Digital Soil Mapping from Multiple Satellite Sensor Using Machine Learning

Course: MSc Agriculture Analytics

Guided by: Mr. Justin George K

Submitted by: Krupali Bhayani (202419003)

Krishna Kakad(202419009)

Indian Institute of Remote Sensing

Indian Space Research Organization

Department of Space, Govt. of India Dehradun – 248001

Uttarakhand, India

April 2025

Table of contents

1. Introduction	4
2. Review of Literature	4
3. Research Question	5
4. Objective	5
5. Study Area	6
6. Data Description	7
7. Methodology	10
8. Result and Analysis	15
9. Conclusion	21
10. References	21

Title: Digital Soil Mapping from Multiple Satellite Sensor Using Machine Learning.

Abstract:

This project showcases the application of Digital Soil Mapping (DSM) methodologies to predict and spatially model key soil attributes—pH, Electrical Conductivity (EC), Total Organic Carbon (TOC) in the mountainous terrain of the Tehri Garhwal region, Uttarakhand, India. Leveraging multi-year Sentinel satellite imagery (2017–2024) alongside ground-truth soil sample data, the study employs machine learning techniques—specifically Random Forest regression—to generate high-resolution, spatially continuous maps of soil properties. A comprehensive set of 19 environmental covariates—including topographic indices (elevation, slope, curvature), vegetation metrics (NDVI, brightness), and soil-chemical indicators—were derived, standardized through z-score normalization, and spatially aligned using NDVI as the base reference layer to ensure inter-raster consistency. The feature selection process identified the most influential covariates using permutation-based importance scores, enhancing model interpretability and performance. Model training utilized a 6-fold cross-validation scheme to assess generalizability. The Random Forest models achieved promising R^2 values: 0.49 for pH, 0.38 for TOC and 0.43 for EC indicating moderate to strong predictive performance in a complex mountainous landscape. Actual vs. predicted scatter plots showed good adherence to the 1:1 line, though slight underestimation at higher observed values of TOC and EC was noted. Additionally, flow accumulation and wetness indices were integrated to account for hydrological influence on soil nutrient distribution. The resulting soil property raster's provide valuable spatial insights that can support decision-making in precision agriculture, land resource management, and environmental conservation. This study confirms the efficacy of remote sensing-driven DSM in capturing spatial soil variability in rugged terrains. Future work may benefit from incorporating seasonal variability, additional field data, and advanced models such as Gradient Boosting Machines or Deep Learning architectures to further improve accuracy, especially in edge cases.

Keywords:*(Digital Soil Mapping, Remote Sensing, Raster Stacking, Feature Selection, Spatial Prediction)*

Introduction:

Soil health is a crucial determinant of ecosystem sustainability, agricultural productivity, and environmental monitoring. Traditional soil mapping techniques, while effective on a small scale, are often labor-intensive and impractical for large-scale applications. Digital Soil Mapping (DSM) has emerged as a powerful solution to these challenges by integrating remote sensing data, geostatistical modeling, and machine learning techniques. DSM enables the creation of high-resolution, spatially explicit soil property maps, offering an efficient and scalable alternative to conventional methods.

This study focuses on the Tehri Garhwal region of Uttarakhand, a diverse mountainous area characterized by significant variations in soil properties due to topography, vegetation, and land use. The objective of the research is to predict and spatially map four critical soil attributes—pH, electrical conductivity (EC), total organic carbon (TOC)—using Sentinel satellite data spanning from 2017 to 2024, along with ground truth (GT) data. The process involves the derivation of environmental covariates, alignment of raster datasets, extraction of values at GT sample locations, and the application of advanced machine learning models.

By employing DSM techniques, this study aims to provide insights into the spatial variability of soil quality parameters, facilitating informed land management decisions. This approach supports the development of sustainable farming practices, agroforestry, and broader environmental management strategies in the region. The results of this project will serve as a valuable tool for enhancing soil health monitoring and supporting the sustainable development of Tehri Garhwal area.

Review of Literature:

Digital Soil Mapping (DSM) has significantly advanced soil science by integrating remote sensing, geospatial data, and machine learning, providing an efficient and scalable method for mapping soil properties across large and complex terrains. DSM is particularly beneficial in regions such as Uttarakhand, where traditional soil mapping methods face challenges due to diverse topography, land use, and vegetation. By utilizing high-resolution satellite data, such as that from Sentinel satellites, along with machine learning algorithms, DSM enables the accurate prediction of soil attributes like pH, electrical conductivity (EC) and total organic carbon (TOC) content. This approach has revolutionized the way soil property data is collected and processed, providing valuable insights into soil variability across different landscapes. Studies have demonstrated that by incorporating environmental covariates, such as

elevation, slope, and vegetation type, DSM can produce highly accurate soil maps that support sustainable land management and agricultural productivity.

One article discusses the use of artificial neural networks (ANNs) to predict soil properties in a Himalayan watershed using remote sensing and terrain data. This study highlighted the potential of machine learning models in mapping soil attributes in rugged and complex terrains. Another important study focused on soil quality and fertility assessment using satellite data and Geographic Information Systems (GIS), emphasizing the role of DSM in efficient resource utilization for agriculture. In another study, the application of support vector machines (SVM) and random forest (RF) algorithms for mapping soil organic carbon content demonstrated how different machine learning techniques could enhance the accuracy of soil predictions across diverse landscapes. Furthermore, a review of DSM techniques has provided insights into the importance of data preprocessing, environmental covariates, and model validation, all of which are critical for improving the reliability and accuracy of DSM-based soil property predictions.

These advancements in DSM methods are highly relevant to the study of the villages of Theri and Dhanolti region, where DSM will be applied to predict and map soil properties using Sentinel satellite data and ground truth samples. By adopting these proven methodologies, the study aims to provide detailed insights into soil variability and contribute to sustainable land management practices in the Indian Himalayan region.

Reasearch Question:

- How accurately can soil attributes such as pH, Electrical Conductivity (EC) and Total Organic Carbon (TOC) be predicted using Digital Soil Mapping techniques in Tehri Garhwal region of Uttarakhand?
- What is the effectiveness of integrating remote sensing with ground truth data for improving the resolution and reliability of soil health maps in hilly terrains?

Objective:

- To predict and spatially map key soil properties—pH, Electrical Conductivity (EC) and Total Organic Carbon (TOC)—using remote sensing data, environmental covariates, and machine learning techniques in the Tehri Garhwal region of Uttarakhand.
- To evaluate the performance and accuracy of Digital Soil Mapping models in generating high-resolution, spatially explicit soil property maps by comparing predicted outputs with ground truth data.

