UNIVERSITY OF MISKOLC

Faculty of Earth Science and Engineering

Institute of Geography and GeoInformatics

Mikoviny Sámuel Doctoral School of Earth Sciences

Head of the Doctoral School: Dr. Péter Szűcs, professor

SOIL MOISTURE CHARACTERIZATION AND SOIL TEXTURE PREDICTION USING MACHINE LEARNING AND MULTI-SOURCE ENVIRONMENTAL DATA

Ph.D. Dissertation Booklet

By:

MOHAMED RAJHI

Scientific supervisor:

Prof. Endre Dobos

Miskolc, 2025

HUNGARY

1. Introduction

The comprehension of soil moisture has evolved from historical agricultural knowledge to a pivotal focus in modern environmental and earth sciences. In the 3rd century BCE, agronomists like Mago of Carthage emphasized the importance of soil type, moisture retention, and irrigation scheduling in sustainable agriculture, principles subsequently reaffirmed by Roman scientists like Columella and Varro (Columella, 1941; Cato the Elder, 1934; Daniel J. Hillel., 1991; White, 1970). These early practices, grounded in empirical observation, demonstrate a fundamental understanding of the soil–water–plant relationship.

The launch of microwave satellites, including NASA's Skylab in the 1970s and following by the passive sensors on SMMR (1978), AMSR-E (2002), SMOS (2009), and SMAP (2015), enabled the integration of soil moisture into the domain application of remote sensing (Njoku & Entekhabi, 1996). These missions made a significant advancement in the global monitoring of surface soil moisture and frequent revisits, however at low spatial resolutions and limited depth sensitivity (Kerr et al., 2010; Zribi et al., 2011). Simultaneously, active radar missions (ERS, Envisat, Sentinel-1) provided greater resolution datasets but are influenced by surface roughness, vegetation, and dielectric characteristics and require advanced algorithms to derive dependable moisture signals. Optical sensors (e.g., Sentinel-2) offer indirect vegetation-derived measures of moisture stress, such as NDVI and NDWI, which can enhance microwave observations but are constrained by cloud cover.

However, the spatial and temporal heterogeneity of soil moisture makes it difficult to observe comprehensively. Soil moisture is influenced by soil texture and structure, weather, vegetation cover, topography, irrigation, and land management practices. These vary not only in space but also across soil depth and time, making point measurements poorly representative of larger areas (Grayson & Western, 1998; Ochsner et al., 2013a). Although in-situ sensors provide great precision, their application is limited to experimental sites and does not account for broad-scale variability.

Soil texture is crucial for calibrating soil moisture sensors in the field, as sensor performance is dependent upon the dielectric characteristics of the soil matrix (Topp et al., 1980; Robinson et al., 2008). Conventional methods to determine soil texture, utilizing laboratory analyses such hydrometer or laser diffraction techniques, are time-consuming, labor-intensive, and costly, particularly when applied over extensive areas (Minasny & McBratney, 2016). This constraint limits the spatial resolution especially in resource-constrained areas.

To address the limitations of traditional laboratory-based texture analysis, inverse modeling approaches have gained increasing attention. In this context, soil texture is inferred indirectly from the dynamics of soil moisture, rather than measured directly. Soil texture influences the rate and magnitude of soil moisture fluctuations following precipitation or irrigation, allowing for the prediction of textural classes or fractions through the analysis of the temporal dynamics of soil moisture profiles (Mohanty et al., 2017; Vereecken et al., 2016). Inversion-based techniques, whether integrated with sensor data or remote sensing soil moisture, present an attractive method for non-invasive, scalable texture mapping. This method is especially beneficial in agricultural monitoring systems where the expense or practicality of frequent ground sampling is constrained.

Research Rationale and the Relevance of the study

The rationale for focusing on two climatically distinct regions, Tunisia (semi-arid, water-stressed) and Hungary (temperate, data-rich), is to evaluate and characterize the soil moisture across heterogeneous landscapes. Tunisia's Merguellil basin offers a relevant test site for assessing the potential of remote sensing in regions with limited monitoring infrastructure, whereas Hungary provides a reference environment with established datasets and field instrumentation. This dual-site strategy allows for rigorous model calibration, validation, and transferability testing under varying soil textures, land cover types, and climatic gradients.

The scientific contribution of this thesis lies in advancing digital soil mapping (DSM) techniques through a fusion of EO data, in-situ measurements, and machine learning. The relevance extends to precision agriculture, drought risk management, and climate adaptation. By providing spatially explicit soil moisture maps and texture predictions, the research supports data-driven decisions in water allocation, crop management, and land use planning.

In summary, this study addresses critical scientific and operational gaps in soil moisture estimation. It demonstrates that remote sensing and AI-based models, when calibrated with reliable in-situ data, can generate high-resolution outputs for sustainable land and water management, particularly valuable in regions where traditional monitoring is scarce or fragmented.

2. Aim and objectives

The primary aim of this doctoral research is to develop an integrated framework for accurate, non-invasive estimation of soil moisture using multi-source data, including Sentinel-1 SAR, Sentinel-2 optical imagery, in-situ sensor networks, and meteorological variables, combined through machine learning techniques. The study targets operational soil moisture monitoring at multiple spatial scales and in contrasting agroecological contexts, specifically the semi-arid Merguellil basin in Tunisia and temperate agricultural zones in Hungary.

