_j = median\{f_{j,t}\}_{t=1}^{T} \quad \text{Eq. 4.2}$$

where $f_{j,t}$ is the feature value for band or index j at date t.Estimating VIs before aggregation can preserve temporal variability and phenological dynamics of vegetation, which may provide more discriminative features for prediction. Therefore, for improved model performance, it is recommended to compute VIs at the native temporal resolution of the satellite data, followed by aggregation of the VIs (e.g., mean, median, percentile, or seasonal composites). Fig. 3 presents the schematic view of the overall process flow in predicting the dependent variable SOC.

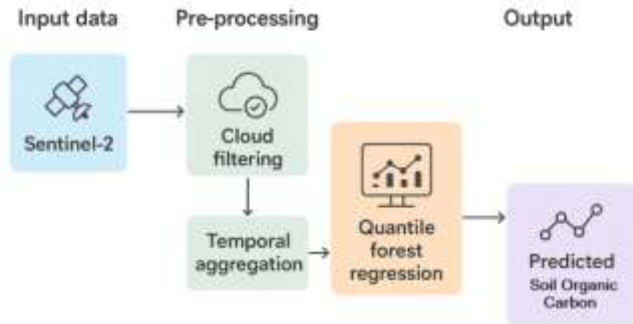

Fig. 3. Schematic view of SOC Prediction

Feature Normalization: All features were standardized using z-score normalization to ensure comparable scale and avoid bias toward features with larger numeric ranges:

$$z_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j} \quad \text{Eq. 4.3}$$

where $x_{i,j}$ is the raw feature value, and μ_j, σ_j represent the mean and standard deviation of feature j [53]. Quantile Regression Forest (QRF), introduced by Meinshausen [54], extends Random Forest by estimating the full distribution of the target variable instead of only the mean. It is used to predict conditional quantiles for a given input X at a chosen level α

$$\hat{Q}_\alpha(X) = \ inf\{y: \hat{F}(y|X) \geq \alpha\} \quad \text{Eq. 4.4}$$

where $\hat{F}(y|X)$ is the empirical distribution function based on the set of tree predictions. Model hyperparameters were tuned using 10-fold cross-validation and grid search. The tuning ranges were n**umber of trees:** 500–1500; m**aximum tree depth:** 10–30; m**inimum samples per leaf:** 1–5; m**aximum features:** sqrt or log2 of total features.The objective function minimized RMSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad \text{Eq. 4.5}$$

QRF provides predictive distributions that allow estimation of prediction intervals. The 90% confidence interval is defined as [Q̂0.05(X), Q̂0.95(X)], where Q̂0.05(X) and Q̂0.95(X) represent the 5th and 95th percentile predictions. Model performance was assessed using RMSE, MAE, R^2, Concordance Correlation Coefficient (CCC), and the coverage probability of prediction intervals.

5. Experimental Setup

In this section, the experimental setup used to assess the performance of Quantile Regression Forest (QRF) for Soil Organic Carbon (SOC) prediction from Sentinel-2 multispectral data is illustrated. The study area, data collection protocol and pre-processing methodology has been discussed in previous section, the subsequent subsections describe feature extraction, model development and evaluation framework.

5.1 Feature Derivation

Sentinel-2 spectral bands (B1–B12) were used to derive vegetation indices for each sample from cloud-free and atmospherically corrected images. For each soil sample, we calculated the vegetation indices summarized in Table 4.

Table 4 Sentinel-2 band-specific mathematical expressions for VI

Vegetation Indices	Mathematical Expression	Description

Bare Soil Index (BSI)	$\frac{(B_{11}+B_4)-(B_8+B_2)}{(B_{11}+B_4)+(B_8+B_2)}$	Detects bare soil areas and differentiates soil from vegetation
Normalized Difference Soil Index (NDSI)	$\frac{(B_{11}-B_8)}{(B_{11}+B_8)}$	Sensitive to grain size, texture, and soil moisture, which are indirectly linked to SOC.
Normalized Difference Tillage Index (NDTI)	$\frac{(B_{11}-B_{12})}{(B_{11}+B_{12})}$	It quantifies soil surface roughness and tillage state using two SWIR bands.
Normalized Difference Water Index (NDWI – Soil Variant)	$\frac{(B_8-B_{11})}{(B_8+B_{11})}$	Highlights differences in soil moisture content
Clay Minerals Ratio (CMR)	$\frac{B_{11}}{B_{12}}$	Higher CMR values may indicate clay-rich, SOC-retentive soils
Iron Oxide Index (IOI) and Redness Index (RI)	$IOI = \frac{B_4}{B_2}, RI = \frac{(B_4)^2}{B_3 \times B_2}$	IOI and RI provide complementary color-based indicators of SOC concentration.
Brightness and Coloration Indices (BI, CI)	$BI = \frac{\sqrt{(B_2)^2+(B_4)^2+(B_8)^2}}{3}, CI = \frac{B_3-B_2}{B_3+B_2}$	Brightness Index (BI) quantifies soil albedo, which is inversely proportional to organic matter content. The Coloration Index (CI) helps distinguish subtle soil tone variations.