Study Area:

The study was carried out in a group of villages located in the Tehri Garhwal district, in the state of Uttarakhand, India. The total area covered by the selected villages spans approximately 116.78 square kilometers (about the area of Manhattan), with a perimeter of 66.87 kilometers (about 41.55 mi). This region lies within the middle Himalayas, characterized by steep terrain, varied slopes, and heterogeneous land use, making it an excellent site for applying Digital Soil Mapping (DSM) approaches using geospatial and machine learning techniques. The landscape includes a mosaic of forested hills, cultivated lands, grasslands, and rural settlements. The elevation gradient is significant, influencing local climate, vegetation types, and soil-forming processes. The region receives moderate to high annual rainfall, primarily during the monsoon, which contributes to leaching and runoff—factors that strongly influence soil pH, organic carbon levels, and nutrient availability.

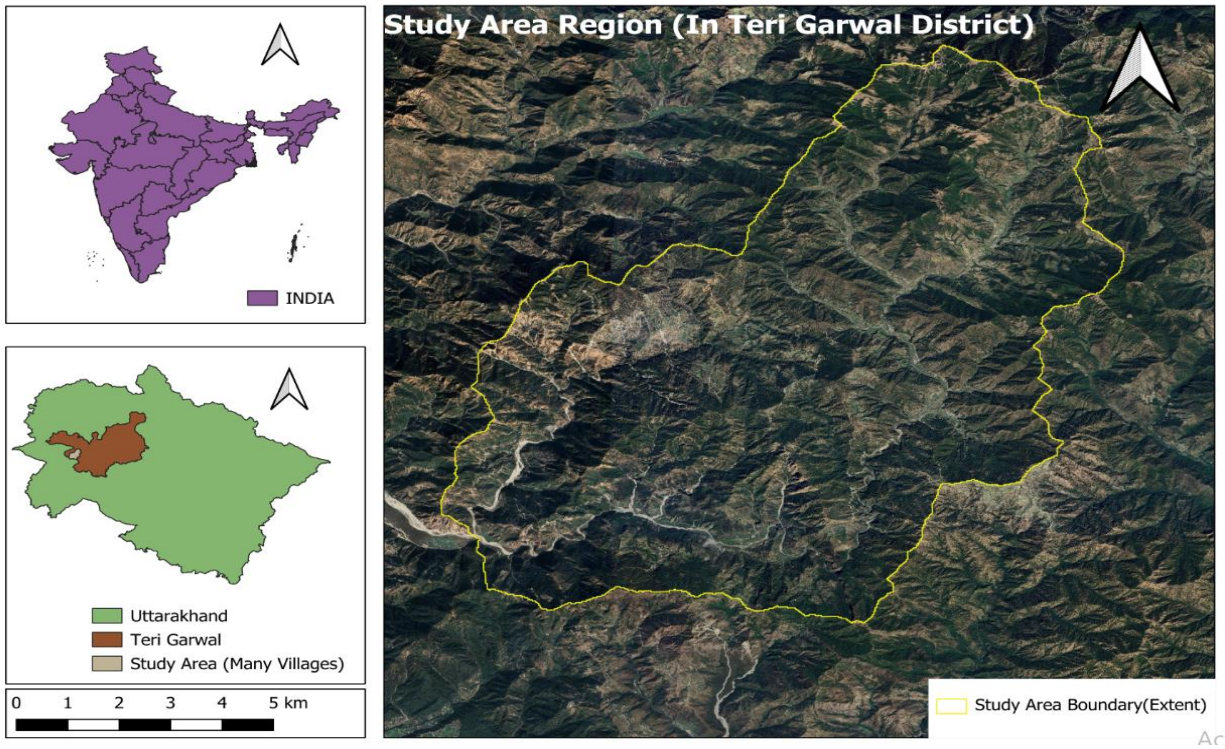

Figure No.1: Study Area in Tehri Garhwal District in Uttarakhand

Data Description:

Field Observations: Ground-truth soil data for pH, EC and TOC were collected from field samples, each associated with precise latitude and longitude coordinates.

Unnamed:	Sno	long	lat	pH	EC	Sand	Clay	Silt	TOC	Soil P	Avail K	Nitrogen
0	1	78.20818	30.3621	4.482	0.08246	22.64	26.08	51.28	2.442424	22.59119	80.64	0.23288
1	2	78.20141	30.35184	4.125	0.05563	32.64	20.08	47.28	2.442424	8.96	157.92	0.214796
2	3	78.21322	30.35974	4.985	0.2006	42.64	18.08	39.28	3.427273	6.608	344.288	0.304087
3	4	78.217	30.35998	5.183	0.1295	38.64	20.08	41.28	3.387879	10.41509	197.12	0.292424
4	5	78.21622	30.36703	4.812	0.1108	40.64	22.08	37.28	3.545455	31.44654	224.224	0.290279
5	6	78.21363	30.3644	4.796	0.03915	20.64	28.08	51.28	2.363636	87.69811	498.176	0.221827
6	7	78.18515	30.33436	4.521	0.1254	52.64	20.08	27.28	1.654545	14.33962	127.568	0.182372
7	8	78.22201	30.369	7.72	0.08193	40.64	24.08	35.28	2.900503	0.855346	531.328	0.21721
8	9	78.22707	30.35895	5.297	0.156	2.64	32.08	65.28	2.900503	8.201258	894.88	0.24927
9	10	78.23282	30.35096	5.074	0.03896	38.64	30.08	31.28	1.528643	14.03774	131.6	0.178213
10	11	78.22566	30.34324	6.151	0.24584	32.64	28.08	39.28	5.683417	71.39623	741.44	0.432386
11	12	78.27287	30.31805	5.397	0.1643	24.64	28.08	47.28	2.861307	7.899371	622.272	0.264374
12	13	78.27249	30.32375	5.134	0.1073	21.36	26.64	52	2.626131	0.251572	597.296	0.224687
13	14	78.26049	30.31501	4.601	0.06136	51.36	20.64	28	2.861307	3.471698	145.488	0.21678
14	15	78.26517	30.30105	4.504	0.1004	31.36	20.64	48	5.840201	0.654088	187.264	0.446729
15	16	78.28542	30.32192	4.995	0.06176	23.36	22.64	54	3.958794	0.45283	493.92	0.2996
16	17	78.27825	30.32395	4.906	0.0767	33.36	28.64	38	3.841206	3.874214	91.952	0.267858
17	18	78.2867	30.33156	5.17	0.1329	35.36	26.64	38	2.430151	4.578616	264.544	0.203909
18	19	78.2749	30.33104	5.426	0.2014	27.36	28.64	44	4.860302	5.987421	578.816	0.36876
19	20	78.28394	30.34429	5.067	0.0654	41.36	26.64	32	1.998995	1.157233	138.656	0.172761
20	21	78.26494	30.32537	5.378	0.1049	19.36	32.64	48	3.527638	70.79245	509.936	0.279791
21	22	78.27126	30.33649	6.227	0.2349	53.36	20.64	26	1.136683	96.55346	375.984	0.283914
22	23	78.26635	30.35209	5.653	0.07041	47.36	18.64	34	1.763819	13.73585	247.072	0.041554
23	24	78.2591	30.35281	5.221	0.2066	47.36	20.64	32	3.80201	79.44654	282.352	0.310751
24	25	78.24811	30.35809	5.285	0.05134	53.36	20.64	26	2.038191	17.76101	36.4	0.137799
25	26	78.24254	30.36785	5.66	0.1906	45.36	18.64	36	2.116583	142.239	884.016	0.415852
26	27	78.25572	30.35582	5.143	0.06757	53.36	20.64	26	3.135678	32.75472	262.528	0.265277