To fulfil this aim, the research is structured around the following objectives:

1. To assess the role of soil physical properties, including texture, porosity, structure, and surface characteristics, in controlling soil moisture dynamics, using both laboratory-measured and remote-sensing-derived parameters.
2. To support the calibration and validation of the in-situ soil moisture measurements from Sentek EnviroSCAN capacitance probes across different depths (10 cm, 20 cm, 30 cm), ensuring reliable ground-truth data for model training and validation.
3. To extract and process Sentinel-1 backscatter coefficients (VV, VH) and Sentinel-2 vegetation indices (NDVI, NDWI) for monitoring vegetation–soil–water interactions at plot and watershed scales.
4. To construct predictive models using Random Forest regression algorithms for estimating soil moisture and soil texture classes (sand, silt, clay) from environmental and satellite-derived covariates.
5. To evaluate the spatio-temporal variability of soil moisture across the study area by integrating sensor data, topographic parameters, and climatic drivers (temperature, rainfall, evapotranspiration).
6. To contribute to the advancement of digital soil mapping (DSM) by demonstrating the effectiveness of combining remote sensing, in-situ data, and machine learning for soil moisture and soil texture estimation.

3. Soil moisture characterization using optical and radar remote sensing data

3.1. Study area

The study area (Figure 1), the Merguellil basin in Kairouan (central Tunisia), spans 658,000 ha and combines a mountainous upstream (El Haouareb Dam watershed) with a downstream alluvial plain, making it a diverse yet erosion-prone landscape.

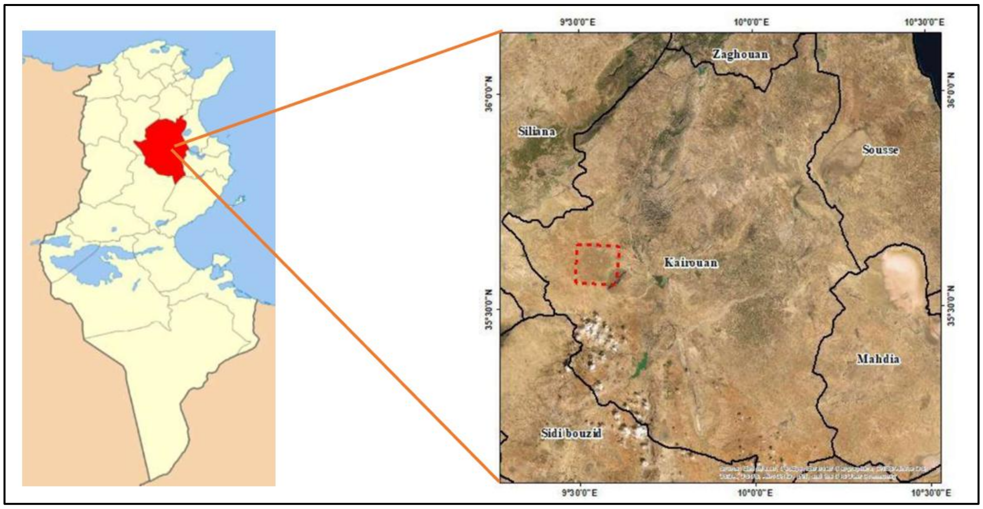


Figure 1. Study area Merguellil downstream basin of the Kairouan plain (Lacombe, 2007)

3.2. Methods

The diagram (Figure 2) illustrates the used workflow for characterizing soil moisture through the integration of multi-source remote sensing and environmental data. It begins with a preliminary area analysis, followed by the collection of three main data types: Sentinel-1 radar imagery, Sentinel-2 optical imagery, and climate data (e.g., ERA5, GPM). Following, covariates were extracted from these sources, the backscatter coefficients from Sentinel-1, the vegetation and water indices (NDVI, NDWI) from Sentinel-2, and environmental parameters from climatic datasets.

These covariates are then integrated and subjected to data analysis, including statistical methods such as linear correlation, to establish relationships between the derived indices and observed soil moisture. The final output of the research is the spatial and temporal characterization of soil moisture conditions across the study area. This integrated approach leverages the complementary strengths of radar, optical, and climate data for a comprehensive understanding of soil hydrological dynamics.