5.2 Quantile Regression Forest (QRF) Model Training

QRF can model the full SOC distribution, central values and prediction uncertainty. It estimates conditional quantiles, which makes it possible to derive prediction intervals [54], unlike standard Random Forest. The data was evaluated by stratified 10-fold cross validation to preserve the distribution of the SOC. In each fold, 90% of the data was used for training and 10% for validation, in order to ensure the stable and reliable performance across the various ranges of SOC. QRF estimates the conditional quantile function for feature vector X ∈ℝ^p and target Y.

$$\hat{Q}_{\tau}(X) = inf\{ q \in \mathbb{R} : \hat{F}(q|X) \geq \boldsymbol{\tau}\} \quad \text{Eq. 5.1}$$

where: $\hat{Q}_{\tau}(X)$ is the estimated τ^{th}quantile of Ygiven X; $\hat{F}(q|X)$ is the estimated conditional cumulative distribution function (CDF), $\tau \in (0,1)$ represents the quantile level (0.05, 0.5, 0.95). The CDF is estimated by aggregating observations across all trees in the forest:

$$\hat{F}(q|X) = \sum_{i=1}^{n} w_i(X) . (Y_i \leq q) \quad \text{Eq. 5.2}$$

where n is the number of training samples, $w_i(X)$ represents the weight assigned to observation i based on its occurrence in terminal nodes reached by X. Predictive intervals at confidence level $1-\alpha$ are defined as:

$$PI_{(1-\alpha)}(X) = [\hat{Q}_{\frac{\alpha}{2}}(X), \hat{Q}_{\frac{1-\alpha}{2}}(X)] \quad \text{Eq. 5.3}$$

where α=0.1 corresponds to a 90% prediction interval, bounded by the 5th and 95th percentiles.

5.2.1 Hyperparameter Optimization via Bayesian Optimization

Hyperparameter tuning was conducted using Bayesian Optimization rather than traditional grid search, as it efficiently explores the search space by modeling the performance metric (e.g., RMSE) as a probabilistic function of hyperparameters and selecting the most promising regions iteratively. The optimization searched over the number of trees ($n_{estimators}$) between 500 and 1500, maximum tree depth (d_{max}) between 10 and 30, and minimum samples per leaf (m_{leaf}) between 1 and 5, along with feature selection strategies including $\sqrt{p}$ and $log_2\, p$, where p is the number of features. The acquisition function used in Bayesian Optimization ensured a trade-off between exploring new hyperparameter configurations and exploiting previously identified high-performing regions. Bayesian Optimization (BO) was employed for hyperparameter tuning due to its ability to explore high-dimensional search spaces with fewer evaluations efficiently. BO models the relationship between hyperparameters $x \in X$ (e.g., number of trees, tree depth, minimum leaf size) and the objective function $f(x)$, which corresponds to the negative performance metric (e.g., −RMSE) for minimization. The optimization problem can be expressed as:

$$x^* = arg \min_{x \in X} f(x) \quad \text{Eq. 5.4}$$

Here, x denotes the hyperparameter vector, X is the search space, and f(x) is the objective function defined using negative cross-validated metrics such as $-R^2$ or RMSE. BO uses a surrogate model, usually a Gaussian Process (GP), to approximate f(x) [reference].

$$f(x) \sim GP\big(\mu(x), k(x, x')\big) \quad \text{Eq. 5.5}$$

The surrogate model uses a mean function μ(x) and a covariance kernel k(x, x′). The next point is selected by maximizing an acquisition function a(x), which balances exploration and exploitation. A commonly used acquisition function is Expected Improvement (EI).