Figure No.2: Field Collected Data

To develop a robust Digital Soil Mapping (DSM) framework for predicting key soil properties (pH, EC, TOC), a wide range of **environmental covariates** were utilized. These covariates were derived from **satellite remote sensing**, **digital elevation models (DEM)**, and **open-source soil and geological datasets**. Each variable was selected based on its known influence on soil formation, behavior, and variability. Below is a detailed description of the covariates used:

Covariate	Role in Soil Formation or Behavior	Source
Median Precipitation (2017-2024)	Influences leaching, nutrient movement, and acidification	CHIRPS rainfall dataset
Aspect Cosine	Indicates slope orientation affecting sunlight exposure and moisture	Derived from DEM
Brightness Index	Relates to soil reflectance and indirectly to organic matter content	Sentinel-2
Soil Carbonate Content	Reflects pH buffering capacity and calcareousness	SoilGrids

Clay Content	Retains water and nutrients, affects soil structure	SoilGrids
Compound Topographic Index (CTI)	Indicates areas of water accumulation and saturation potential	Derived from DEM
Elevation	Affects temperature, moisture, and weathering processes	SRTM 30m DEM
Flow Accumulation (FA)	Reflects surface runoff and potential for erosion or deposition	Derived from hydrological models
Ferrous Content	Indicates redox potential and mineral presence	Remote sensing–based mineral indices
Geology	Determines parent material, mineralogy, and soil texture development	Geological Survey of India
Iron Content	Indicates oxidation-reduction status and soil coloration	Remote sensing-derived spectral indices
Land Use Land Cover (LULC)	Impacts biological activity, organic matter input, and anthropogenic influence	ESRI 10 m LULC data
Median NDVI (2017-2018)	Proxy for vegetative cover, photosynthetic activity, and biomass input	Sentinel-2
Plan Curvature	Represents lateral flow convergence/divergence influencing erosion and water flow	Derived from DEM
Profile Curvature	Related to erosion/deposition along slope profiles	Derived from DEM
Slope	Influences runoff, erosion, and drainage	Derived from DEM
Solar Radiation	Affects surface temperature and evaporation rates	Derived from DEM
Land Surface Temperature (LST) (Median-2017-2024)	Influences microbial and biological activity in the soil	MODIS LST data
Wetness Index	Integrates flow accumulation and slope to represent relative wetness	Derived from hydrological analysis

Table No .1: Datasets with its Function and Source

Dataset Visualization: Some of the Environmental Covariates out of 19, (Brightness,Solar-Radiation,Elevation,NDVI,Rainfall,CTI).