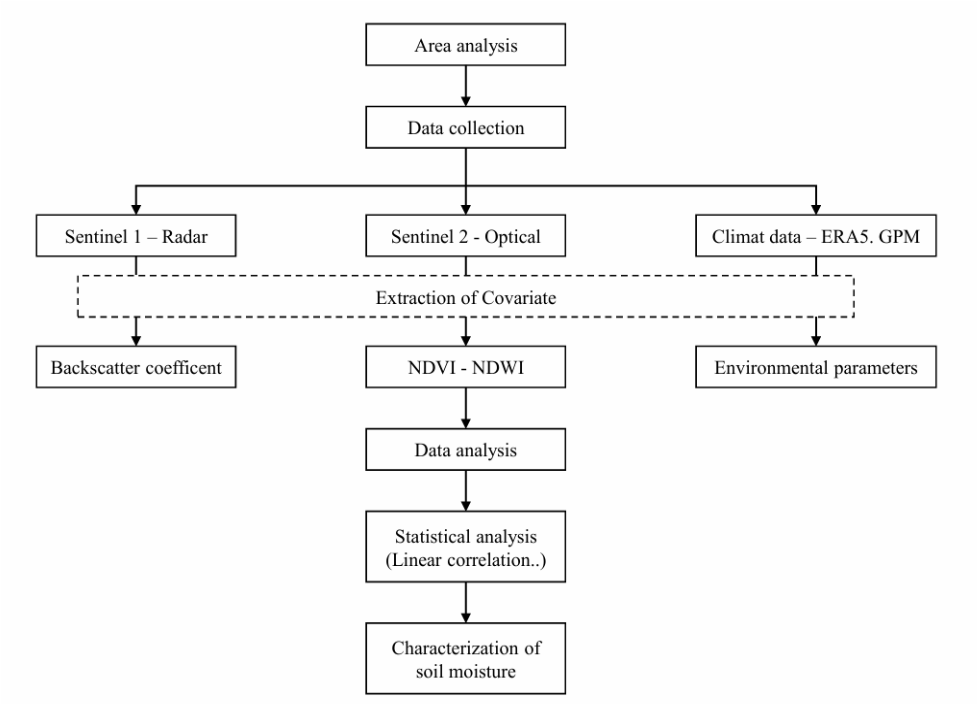


Figure 2. Methodological diagram of the soil moisture characterization

3.2.1. Data collection

The environmental covariates, Sentinel-1 (Backscatter coefficient), Sentinel-2 (NDVI and NDWI) and Digital terrain data were chosen as target explanatory variables to integrate. NDI (Normalized Difference Index) is calculated as the difference between reflectance values in two bands divided by the sum of those values (Dobos & Luca Montanarella, 2010)

3.2.2. Extraction of Covariates

A combination of literature and statistical processing was used to select certain factors. According to the literature, covariates had to follow three criteria: first, they had to represent soil-forming elements; second, they had to have a direct link with SM; and third, they had to be readily available (Dobos et al., 2000).

3.2.3. Characterization of Soil Moisture

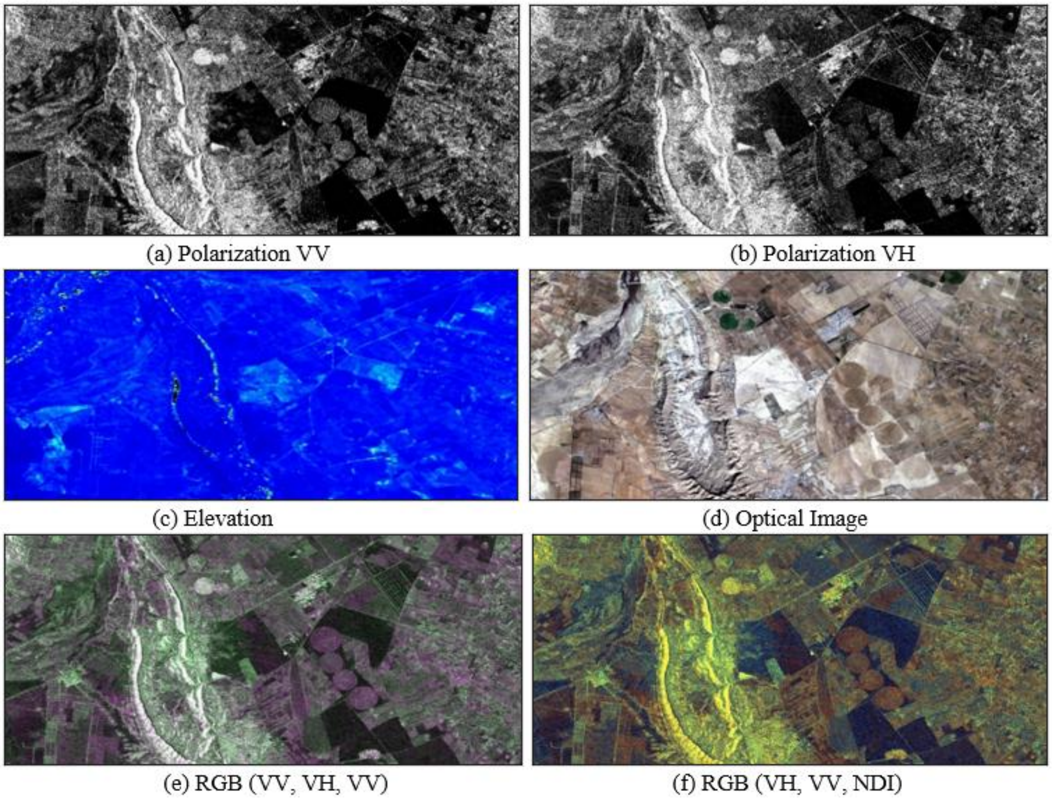
Many researchers have developed models to characterize soil moisture, and these models showed that soil moisture and backscattering coefficient have a fundamental relationship, e.g. (Zhang et al., 2021). The backscatter coefficient and Sentinel-2 data are considered main parameters that can help to characterize the soil moisture.

3.3. Results and Discussion

3.3.1. Radar remote sensing

An optical image and a Sentinel-1 radar image shown in Figure 3 to visually analyze the area. The radar image (polarizations) is colored with an RGB composite: red = VH, green = VV, and blue = NDI, where NDI refers to the Normalized Difference Index of the VH and VV polarizations. The dominance of bluish tones, greenish tones, and the presence of reddish tones in the southern part of the study area can be immediately distinguished.