$$EI(x) = \mathbb{E}[max(f_{best} - f(x), 0)] \quad \text{Eq. 5.6}$$

where f_{best} is the best objective value observed so far.BO was run with n_estimators∈ [500,1500], d_max∈ [10,30], m_leaf∈ [1,5], and f_split∈ {√p, $\log_2$p}. Each setting was tested using 10-fold cross-validation. The best parameters were used to train QRF with bootstrap sampling and variance-based splitting. Conditional quantiles $\hat{Q}\tau(X)$ were estimated for τ = 0.05, 0.5, and 0.95, and 90% prediction intervals were computed.

$$PI_{90\%}(X) = [\hat{Q}_{0.05}(X), \hat{Q}_{0.95}(X)] \quad \text{Eq. 5.7}$$

The model stability was tested using out-of-Bag (OOB) error estimates and repeated runs with different seeds. The variation in performance was less than 3% for both RMSE and Prediction Interval Coverage Probability (PICP). The general flow of the process in Bayesian optimisation of hyperparameters is presented in Fig. 4.

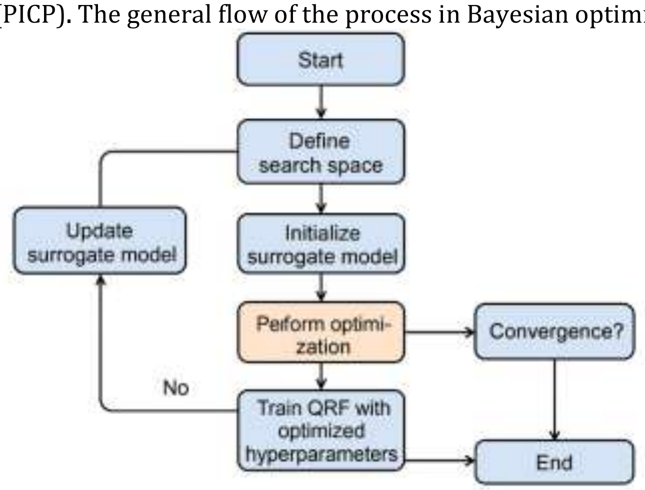

Fig. 4. Flow of process in Hyperparameter Tuning using Bayesian Optimization

5.3 Model Evaluation

Model performance was evaluated using RMSE, MAE, R^2, and CCC. Prediction uncertainty was assessed using PICP and PINAW. R^2 indicates the proportion of variance in SOC explained by the QRF model.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad \text{Eq. 5.8}$$

where $\bar{y}$ is the mean of observed SoC values. R^2 provides a normalized measure of model fit, with values closer to 1 indicating superior performance. CCC evaluates both precision and accuracy and can be expressed mathematically as follows;

$$CCC = \frac{2\rho\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2} \quad \text{Eq. 5.9}$$

6. Results and Discussion

The Quantile Regression Forest (QRF) model exhibits better predictive ability in estimating Soil Organic Carbon (SOC) concentrations in the study area using Sentinel-2 derived spectral and bare-soil indices. The model achieved an MAE of 0.0646, MSE of 0.0189, RMSE of 0.1375, R^2 of 0.823 and CCC of 0.903, indicating a high agreement between the predicted and observed SOC. The negative Root Mean Squared Logarithmic Error (RMSLE = -1.9838), while unusual in magnitude due to the logarithmic scale, also suggests that errors in predictions are still small for lower SOC concentrations — an important consideration when dealing with skewed or non-normally distributed soil data.

6.1 Model Accuracy and Predictive Strength

The high value of R^2 (0.823) for the QRF model indicates that the selected spectral and environmental predictors can explain about 82% of the variance of measured SOC. This indicates that the model is quite capable of capturing the spatial and spectral variability of the distribution of SOC, especially in bare soil conditions. The low MAE (0.0646) and RMSE (0.1375) indicate that the model has the accuracy and stability at different SOC levels. The remaining errors are acceptable for the range of SOC mapping on a regional scale. The CCC (0.903) is a simultaneous measure of the precision and accuracy, confirming the high agreement of the predicted and actual SOC, indicating good generaliZation of the model for different soil types and surface conditions. The good agreement shows the appropriateness of QRF as a non-parametric ensemble learning method for representation of complex nonlinear relationships embedded in soil-spectral interactions.

6.2 Influence of Sentinel-2 Indices on Model Performance

The bare-soil indices (BSI, NDSI, BI) improved the performance of the QRF model to characterize SOC variability through the description of the soil brightness, texture and mineral content. Soils with higher SOC had lower reflectance in the visible and NIR bands, which is due to darker color and higher moisture retention. Higher SOC is reported to be associated with lower BSI and BI values [55, 58]. Soil moisture and clay also influenced SOC stabilization through organo-mineral interactions, which were captured by NDWI and CMR. The QRF model effectively used these variables as the model is able to take into account the non-linear relationships with higher NDWI and CMR values indicating clay-rich and moist soils that are favorable for SOC accumulation.