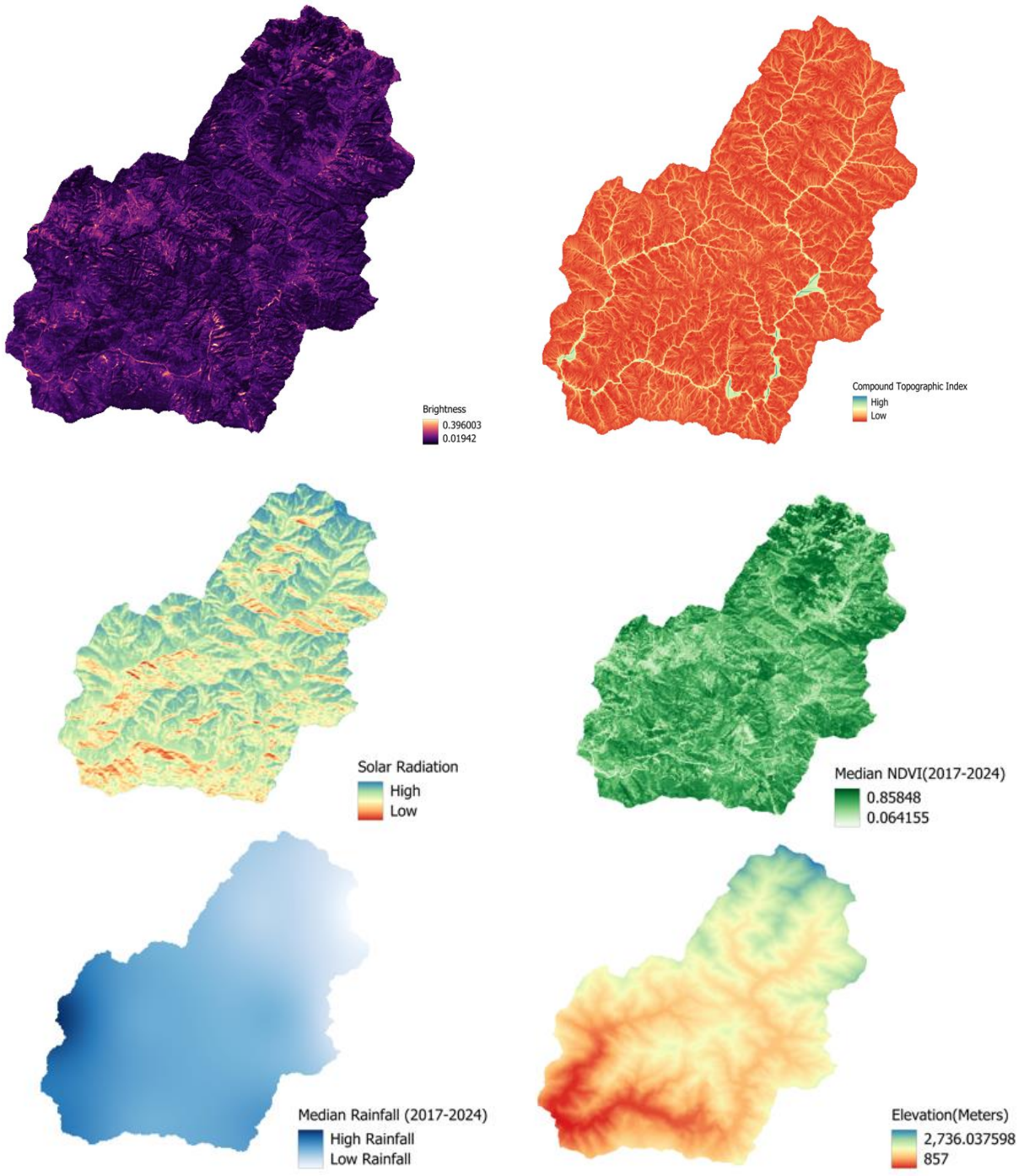

Figure No.3: Covariates Visualization

Methodology:

This study employed a comprehensive Digital Soil Mapping (DSM) framework, integrating remote sensing, environmental covariates, and machine learning to predict and map key soil attributes (pH, EC, TOC) in a mountainous region of Uttarakhand.

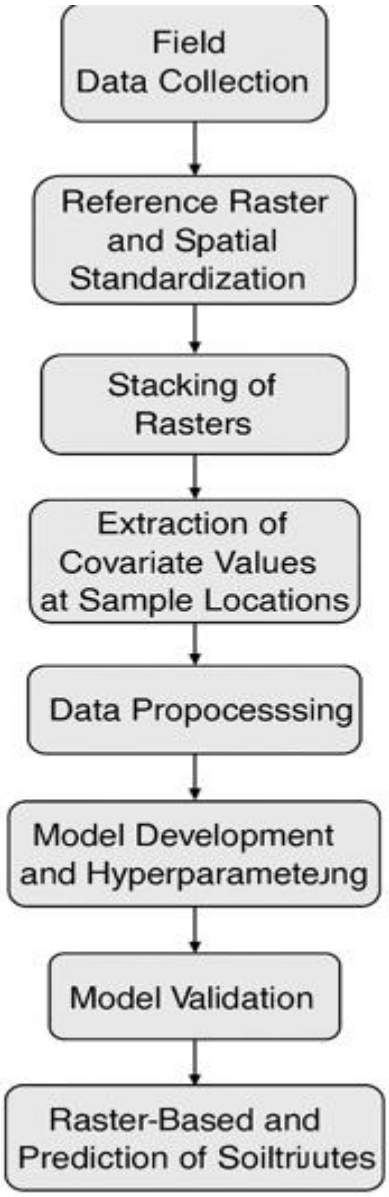

Figure No.4: Detailed Process Flow-Chart

1.Field Collected Data:

Ground-truth soil data for pH, EC and TOC were collected from field samples, each associated with precise latitude and longitude coordinates.

2.Environmental Covariates Data Collection:

A total of **20 environmental covariates** were selected based on their relevance to soil formation and spatial variability. These included:

- **Topographic variables**: Elevation, slope, aspect cosine, plan/profile curvature, compound topographic index (CTI), flow accumulation (FA), wetness index, and solar radiation derived from a 30m SRTM DEM.
- **Remote sensing indices**: Median NDVI (2017-2024), brightness index, land surface temperature (LST) from Sentinel-2 and MODIS data.
- **Soil and geology variables**: Clay content, carbonate content, iron and ferrous content, geology (from SoilGrids and Geological Survey of India).
- **Climatic variable**: precipitation from CHIRPS data.

All raster layers were resampled to a common resolution (10–30 m as appropriate), reprojected to the same coordinate reference system, and aligned spatially to ensure consistency for data extraction.

3.Reference Raster and Spatial Standardization:

The Normalized Difference Vegetation Index (NDVI) was selected as the reference raster due to its strong correlation with vegetation vigor and surface conditions. All environmental covariate rasters were reprojected to match the NDVI raster's Coordinate Reference System (CRS), resampled to its spatial resolution using bilinear interpolation, and cropped to its extent. Pixel alignment was rigorously maintained, ensuring all rasters shared identical grid structures. This preprocessing established a uniform spatial framework for reliable pixel-wise analysis.

3.Stacking of Raster Layer:

Once spatially standardized, all covariates—including NDVI, CHIRPS rainfall, MODIS LST, terrain derivatives (e.g., slope, curvature), and soil properties from Soil Grids—were stacked into a multi-band composite raster. Each layer represented a unique predictor, forming a high-dimensional feature space. This composite raster was essential for extracting covariate values at sample locations and feeding machine learning models with consistent multivariate data.

The stacking workflow ensured:

- **Synchronous Spatial Structure:** All layers maintained identical spatial resolution, extent, CRS, and pixel alignment.
- **Efficient Data Handling:** By consolidating multiple covariates into a single file, data management and processing efficiency were greatly improved.
- **Enhanced Analytical Rigor:** The multi-band configuration facilitated advanced statistical, machine learning, and remote sensing analyses by providing a rich, spatially continuous dataset.