First, a visualization exercise was performed using the RGB composite and the different polarizations (Figure 10) were then carried out as follows:


(a) Polarization VV

(b) Polarization VH

(c) Elevation

(d) Optical Image

(e) RGB (VV, VH, VV)

(f) RGB (VH, VV, NDI)

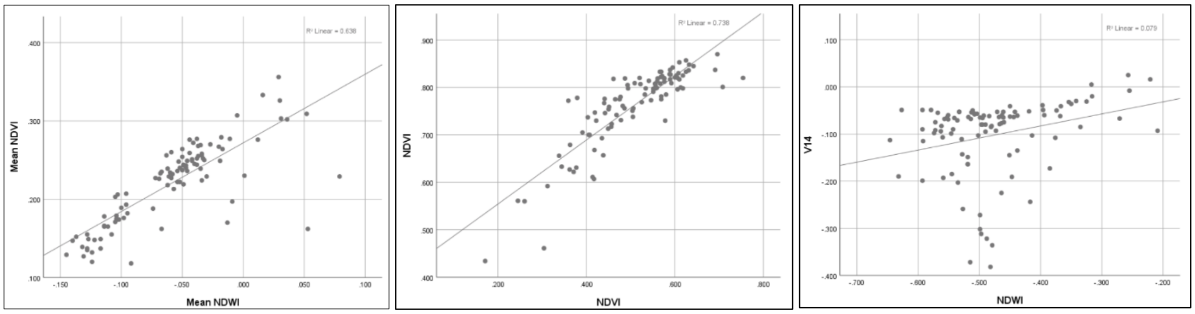
Figure 3. (a) Polarization VV; (b) Polarization VH; (c) Elevation; (d) Optical Image; (e) RGB (VV, VH, VV); (f) RGB (VH, VV, NDI)

Sentinel-1 C-band VV and VH polarizations provided complementary insights into soil moisture and land cover: VH enhanced water and soil wetness detection, while VV was more effective for vegetation and land use analysis. False-color composites distinguished cultivated soils, ploughed terrain, biomass, and bedrock structures, highlighting the sensitivity of SAR backscatter to moisture and surface conditions. However, the complexity of multiple scattering in vegetation and terrain underscores the need for ground-truth validation and advanced analysis to improve SAR-based soil moisture assessment.

3.3.2. Optical Remote Sensing

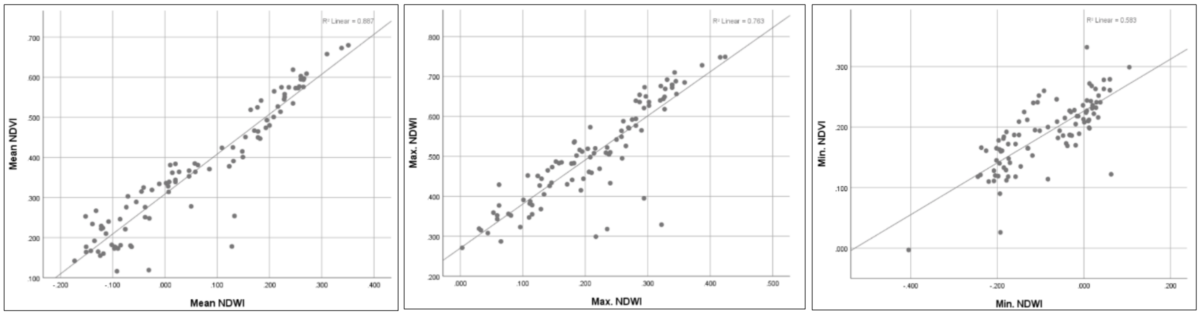
Correlation between the NDVI and the NDWI

The results show that maximum and mean NDWI–NDVI values exhibit strong correlations with soil moisture, while minimum values display weaker and less consistent relationships, likely influenced by free water, urban areas, or heterogeneous land covers. Validation using additional pilots (olive groves and horticultural crops) confirmed this pattern, with strong correlations for mean and maximum values but lower predictability for minimums. Outlier points, consistently observed across pilots, are interpreted as noise arising from technical errors or non-representative conditions.



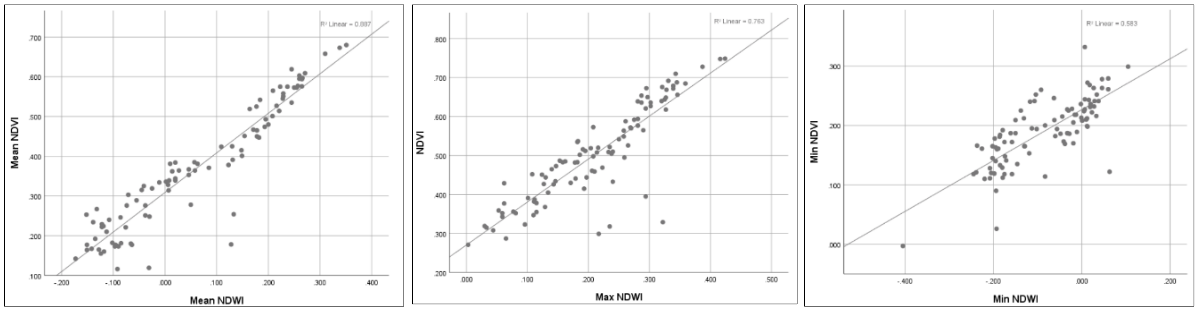
R = 0.79 R² = 0.64 R = 0.859 R² = 0.638 R = 0.28 R² = 0.079

Figure 4. Correlation between the NDWI and NDVI of the study area



R = 0.87 R² = 0.89 R = 0.94 R² = 0.76 R = 0.764 R² = 0.58

Figure 5. Correlation between the NDWI and NDVI of Pilot 3



R = 0.94 R² = 0.89 R = 0.87 R² = 0.76 R = 0.76 R² = 0.58

Figure 6. Correlation between the NDWI and NDVI of Pilot 4

3.3.3. Conclusion

This study demonstrated that integrating Sentinel-1B SAR data (VV, VH), NDVI, NDWI, DEM, and soil type information provides a robust framework for characterizing soil moisture regimes across heterogeneous land uses. Correlation analysis across multiple pilot sites revealed that vegetation indices, radar backscatter, and soil properties are strongly linked to soil wetness and vegetation conditions, confirming the comparative potential of optical and radar datasets. The results highlight that combining cognitive and statistical approaches with local knowledge is effective for understanding soil–water–vegetation interactions, and further statistical testing of environmental covariates is essential to identify the most influential drivers of soil moisture variability. Based on the conclusions above the following theses were formulated:

Thesis 1

The synergistic integration of Sentinel-1 C-band SAR backscatter (VV/VH), Sentinel-2-derived vegetation and water indices (NDVI, NDWI), and digital elevation data enables robust spatiotemporal characterization of soil moisture in the semi-arid Merguellil Basin. By capturing complementary signals related to vegetation structure, surface roughness, water content and water retention variability, this approach delineates soil wetness patterns across physiographic zones and land use types with minimal ground-based input. The method is particularly effective for characterizing topsoil moisture (0–5 cm), where SAR backscatter is most responsive, and its performance remains sensitive to vegetation density. However, limitations may arise under dense canopy cover or highly rough surfaces, where multiple scattering and signal saturation can reduce accuracy.

Thesis 2

Strong statistical correlations between radar backscatter coefficients, spectral vegetation and water indices, and topographic parameters demonstrate the capacity of multi-source remote sensing to infer soil moisture regimes under data-scarce conditions. The observed divergence in NDVI–NDWI dynamics across pilot sites further reveals the influence of soil texture, land use, and water retention variability, supporting the use of integrated optical-radar analysis for hydrological monitoring in heterogeneous agro-ecosystems. Nevertheless, performance declines under dense vegetation or highly rough surfaces, where signal penetration and retrieval accuracy are reduced.

4. Non-Invasive Soil Texture Prediction Using Random Forest and Multi-Source Environmental Data

The study area of this chapter is Hungary, a Central European country situated within the Carpathian Basin, spanning latitudes 45°48′ to 48°35′ N and longitudes 16°05′ to 22°58′ E.

4.1. Materials and Methods

4.1.1. Materials

Soil particle size distribution was assessed from 0–30 cm depth samples collected at 25 sites across Hungary, chosen to capture diverse geographic and geological conditions. Standard granulometric analysis determined the proportions of sand, silt, and clay, which were classified using the USDA soil texture system. The samples span multiple textural classes, including sandy loam, loam, clay loam, sandy clay loam, and clay, with most clustering in the loam to clay loam categories. This distribution highlights the heterogeneity of soil properties across the study area and provides a robust basis for texture-based classification and analysis.

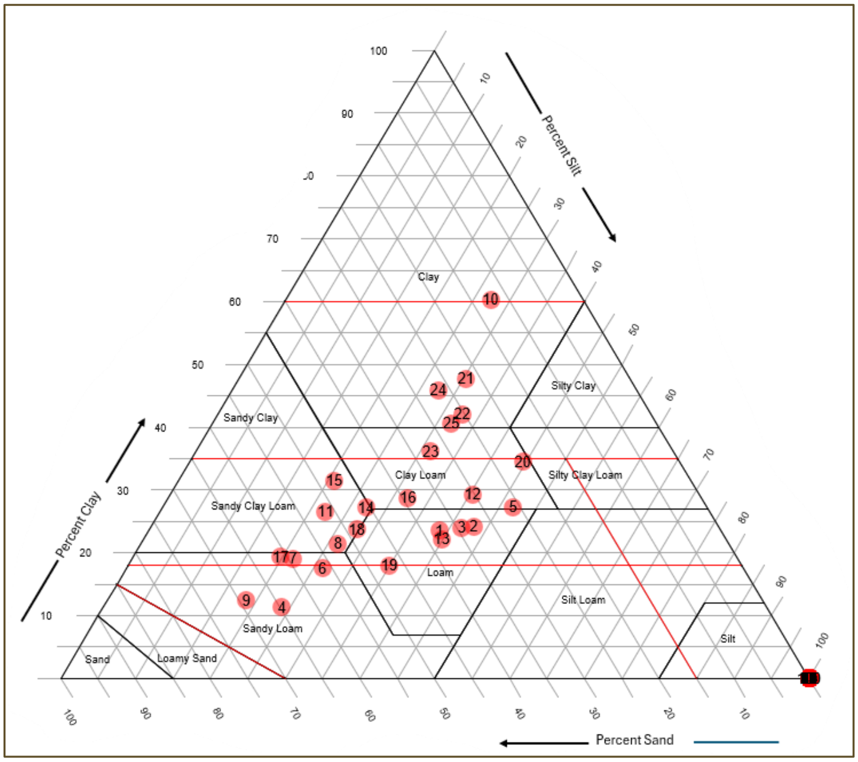


Figure 7. Soil Textiure Distribution of Soil Samples Plotted on the USDA Soil Texture Triangle

A total of twenty-five distinct field sites across Hungary, representing different soil textures, were selected for this study. At each site, one Sentek EnviroScan sensor access pipe was installed.
To ensure minimal soil disturbance and optimal sensor-soil contact, the installation process followed a rigorous methodology.

Data Description

Figure 8 presents the set of input variables used in the machine learning framework together with the target feature, soil texture.