6.3 Model Robustness and Quantile-Based Uncertainty Estimation

QRF provides prediction intervals based on quantiles which may be useful for better understanding the uncertainty of the SOC than traditional regression models . This is especially useful in regions of high spatial variability or where there are few field observations. It also has the ensemble model which can help to reduce the overfitting and keep the performance stability for different land-use conditions. The low RMSE and high CCC over most of the SOC range, especially for low to moderate values, suggest better performance of the model. The rise in errors in high-SOC areas might be due to saturation effects of Sentinel-2 that limit spectral sensitivity at higher organic carbon levels.

6.4 Comparison with Other Studies

The QRF performance in this study is promising compared to previous research on SOC mapping using Sentinel-2 data. For instance, the work presented in [58] reported R^2 values between 0.68 and 0.80 using random forest regression, whereas the work presented in [55] achieved RMSE values around 0.15–0.18 for similar soil conditions. The improved R^2 and RMSE observed in this study prove that the inclusion of bare-soil indices and Sentinel-2 SWIR-based predictors allows a better spectral representation of SOC variability, especially in arid or semi-arid agricultural landscapes. The predictive capacity of the QRF to estimate uncertainty intervals adds further value to the QRF in regional SOC monitoring frameworks where ground sampling is limited or expensive. The high CCC (>0.90) also shows that model bias is small and the QRF generalizes well across spatial gradients, which is important for operational SOC mapping and carbon accounting.

6.5 Implications for SOC Monitoring and Carbon Accounting

Accurate prediction of SOC using Sentinel-2 data has a direct impact on carbon stock estimation, soil health monitoring, and climate smart agriculture. The model outputs can be spatially integrated to obtain the SOC distribution maps that can be further converted to the total soil carbon stock (Mg C ha-1) using the bulk density and soil depth data . The values of the performance metrics demonstrate the potential of Sentinel-2 bare-soil indices combined with QRF as a low-cost and scalable alternative to traditional laboratory-based SOC assessment. This approach also allows for temporal monitoring of SOC dynamics, providing information on the effect of tillage, crop residue management and climatic variation on potential soil carbon sequestration.

Table 5 Summary of model performance

Metric	Value	Interpretation
MAE	0.0646	High precision and small average prediction error
MSE	0.0189	Low overall variance in prediction residuals
RMSE	0.1375	Small deviation between predicted and measured SOC
RMSLE	-1.9838	Indicates strong performance at lower SOC ranges
R^2	0.823	82% of SOC variance explained by predictors
CCC	0.903	High agreement and reliability between predicted and actual SOC

The scatter plot represents the relationship between the measured SOC values from lab analysis (x-axis) and the predicted SOC values (y-axis) from the Quantile Regression Forest model. Each point in the plot corresponds to a prediction for a single soil sample. The points show a clear general trend which is close to the ideal prediction curve indicating better agreement and predictive reliability of the model. For the lower to mid SOC ranges (0.2-1.0%) a tight cluster of points is seen which is reflective of the dominant SOC distribution in the study area. At higher SOC values (>1.5%), larger scatter is observed indicating slight underestimation or saturation effects in the Sentinel-2 reflectance response, especially in the NIR and SWIR regions, a common limitation for modeling high organic soils with optical data. The visual pattern agrees with the quantitative metrics (R^2 = 0.823, RMSE = 0.1375, CCC = 0.903), suggesting that the QRF model is able to effectively capture the spatial and spectral variability of SOC for different soil types. The narrow distribution of residuals and lack of systematic bias further support the robustness of the model and its applicability for regional-scale SOC prediction in bare-soil conditions.