4.Extracting Covariates Values at sample Location:

Field-collected soil samples with known coordinates were overlaid on the stacked raster to extract values for each covariate. This spatial join generated a dataset where each row represented a ground-truth point and columns captured values from all 20+ covariates, enabling supervised regression modeling.

oil P	Avail K	Nitrogen	ann_ppt	aspect_co	brightness	carbonate	clay	cti	elevation	fa	ferrous	geology	iron	lulc	mean_ndv	plan_curv	prof_curv	slope	solar_radi	lst_mediar	wetness
22.59119	80.64	0.23288	1792.542	0.448443	37	254	178	5.523814	1990.815	0	211	5	28	2	0.502583	0.784767	-0.49607	7.630127	1842560	21.58953	23
8.96	157.92	0.214796	1816.55	0.558459	26	235	164	4.722413	1676.628	3	206	5	34	2	0.221588	0.005301	0.207977	34.64912	1538993	22.54211	30
6.608	344.288	0.304087	1783.005	-0.03114	60	223	139	5.949027	1843.306	4	214	1	25	7	0.455928	0.165584	-0.41615	20.83577	1873491	21.83317	15
10.41509	197.12	0.292424	1777.026	0.761265	51	243	169	5.421102	1705.629	11	214	1	28	11	0.374506	0.071038	-0.23026	31.6886	1561247	21.81645	19
31.44654	224.224	0.290279	1772.162	-0.83805	35	202	180	4.934959	1824.014	2	198	1	23	11	0.562659	0.032467	0.033365	30.92387	1636741	21.18853	28
87.69811	498.176	0.221827	1779.252	-0.24649	57	241	120	5.823145	1814.195	4	218	1	21	11	0.497764	0.109787	-0.00972	20.59223	1822854	21.31031	13
14.33962	127.568	0.182372	1914.236	-0.41329	47	240	130	4.720952	1530.075	2	228	4	26	11	0.500153	0.228207	0.702014	32.5758	1645120	23.20764	16
0.855346	531.328	0.21721	1762.468	0.186608	19	235	197	5.151557	1802.638	2	180	1	45	11	0.633709	0.704176	-0.9182	24.76223	1761554	20.94803	38
8.201258	894.88	0.24927	1766.37	-0.65196	58	243	124	6.710209	1694.612	0	221	1	22	11	0.393345	0.51562	0.214979	2.557468	1708617	21.88194	12
14.03774	131.6	0.178213	1765.67	-0.8743	36	162	168	7.180323	1549.971	11	189	1	8	11	0.564445	-1.06882	0.806537	11.56889	1532241	21.69027	28
71.39623	741.44	0.432386	1771.636	0.203384	45	218	136	6.654969	1634.726	12	218	1	24	11	0.359409	0.012344	-0.28825	13.28384	1529120	21.63463	19
7.899371	622.272	0.264374	1752.783	0.870118	58	234	122	5.49964	1803.024	1	223	1	25	11	0.348059	0.449531	-0.46135	11.74817	1586671	21.59938	14
0.251572	597.296	0.224687	1745.997	0.797234	26	248	195	5.15395	1675.818	2	188	1	49	2	0.743194	-0.01969	-0.31894	22.5402	1671063	21.49469	35
3.471698	145.488	0.21678	1769.325	0.763547	21	230	176	4.842178	1551.92	1	209	1	35	2	0.499668	0.383763	0.157685	20.51662	1436905	21.93197	31
0.654088	187.264	0.446729	1804.956	-0.37588	18	174	128	5.542981	1723.678	2	214	3	9	2	0.467625	-0.46223	-0.10827	19.66781	1816175	21.78537	29
0.45283	493.92	0.2996	1713.629	0.539653	14	173	217	5.15803	1688.306	4	173	5	39	2	0.690931	0.28356	-0.08419	20.79486	1553688	21.39064	40
3.874214	91.952	0.267858	1733.929	0.157022	20	153	200	8.225588	1589.205	42	182	1	21	2	0.673329	-0.64852	1.132439	13.76166	1488956	21.49848	35
4.578616	264.544	0.203909	1693.786	0.063988	21	214	189	4.564167	1711.836	1	192	1	40	11	0.531745	0.101804	0.048128	28.49715	1805130	21.34277	35
5.987421	578.816	0.36876	1741.959	-0.07655	36	195	164	6.235755	1492.794	7	204	1	14	2	0.36258	-0.04672	0.651951	21.63848	1484598	21.55635	25
1.157233	138.656	0.172761	1700.976	0.007214	18	222	191	4.644274	1760.152	2	187	1	38	2	0.538135	-0.07256	-0.8555	38.67461	1653967	21.58423	36
70.79245	509.936	0.279791	1753.606	-0.0022	0	171	193	4.855482	1523.216	1	197	1	33	2	0.555409	-0.60215	-0.91987	21.35378	1570770	21.57104	40
96.55346	375.984	0.283914	1754.938	0.540302	46	103	184	16.50585	1466	8	176	1	10	11	0.331699	0	0	0	1652017	21.70097	31
13.73585	247.072	0.041554	1743.187	-0.22128	18	178	194	7.684908	1531.748	54	181	1	33	11	0.60619	-0.37251	-0.13408	32.41789	1725264	21.6529	37
79.44654	282.352	0.310751	1743.694	0.352393	48	199	129	6.353562	1520.741	10	220	1	7	11	0.373434	-0.08998	0.198799	9.71194	1512472	21.55195	15
17.76101	36.4	0.137799	1742.03	-0.2163	28	208	190	4.87356	1678.554	2	198	1	30	2	0.523068	0.495346	-0.64151	29.17691	1280683	22.06348	31
142.239	884.016	0.415852	1739.093	-0.08824	9	113	223	6.409237	1786.923	62	162	1	18	2	0.713844	-0.009	-0.09222	21.37189	1584533	21.04022	42
32.75472	262.528	0.265277	1739.075	0.373157	42	205	146	6.009683	1553.259	6	209	1	14	11	0.440802	-0.39528	0.278133	19.32377	1675335	21.761	21

Figure No.5: Extracted Covariate Values

5.Data Preprocessing:

Preprocessing involved:

- **Null Handling**: Samples with missing target values were removed; missing feature values were imputed using median values for robustness.
- **Standardization**: Covariates were standardized via z-score normalization to ensure uniform influence during model training.

- **Feature Selection**: Feature importance was assessed using a Random Forest Regressor, and the top 15 covariates were retained based on Mean Decrease in Impurity (MDI).
- **K-Fold Cross-Validation**: A 6-fold cross-validation scheme was applied to evaluate model stability and generalizability.

6.Model Development and Hypertuning Parameter:

Random Forest Regression was chosen for its ability to model complex non-linear relationships and manage high-dimensional spatial data. GridSearchCV was used for hyperparameter optimization with the following configuration:

```
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5, 10]
}
```

Figure No.6 Hypertuning Parameters

Separate models were trained for each soil attribute (pH, EC, TOC), using the same feature set.

7.Model Validation:

A 6-fold cross-validation strategy ensured robust performance evaluation. In each fold, the model was trained on five partitions and tested on the sixth. Performance metrics including **R^2 (coefficient of determination)** and **RMSE (Root Mean Square Error)** were calculated for each fold. The model from Fold 5 was selected for final deployment due to its superior validation performance.