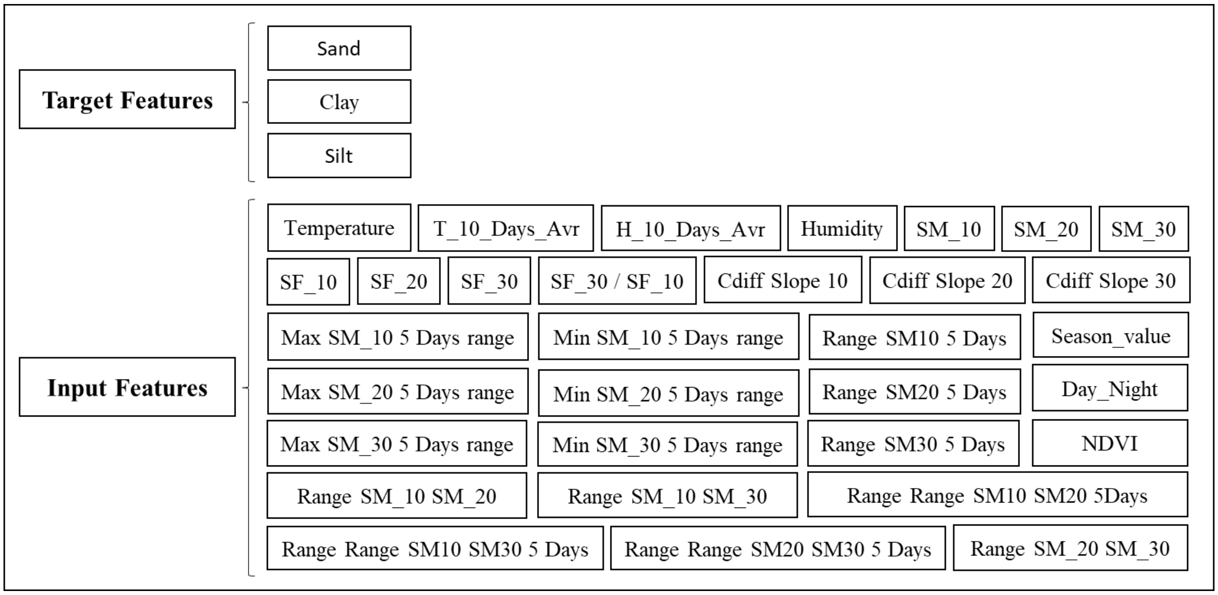


Figure 2. Input Features and Target Variable for Soil Texture Prediction

In this study, the target variables are the soil particle size classes. sand, silt, and clay, defined by USDA standards, which form the basis of soil texture classification and strongly influence water retention, aeration, and nutrient dynamics. Input features include both in-situ and remote sensing variables. Multi-depth soil temperature, humidity, and scaled frequency (SF) measurements were obtained using the Sentek EnviroSCAN probe, from which soil moisture at 10, 20, and 30 cm depths (SM_10, SM_20, SM_30) and their short-term fluctuations (5-day ranges, vertical gradients, and slopes) were derived. Additional temporal features included normalized seasonal progression (Season_Value) and diurnal cycles (Day_Night). Vegetation status was captured by the Normalized Difference Vegetation Index (NDVI) from Sentinel-2 imagery, serving as a proxy for soil–plant water interactions. Together, these variables integrate soil physical properties, hydroclimatic dynamics, and vegetation responses, providing a comprehensive dataset to assess the relationships between soil texture and soil moisture variability across depths and timescales.

Variable importance and Model Performance Evaluation

The feature importance of predictor variables was evaluated using the Random Forest Regressor (RFR) algorithm. The importance of each feature was determined based on the mean decrease in impurity (MDI), which is computed as the total reduction in variance attributed to each feature across all decision trees within the ensemble model. Specifically, the scikit-learn implementation of Random Forest calculates feature importance by assessing how much each predictor variable contributes to improving the accuracy of the model's predictions.

Since the Random Forest Regressor (RFR) is used for continuous variable prediction rather than classification, regression-specific evaluation metrics were employed to assess model performance. The key performance indicators used in this study include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2 Score). These metrics provide a quantitative measure of how well the predicted soil properties (Sand, Silt, and Clay fractions) align with actual observed values.

4.1.2. Methods

The flowchart (Figure 3) illustrates the methodological framework adopted in this study, organized into three main phases: data preparation, model construction, and data prediction. In the data preparation phase, field-based measurements, satellite observations, and ancillary environmental datasets were collected, pre-processed, and harmonized into a consistent spatial and temporal framework. Relevant features were extracted and compiled to create the input dataset. The model construction phase involved partitioning the dataset into training and testing subsets, followed by the implementation of the machine learning algorithms, Random Forest Regressor (RF). Hyperparameter tuning and cross-validation were applied to optimize model performance and minimize overfitting. In the data prediction phase, the trained model was applied to generate spatially continuous predictions of soil texture, which were then validated against independent in situ measurements.