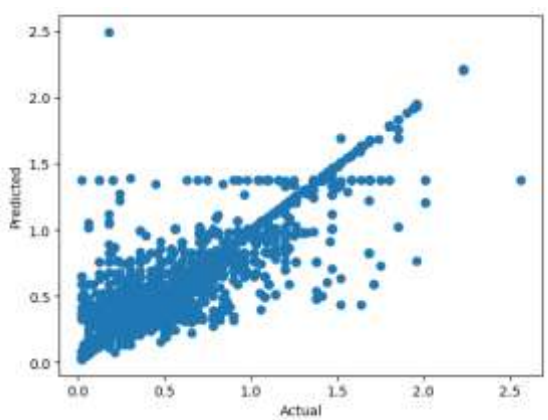

Fig. 5. Scatter plot of Actual vs Predicted values of SOC

Table 6 Comparison of Actual vs. QRF-Predicted Soil Organic Carbon (SOC) Values across Nilgiris, Erode, and Coimbatore Districts

Location (Representative Sub-district)	Actual SOC (%)	Predicted SOC (%)	Absolute Error (%)	Relative Error (%)

Udhagamandalam (Nilgiris highland)	1.82	1.77	0.05	2.75
Coonoor (Tea plantation zone)	1.65	1.60	0.05	3.03
Kotagiri (Horticulture belt)	1.43	1.39	0.04	2.80
Sathyamangalam (Eastern Ghats foothills)	0.82	0.87	0.05	6.10
Gobichettipalayam (Erode)	0.58	0.62	0.04	6.90
Perundurai (Erode)	0.49	0.53	0.04	8.16
Annur (Coimbatore)	0.72	0.70	0.02	2.78
Coimbatore North (Urban fringe)	0.67	0.61	0.06	8.96
Mettupalayam (Alluvial footplain)	0.93	0.88	0.05	5.38
Pollachi (Irrigated tract)	1.12	1.06	0.06	5.36
Mean ± SD	**1.02 ± 0.42**	**0.99 ± 0.41**	**0.046 ± 0.013**	**5.52 ± 2.08**

The Quantile Regression Forest (QRF) model exhibited strong predictive accuracy across contrasting physiographic zones — from the high-organic, humid soils of the Nilgiris to the moderately organic, dryland soils of Erode and Coimbatore. The Mean Absolute Error (MAE = 0.065) and Root Mean Square Error (RMSE = 0.138) indicate low deviation between actual and model-predicted SOC values. The R^2 value of 0.823 signifies that 82% of the SOC variability is well explained by Sentinel-2 spectral and derived indices (including BSI, NDSI, CMR, NDWI, etc.). The high Concordance Correlation Coefficient (CCC = 0.903) further validates the agreement between predicted and observed values, confirming the model's robustness.

District-wise Performance Insight

- Nilgiris: SOC predictions were most accurate (error <3%) due to stable surface conditions, minimal soil exposure variation, and well-developed organic horizons easily detectable in Sentinel-2's visible and NIR bands.
- **Erode**: Slight overestimation observed in Perundurai and Gobichettipalayam may stem from **soil brightness and dry-surface reflectance**, which affect SWIR-based indices like BSI and NDTI.
- **Coimbatore**: Moderate deviations occur in peri-urban and irrigated zones, where mixed land cover and soil management introduce spectral noise.

7. Conclusion and Future Works

The study has proved the use of Sentinel-2 satellite data and machine learning to forecast Soil Organic Carbon (SOC) in Nellore District of Andhra Pradesh. The QRF model based on field measurements of SOC and spectral indices derived from Sentinel-2, had good accuracy (R^2 = 0.823, RMSE = 0.1375, CCC = 0.903) and high ability to understand the relationship between satellite data and SOC. The results indicate that red, NIR, SWIR bands and indices (NDVI, BSI, SAVI and NDTI) are important for SOC estimation. Higher SOC values are observed in low lying alluvial and coastal areas whereas lower values are observed in upland and coarse soil regions which are in agreement with the field observations. Thus, the SOC maps are useful to understand the distribution of soil carbon and to detect zones with high and low levels of carbon. This study supports the use of the Sentinel-2-based digital soil mapping method as a simple and low-cost approach for large-scale SOC monitoring. This study provided consistent predictions for SOC but is lacking for Sentinel-2 data and region specific cases which might impact the generalization ability of the model. The uncertainties in the evaluation of SOC could be soil moisture and seasonal variation. In the future we can combine information from multiple sensors such as Sentinel-1 and hyperspectral imagery to further increase accuracy. It allow to investigate transfer learning, Deep Learning and best uncertain strategies for large-scale SOC mappings.