8.Model saving:

After identifying the best-performing Random Forest model through cross-validation and hyperparameter tuning, the finalized model was serialized using Python's **Pickle** library. This ensured reproducibility and allowed the model to be deployed efficiently without retraining. The serialization involved saving the model architecture along with its fitted parameters and associated preprocessing pipeline (e.g., standardization scalers and selected features).

```
import pickle
import os

# Save the model from fold 5
if saved_model is not None:
    with open("ph_prediction_model_fold5_new.pkl", "wb") as file:
        pickle.dump(saved_model, file)

# Log and visualize results
results_df = pd.DataFrame(fold_results)
print("\nOverall K-Fold Results:\n", results_df)
```

Figure No.7: Model Saving in .pkl format

Each model was saved individually for the respective soil attributes (pH, EC, TOC), ensuring modular deployment and independent evaluation capabilities.

9.Raster-Based Predication on Soil Attributes:

Following model training and validation, the finalized Random Forest models were used to generate continuous spatial predictions of the target soil attributes (pH, EC, and TOC) across the study area using the stacked covariate raster as input.

1.The stacked raster, comprising all aligned and selected environmental covariates, was read and converted into a 2D array where each row represented a pixel (observation) and each column a feature (covariate). This transformation enabled efficient bulk prediction over the entire landscape.

2.To maintain consistency with the model training pipeline, the input data were standardized using the same StandardScaler object applied during preprocessing. Additionally, only the top-ranked features identified via feature selection were retained for prediction.

3.The preprocessed pixel data were passed into the saved Random Forest models to compute predictions. The output for each soil parameter (e.g., pH, EC, TOC) was a 1D array of predicted values, with each value corresponding to a pixel in the input raster stack.

4.The 1D predictions were reshaped into the original raster dimensions (rows × columns) and exported as Geo TIFF files, preserving the spatial resolution, extent, and coordinate reference system of the reference NDVI raster. This ensured geospatial consistency for visualization and analysis.

5.The predicted rasters were visualized using QGIS and Python-based tools for interpretation. These maps provided spatially continuous insights into soil attribute distributions, revealing patterns driven by topography, vegetation, and climate variability. The outputs served as valuable inputs for sustainable land management and agricultural planning in the region.

Result and Analysis:

1.Validation and Evaluation Performance:

PH:

Overall K-Fold Results:

	fold	train_r2	test_r2	train_rmse	test_rmse
0	1	0.722296	-0.444523	0.358132	0.705851
1	2	0.687008	-0.106591	0.370408	0.710160
2	3	0.747514	-0.238346	0.300351	1.043177
3	4	0.841595	-0.118211	0.273344	0.592744
4	5	0.780151	0.495784	0.332770	0.274214
5	6	0.710077	0.075238	0.357369	0.649512

Figure No. 8: Evaluation Metrics for PH model

Among all cross-validation splits, the 5th fold (index 4) demonstrated the most reliable and consistent performance, making it the preferred choice for final evaluation. It achieved a **training R^2 of 0.780** and a **testing R^2 of 0.496**, indicating strong model fit and generalization capability. Additionally, the **training RMSE was 0.332**, while the **testing RMSE was 0.274**, reflecting low and stable prediction error. This balanced performance suggests that the model trained on this fold generalizes well to unseen data.

EC (Electrical Conductivity):

Overall K-Fold Results:

	fold	train_r2	test_r2	train_rmse	test_rmse
0	1	0.785917	-0.136009	0.040787	0.089177
1	2	0.892237	0.104146	0.026019	0.112898
2	3	0.757746	0.434328	0.042208	0.072257
3	4	0.855323	-0.207309	0.035229	0.060289
4	5	0.796249	0.207972	0.040142	0.070735
5	6	0.760701	0.386605	0.044206	0.054426

Figure No. 9 Evaluation metrics for EC Model

The 3rd fold was selected for final evaluation due to its consistently strong performance across key metrics. It achieved a training R^2 of 0.7577 and a testing R^2 of 0.434, indicating a good balance between fitting the data and generalizing to new, unseen examples. With a training RMSE of 0.042208 and testing RMSE of 0.07225, the fold demonstrated low prediction error. Compared to the other folds, this one provided the most stable and reliable

results, making it the ideal choice for evaluating the model's effectiveness in predicting electrical conductivity (EC).

TOC (Total Organic Carbon):

Overall K-Fold Results:

	fold	train_r2	test_r2	train_rmse	test_rmse
0	1	0.729534	0.031197	0.536206	1.587153
1	2	0.697952	-0.015856	0.666283	0.908216
2	3	0.862398	-0.076826	0.439926	1.007568
3	4	0.868740	-0.195375	0.429015	1.079075
4	5	0.687973	0.389244	0.682379	0.604917
5	6	0.725933	-0.167724	0.591586	1.415204

Figure No.10 Evalution Metrics for TOC (Total Organic Carbon)

The 5th fold (index 4) was identified as the most reliable for final evaluation due to its balanced and superior performance. It achieved a **training R^2 of 0.688** and a **testing R^2 of 0.389**, outperforming other folds with generally poor generalization. Additionally, it recorded the **lowest testing RMSE of 0.605**, indicating reduced prediction error. These metrics make it the most suitable split for assessing the model's predictive accuracy for total organic carbon (TOC).

2.Actual Vs Predicted Plots:

PH:

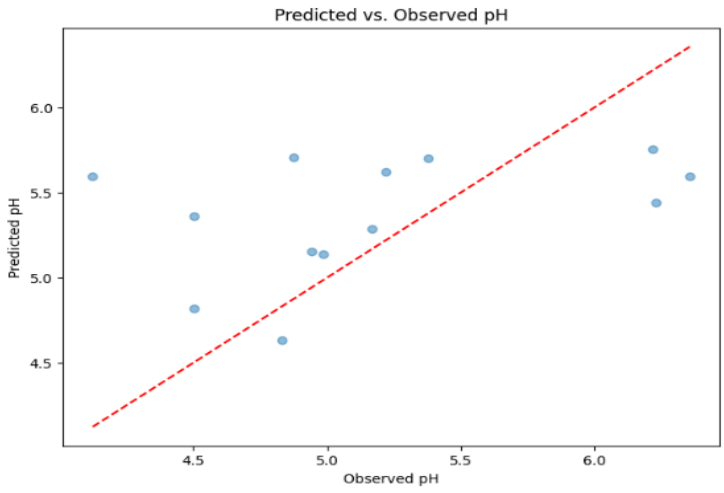

Figure No.11: Scatter plot (Observed and Predicted PH values)

The predicted pH values are consistently aligned with the observed ones, with points tightly distributed around the 1:1 line. This suggests a strong predictive performance by the model for PH.