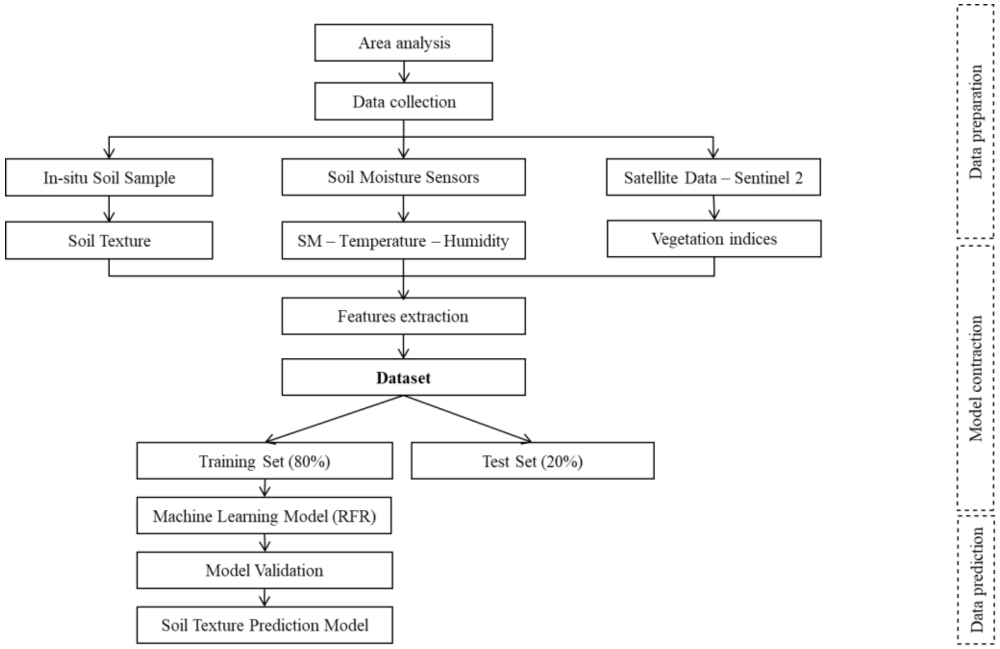


Figure 3. Soil texture prediction workflow

Predictive Modeling

The predictive modeling process using the Random Forest Regressor (RFR) for soil composition estimation follows a structured approach, summarized as follows:

Data Acquisition and Preprocessing: The dataset contains a range of environmental and geospatial predictor variables along with soil texture components (sand, silt, and clay fractions). Initial preprocessing steps include handling missing values, standardizing variables, and ensuring consistency across all observations.

Data Splitting and Model Training: The preprocessed dataset is partitioned into training and test subsets to assess model generalizability. The Random Forest algorithm is implemented, where multiple decision trees are trained using randomly selected samples from the training set. The predictions from these trees are aggregated, reducing variance and improving predictive accuracy.

Model Evaluation and Optimization: The trained model is evaluated using Mean Squared Error (MSE), R^2 score, and feature importance analysis to assess prediction reliability. Hyperparameter tuning, including adjustments to tree depth, the number of estimators, and feature selection, is performed to optimize model performance.

Feature importance analysis was conducted to identify the most influential predictors. The model selection module was used for data partitioning, cross-validation, and hyperparameter tuning, while the metrics module computed accuracy scores for performance evaluation. The ensemble module facilitated the implementation of the Random Forest Regressor, ensuring robust prediction capabilities.

Model Training

The Random Forest model was selected as the predictive framework for its robustness in capturing nonlinear relationships and managing high-dimensional datasets. It operates by constructing multiple decision trees, each trained on randomly sampled subsets of the data, enabling diverse perspectives on feature–target interactions. These trees individually learn the complex dependencies between input features and soil particle size fractions (sand, silt, clay). The ensemble then aggregates the predictions across all trees, averaging results to reduce overfitting, enhance stability, and improve the overall accuracy and generalization of soil property predictions.

4.2. Results & Discussion

Soil particle size class Prediction Using Random Forest Regressor: Model Performance and Feature Importance Analysis

The Random Forest model achieved strong predictive performance for soil particle size fractions, with R^2 values above 0.85 across sand, silt, and clay, confirming its reliability in modeling soil composition. Silt prediction was the most accurate ($R^2 = 0.93$, RMSE = 2.20), reflecting its more uniform properties and consistent model behavior. Sand also performed well ($R^2 = 0.92$, RMSE = 3.87) but showed slightly higher variability due to its coarse texture and rapid drainage. Clay prediction was comparatively weaker ($R^2 = 0.85$, RMSE = 3.44), highlighting the complexity of capturing its nonlinear retention and shrink–swell dynamics. Overall, these results emphasize the model's robustness while underscoring the need for soil texture-specific refinements, particularly for clay-rich soils.

Table 1. Model Performance Metrics for Soil particle size class Prediction

Target Variable	MSE	RMSE	R^2 Score
Sand	14.97	3.87	0.92

Silt	4.85	2.2	0.93
Clay	11.8	3.44	0.85

Feature Importance Relates to Soil particle size class (Sand, Silt, Clay)