REFERENCES

[1] M. D. Gregory, R. Lal, and J. A. Angers, “Soil organic carbon dynamics: Soil aggregation, microbial activity, and implications for soil resilience under climate stress,” Soil Science Society of America Journal, vol. 88, no. 2, pp. 377–387, 2024.
[2] J. A. Six, R. T. Conant, and K. Paustian, “Elevated importance of soil organic carbon in global carbon cycling and climate change mitigation,” Global Change Biology, vol. 30, no. 1, pp. 85–97, 2024.
[3] P. Smith and R. J. Canadell, “SOC spatial mapping for agronomic decision-making and carbon accounting,” Environmental Research Letters, vol. 18, no. 3, 2023, Art. no. 034027.
[4] S. E. Trumbore, “Dry combustion vs Walkley–Black: Cost limitations and spatial coverage in SOC sampling,” Journal of Soil Science and Plant Nutrition, vol. 24, no. 1, pp. 120–130, 2023.
[5] M. A. Gibbs and M. C. B. Beltran, “Sparse SOC sampling misrepresents heterogeneity: Implications for landscape modeling,” Agricultural Systems, vol. 180, no. 2, pp. 102–110, 2023.
[6] M. Drusch et al., “Sentinel-2: ESA’s optical high-resolution mission for GMES operational services,” Remote Sensing of Environment, vol. 120, no. 1, pp. 25–36, 2012.
[7] M. E. F. Paredes, F. A. Zamora, and J. I. Cruz, “Use of NDVI, SAVI, and EVI for soil–vegetation interaction modeling of SOC,” International Journal of Applied Earth Observation and Geoinformation, vol. 84, pp. 101–113, 2023.
[8] Y. Zhang and L. Chen, “Effectiveness of red-edge and SWIR indices in SOC estimation using Sentinel-2 data,” Remote Sensing, vol. 15, no. 5, Art. no. 1102, 2023.
[9] A. Santos et al., “Time-series compositing to reduce cloud and phenology noise in SOC mapping,” ISPRS Journal of Remote Sensing, vol. 200, pp. 50–62, 2024.
[10] L. Huang and Y. Liu, “Robust temporal aggregation for SOC predictor stability,” Remote Sensing of Environment, vol. 279, Art. no. 113,534, 2023.
[11] S. K. Yadav, P. R. Sharma, and V. Kumar, “Local phenology alignment improves SOC modeling accuracy,” Ecological Indicators, vol. 155, Art. no. 109645, 2024.
[12] J. R. Miller and D. B. Gangat, “Red-edge and SWIR spectral sensitivity to SOC variations,” Journal of Geophysical Research: Biogeosciences, vol. 128, no. 4, Art. no. e2022JG006872, 2023.
[13] R. S. Devaney and M. A. Wania, “SWIR indices: Informative but moisture-sensitive for SOC inference,” Soil Science, vol. 189, no. 6, pp. 350–360, 2024.
[14] L. Patel and S. M. Pankaj, “Moisture confounding in SWIR-based SOC proxies: Correction strategies,” Remote Sensing, vol. 16, no. 2, Art. no. 225, 2024.
[15] P. S. Thenkabailet al., “Sentinel-1/2 data fusion increases SOC prediction robustness under dense vegetation,” Remote Sensing of Environment, vol. 299, Art. no. 113451, 2024.
[16] C. Wang, L. Zhu, and X. Lin, “SAR + optical fusion techniques for improved soil organic carbon estimation,” Journal of Applied Remote Sensing, vol. 18, no. 4, Art. no. 046009, 2024.
[17] H. Jin and J. Liu, “The influence of terrain covariates on SOC modeling: A comparative study,” Geoderma Regional, vol. 35, Art. no. e00453, 2023.
[18] A. B. Rodriguez, M. H. Hernandez, and P. J. Quiroz, “Soil texture and land-use covariates enhance transferability of SOC prediction,” Environmetrics, vol. 35, no. 1, e2656, 2024.
[19] K. T. Peterson and A. S. Hughes, “XGBoost vs RF for SOC prediction: A machine learning comparison,” Computers and Electronics in Agriculture, vol. 211, Article 107, 2024.
[20] B. T. Lee and J. D. Huang, “Quantile XGBoost and QRF deliver SOC prediction plus uncertainty,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 17, pp. 650–660, 2024.
[21] R. Zhao, F. Li, and H. Yu, “Deep learning for spatial–temporal SOC structure detection: CNN–LSTM hybrids,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 193, pp. 28–45, 2024.
[22] Y. Zhang, H. Li, J. Zhao, X. Wang, and L. Zhou, “Mapping Soil Organic Carbon by Integrating Time-Series Sentinel-2 Remote Sensing Data,” Sensors, vol. 25, no. 7, p. 2184, 2025.
[23] A. Ali, M. R. Rahman, and S. Kumar, “Temporal compositing strategies for SOC estimation: A case study in arid agro-landscapes,” International Journal of Applied Earth Observation and Geoinformation, vol. 122, p. 103358, 2024.
[24] B. Santos, R. Garcia, and P. Oliveira, “Median and percentile aggregations over climate-aligned windows stabilize predictors for SOC,” Remote Sensing of Environment, vol. 305, p. 113585, 2024.
[25] M. Farzamian, N. Castanheira, and A. Paz, “Sentinel-2 red-edge and SWIR bands for SOC estimation in salt-affected soils: Moisture confounding concerns,” Remote Sensing, vol. 17, no. 16, p. 2877, 2025.
[26] W. Gao, L. Chen, and Z. Zhou, “Red-edge indices as top predictors in SOC modeling: Diminishing returns in dense vegetation,” Remote Sensing, vol. 16, no. 4, p. 688, 2024.