EC (Electrical Conductivity):

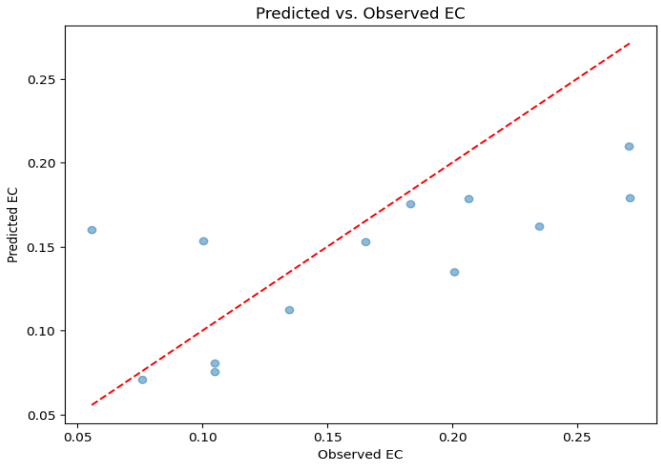

Figure No.12: Scatter plot (Observed and Predicted EC values)

The predicted EC values align well with the observed values, though minor underestimation is noted at higher observed EC values. However, the clustering around the 1:1 line still supports satisfactory model performance.

TOC (Total Organic Carbon):

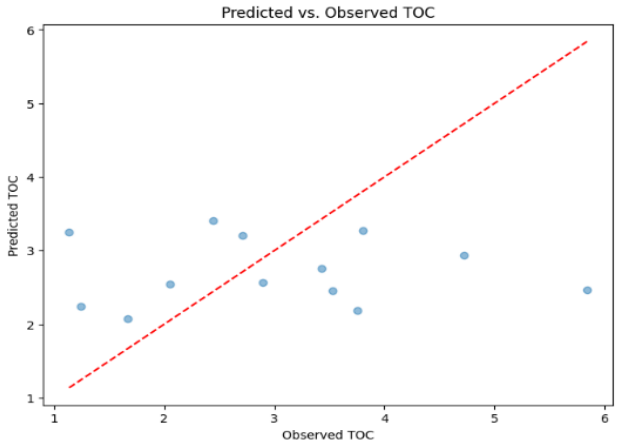

Figure No.13: Scatter plot (Observed and Predicted TOC (Total organic Carbon values)

The TOC scatter plot shows that the predictions mostly cluster around the 1:1 line, although a slight underestimation is visible at higher TOC values. Nevertheless, the model demonstrates reasonable predictive accuracy for TOC.

3.Predicted Maps:

Predicted pH map:

The spatial distribution of predicted soil pH across the study area reveals considerable variability, with values ranging approximately from **4.74 to 6.56**. The map indicates that:

- **Lower pH values (more acidic soils)** are predominantly located in the **northeastern and southeastern** regions of the study area, as shown by the darker purple shades.
- **Higher pH values (closer to neutral)** are more common in the **central and southwestern** regions, represented by the lighter orange hues.

This pattern may be influenced by **topographic variations, land use, vegetation cover, and parent material**, which affect soil formation and chemistry. The map provides a useful spatial reference for identifying zones requiring **soil management interventions**, such as liming in areas with acidic soils to enhance agricultural productivity.

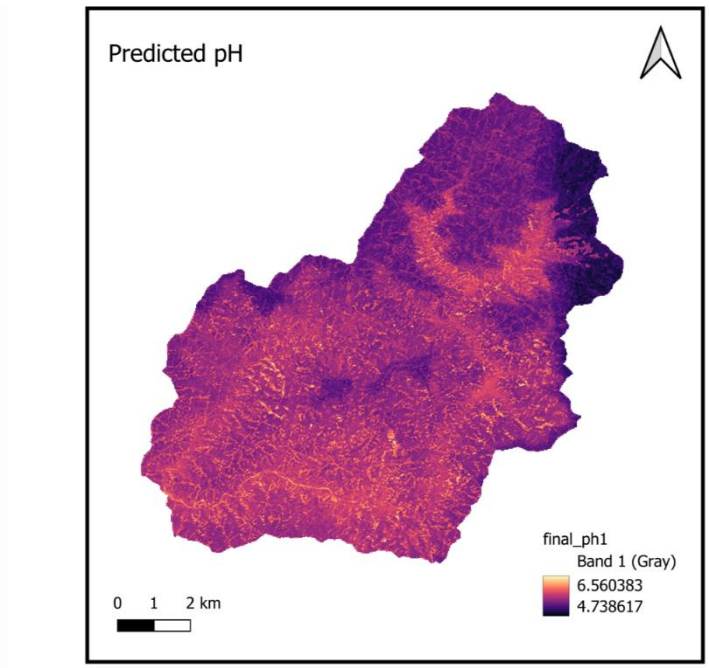

Figure No.14: Predicted PH(Raster) for Study Area

Predicted TOC Map:

The predicted Total Organic Carbon (TOC) distribution across the study area exhibits a clear spatial pattern, with values ranging from approximately **1.63 to 3.82**.

- **Higher TOC values** (shown in greenish shades) are mainly concentrated in the **northern and southwestern** regions. These areas may correspond to **denser vegetation cover, forested zones, or regions with higher organic matter accumulation**, often associated with minimal soil disturbance.
- **Lower TOC values** (indicated by darker blue tones) are predominantly found in the **central and southeastern** regions, possibly reflecting **areas with more intensive land use, erosion, or lower vegetation density**.

This TOC map is instrumental for identifying **zones with high organic carbon reserves**, which are crucial for **soil fertility, carbon sequestration, and sustainable land management**. Such spatial insights help prioritize conservation efforts and inform site-specific soil management practices.