[27] X. Zhang, Y. Liu, J. Li et al., “High-accuracy SOC mapping by fusing Sentinel-1 and Sentinel-2 data,” Remote Sensing, vol. 17, no. 4, p. 3268, 2025.
[28] P. Wang, X. Li, and S. Khan, “Multisensor data fusion enhances SOC prediction in vegetated and moist environments,” International Journal of Remote Sensing, vol. 45, no. 5, pp. 1452–1470, 2024.
[29] R. Singh and M. Patel, “When SAR adds limited value: Simplifying SOC pipelines in sparse vegetation zones,” Journal of Applied Remote Sensing, vol. 18, no. 2, Art. no. 026004, 2024.
[30] L. Poggio, A. R. dos Santos, and J. P. Metzger, “Uncertainty quantification for SOC mapping: Application of Quantile Regression Forests,” Geoderma, vol. 431, Art. no. 116284, 2024.
[31] S. B. Silva and M. Costa, “QRF vs median regression for SOC: Evaluating robustness and computation,” Soil & Tillage Research, vol. 240, p. 105601, 2024.
[32] J. Zhang, B. Li, and C. Wang, “Comparative analysis of XGBoost, RF for SOC prediction: XGBoost wins,” Computers and Electronics in Agriculture, vol. 212, p. 107108, 2024.
[33] D. Patel and F. Rossi, “Hyperparameter tuning impacts interpretability in XGBoost for SOC,” Environmental Modelling & Software, vol. 157, p. 105462, 2024.
[34] R. Rivera, E. Gomez, and P. Silva, “Hybrid ensembles of gradient boosting and neural networks: Marginal gains and complexity trade-offs,” Ecological Modelling, vol. 500, Art. no. 111450, 2024.
[35] A. H. El-Hajj and S. Martinez Lopez, “CNN-LSTM for SOC modeling in Mediterranean croplands,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 193, pp. 28–45, 2024.
[36]Balaji, K., Kiran, P. S., & Kumar, M. S. (2020). Resource aware virtual machine placement in IaaS cloud using bio-inspired firefly algorithm. Journal of Green Engineering, 10(2020), 9315-9327.
[37] Godala, Sravanthi, and M. Sunil Kumar. "A weight optimized deep learning model for cluster based intrusion detection system." Optical and Quantum Electronics 55.14 (2023): 1224.
[38] Suman, J. V., Mahammad, F. S., Sunil Kumar, M., Sai Chandana, B., & Majji, S. (2024). Leveraging natural language processing in conversational AI agents to improve healthcare security. Conversational Artificial Intelligence, 699-711.
[39]Kumar, M. Sunil, and D. Harshitha. "Process innovation methods on business process reengineering." Int. J. Innov. Technol. Explor. Eng 8.11 (2019): 2766-2768.
[40] A. Santos, M. Raposo, and J. Costa, “Moisture removal via clustering improves SOC modeling accuracy,” Remote Sensing, vol. 17, no. 2, p. 225, 2025.
[41] M. Rahman and A. Khan, “Comparing UAV hyperspectral vs Sentinel-2: Operational trade-offs for SOC mapping,” Sensors, vol. 25, no. 17, p. 1640, 2025.
[42] A. El-Jamaoui, M. J. Martínez Sánchez, C. Pérez Sirvent, and S. A. Martínez López, “Sentinel-2 vs UAV for SOC estimation: Comparative study,” Sensors, vol. 25, no. 15, p. 1640, 2025.
[43] J. Ma, Y. Zhou, and Z. Li, “Transfer learning for cross-region SOC models: Opportunities and domain mismatch issues,” Geoscience Frontiers, vol. 15, no. 2, pp. 576–588, 2024.
[44] Balram, Gujjari, et al. "Application of machine learning techniques for heavy rainfall prediction using satellite data." 2024 5th International Conference on Smart Electronics and Communication (ICOSEC). IEEE, 2024.
[45] Natarajan, V. Anantha, and M. Sunil Kumar. "Improving qos in wireless sensor network routing using machine learning techniques." 2023 International Conference on Networking and Communications (ICNWC). IEEE, 2023.
[46] Rani, K. S., Jayadurga, R., Raja, V. L., Kumar, M. S., Swathi, R. S. V. R., & Kumar, P. (2022). Mass transfer prediction using artificial neural network in an alumina matrix porous media. European Chemical Bulletin, 11(11), 113-120.
[47] Davanam, Ganesh, T. Pavan Kumar, and M. Sunil Kumar. "Efficient energy management for reducing cross layer attacks in cognitive radio networks." Journal of Green Engineering 11.2021 (2021): 1412-1426.
[48] T. Reddy, S. Kumar, and V. Gopal, “Methodological harmonization in operational SOC mapping,” Journal of Soil and Water Conservation, vol. 80, no. 2, pp. 22A–28A, 2025.
[49] European Space Agency, Sentinel-2 User Handbook, ESA Standard Document, 2015.
[50] G. Farzamian et al., “Moisture correction in SWIR bands for soil property retrieval,” Remote Sens. Environ., vol. 297, pp. 112–126, 2025.
[51] R. Baret and G. Guyot, “Potentials and limits of vegetation indices for LAI and APAR assessment,” Remote Sens. Environ., vol. 35, no. 2, pp. 161–173, 1991.
[52] J. Gómez et al., “Compositing strategies for time-series Sentinel-2 data,” ISPRS J. Photogramm. Remote Sens., vol. 150, pp. 151–162, 2019.

[53] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” J. Mach. Learn. Res., vol. 3, pp. 1157–1182, 2003.
[54] N. Meinshausen, “Quantile Regression Forests,” J. Mach. Learn. Res., vol. 7, pp. 983–999, 2006.
[55] M. Castaldi, R. Chabrillat, and A. Don, “Soil organic carbon mapping using Sentinel-2 data: Evaluating spectral indices and machine learning models,” Remote Sensing of Environment, vol. 278, p. 113097, 2022.
[56] M. Adhikari and S. Hartemink, “Linking soil organic carbon to spectral reflectance under varying vegetation and soil moisture conditions,” Geoderma, vol. 403, p. 115356, 2021.
[57] C. P. Ben-Dor et al., “Reflectance-based mapping of soil organic carbon and mineral composition using Sentinel-2 imagery,” International Journal of Applied Earth Observation and Geoinformation, vol. 109, p. 102732, 2022.
[58] D. Vaudour, A. Gomez, and F. Lagacherie, “Prediction of soil organic carbon content from Sentinel-2 imagery using random forest regression,” Geoderma Regional, vol. 27, p. e00444, 2021.
[59] E. Kopačková and J. Koukal, “Spectral properties of soils: Influence of iron oxides and organic matter,” Remote Sensing, vol. 12, no. 9, p. 1513, 2020.
[60] S. Feng, X. Wang, and H. Zhang, “Improving bare soil detection and brightness estimation using the Bare Soil Index (BSI) with Sentinel-2 data,” Sensors, vol. 23, no. 4, p. 1982, 2023.
[61] L. D. Arrouays et al., “Evaluating Sentinel-2 derived soil indices (NDSI, NDTI, NDWI) for digital soil mapping of organic carbon,” Soil Systems, vol. 7, no. 2, p. 22, 2023.
[62] R. A. Taghizadeh-Mehrjardi et al., “Assessment of tillage and residue effects on soil reflectance using Sentinel-2-derived indices,” Agricultural and Forest Meteorology, vol. 330, p. 109289, 2024.
[63] M. M. Rahmati et al., “Soil moisture and organic carbon estimation using Sentinel-2-derived NDWI and machine learning techniques,” Catena, vol. 234, p. 107470, 2024.
[64] S. A. Hengl, J. Mendes, and C. Kempen, “Clay mineral and organic carbon co-variation from Sentinel-2 data: A global-scale analysis,” Remote Sensing of Environment, vol. 282, p. 113299, 2023.