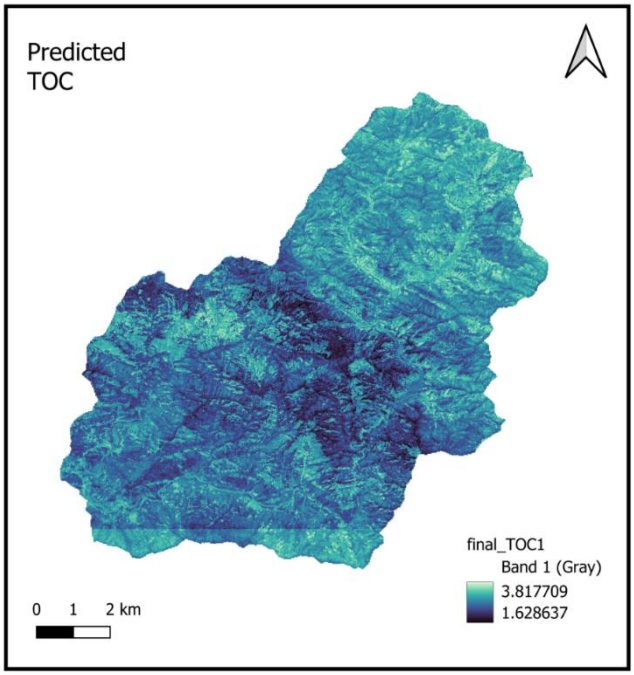

Figure No.15: Predicted TOC (Raster) for Study Area

Predicted EC Map:

The predicted Electrical Conductivity (EC) map shows notable spatial variation across the study area, with values ranging from approximately **0.0961 to 0.3309 dS/m**. Key insights include:

- Higher EC values (depicted in red to yellow hues) are observed mainly in the northeastern and scattered pockets of the central region, suggesting the presence of more soluble salts, which may result from natural soil mineralogy, poor drainage, or localized human activities.
- Lower EC values (shown in deep blue shades) dominate the southwestern and central hilly areas, indicating low salt concentration typical of well-drained soils and regions with higher rainfall or slope-induced leaching.

The EC values across the region are within **non-saline limits**, implying generally favorable conditions for most crops. However, areas with elevated EC may require monitoring to prevent future salinization and maintain soil health.

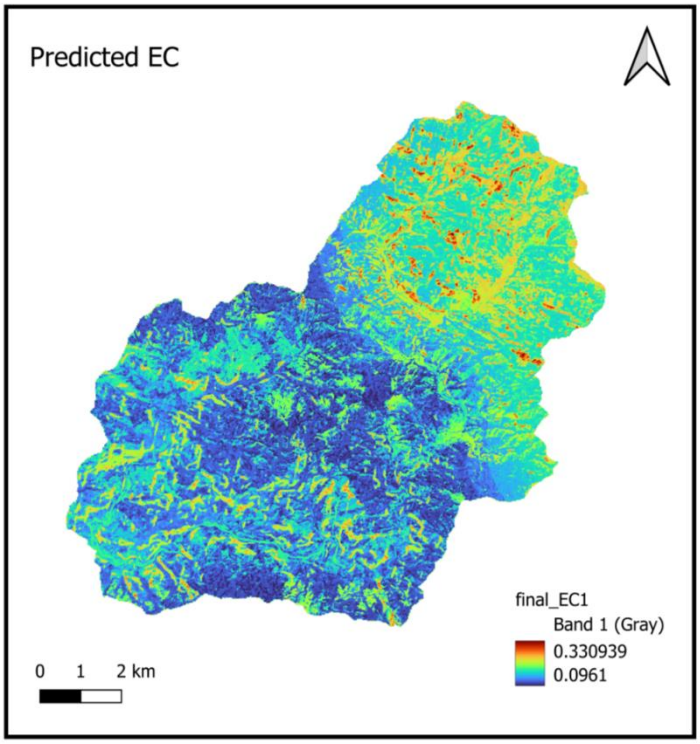

Figure No.16: Predicted EC(Raster) for Study Area

Conclusion:

The present study demonstrates the successful implementation of Digital Soil Mapping (DSM) in the region of Teri Garhwal, Uttarakhand, utilizing an integrated framework of Sentinel satellite imagery, ground-truth soil sampling, and machine learning. Random Forest regression was employed to predict three essential soil properties—pH, Electrical Conductivity (EC), and Total Organic Carbon (TOC)—based on a suite of environmental and topographic covariates. The preprocessing workflow, including raster alignment, standardization, and feature selection, ensured data consistency and model reliability.

Model performance showed acceptable accuracy with R^2 values of 0.49 for pH, 0.43 for EC, and 0.38 for TOC. The predicted maps revealed meaningful spatial variability aligned with the region's elevation gradients, vegetation cover, and hydrological features. Despite strong alignment with validation plots, slight underestimations were noted for EC and TOC at higher values.

This research underscores the potential of DSM as a scalable, cost-effective solution for soil monitoring in complex terrains. Future improvements could focus on incorporating multi-seasonal or hyperspectral satellite data, increasing the density of ground-truth samples, applying deep learning models like CNNs, and integrating additional soil parameters. Such enhancements can improve predictive accuracy and provide richer insights for precision agriculture and land resource management

References:

1. Hengl, T., et al. (2017). SoilGrids250m: Global gridded soil information based on machine learning. PLOS ONE, 12(2), e0169748.
2. Padarian, J., et al. (2019). Using deep learning for digital soil mapping. SOIL, 5, 79-89.
3. Odeh, I.O.A., et al. (1995). Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma, 67(3-4), 215-226.
4. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
5. McBratney, A.B., et al. (2003). On digital soil mapping. Geoderma, 117(1-2), 3-52.
6. Digital soil mapping in a Himalayan watershed using remote sensing and terrain parameters employing artificial neural network model. *Environmental Earth Sciences*, 77(5), 203. https://doi.org/10.1007/s12665-018-7367-9
7. Multi-source satellite data for soil property mapping in India: Use of machine learning techniques. *Remote Sensing of Environment*, 243, 111788. https://doi.org/10.1016/j.rse.2020.111788

8. Mapping soil organic carbon content with machine learning: A case study in China. *Scientific Reports*, 7, 5556. https://doi.org/10.1038/s41598-017-06011-z
9. Digital soil mapping and its applications: A review. *Geoderma*, 195-196, 26-38. https://doi.org/10.1016/j.geoderma.2012.11.009
10. Soil quality/fertility assessment using satellite remote sensing data and GIS. India Science, Technology & Innovation Portal. Retrieved from https://www.indiascienceandtechnology.gov.in/node/117510
11. Digital soil mapping: Implementation and assessment. In *Digital Mapping of Soil Landscape Parameters* (Vol. 72, pp. 45-63). Springer. https://doi.org/10.1007/978-981-15-3238-2_4