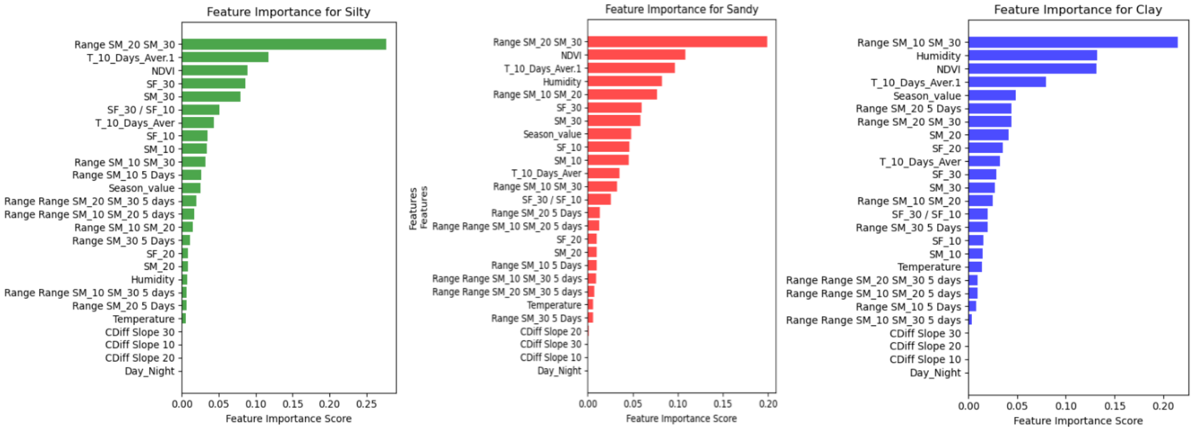


Figure 10. Feature Importance Ranking for Silt, Clay and Sand classes

The feature importance analysis reveals that soil moisture variability across depths is the dominant driver of soil particle size prediction, with Range SM_20–SM_30 being most critical for sand and silt, and Range SM_10–SM_30 for clay. These vertical gradients capture key differences in water transport and retention, reflecting the rapid drainage of sandy soils versus the slower, retentive behavior of clays. NDVI and humidity also emerged as important proxies, linking vegetation health and atmospheric demand to soil moisture dynamics, while temperature indicators (e.g., 10-day averages) highlight the role of evapotranspiration pressure. Together, these results demonstrate that soil texture governs the soil–water–plant continuum, where coarse soils amplify moisture fluctuations under temperature stress and fine soils buffer water availability, and that integrating depth-specific moisture ranges with climatic and vegetation features provides a physically consistent and statistically robust basis for predicting soil texture classes.

Evaluation of Model Predictions: True vs. Predicted Soil Composition

The scatter plots (Figure 11) confirm that the Random Forest model effectively captures soil composition trends, with silt predictions showing the highest accuracy and closest alignment to the 1:1 line, sand predictions performing moderately with some scatter at low values, and clay predictions exhibiting the greatest dispersion and uncertainty due to complex moisture retention properties. Overall, the model proves most reliable for silt, followed by sand, while clay remains the most challenging to predict.

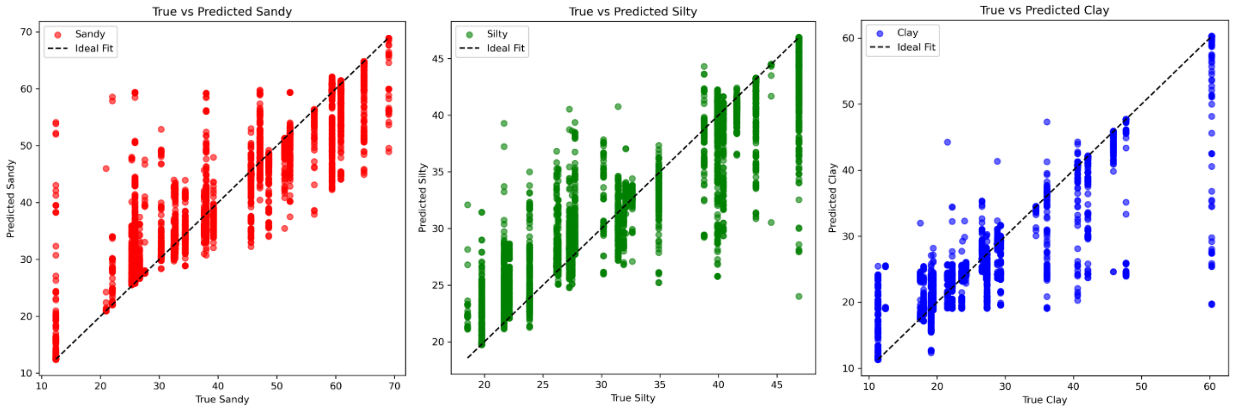


Figure 11. True vs Predicted Values of Sand, Silt, and Clay class

Hyperparameter Optimization Results

The optimized models were subsequently evaluated using Mean Squared Error (MSE) and the coefficient of determination (R2), key statistical metrics for assessing regression model performance. The results demonstrated high predictive accuracy, with R^2 values close to 1, indicating a strong correlation between predicted and observed values:

Table 2. Model Performance Metrics

Soil particle size class	MSE	RMSE	R^2 Score
Sand	5.13	2.27	0.97
Silt	1.01	1	0.99
Clay	4.88	2.21	0.94

4.3. Conclusions

This study presents a robust machine learning framework for predicting soil texture composition, by predicting the sand, silt, and clay fractions, by integrating in-situ soil moisture data with Sentinel-2 derived vegetation indices and environmental variables. Among the tested models, Random Forest Regression (RFR) demonstrated a good performance, achieving coefficient of determination (R^2) values exceeding 0.90 for all soil particle size classes and peaking at 0.99 for silt class, underscoring the model's predictive accuracy and generalization capability.

The feature importance analysis revealed that short-term soil moisture dynamics at multiple depths, in combination with Range of the SM, NDVI and humidity, are key determinants of soil texture variability. These results validate the hypothesis that soil water availability and vegetative response are reliable proxies for underlying textural properties, offering a novel perspective on indirect soil characterization.

Thesis 3

The integration of multi-depth in-situ soil moisture dynamics from capacitance-based Sentek EnviroSCAN sensors with Sentinel-2-derived vegetation indices and environmental parameters enables the indirect, non-invasive prediction of soil texture composition. By leveraging Random Forest regression, this framework captures the complex interactions between soil moisture behavior, vegetative response, and particle size distribution, providing a scalable alternative to traditional laboratory-based soil texture analysis. Furthermore, the potential for site-specific model transferability remains an important consideration, as variations in parent material, climatic conditions, and land use may influence model performance.

Thesis 4

Vertical gradients and short-term variability in soil moisture, quantified through sensor-derived scaled frequency metrics and multi-depth moisture fluctuation indices, exhibit strong predictive relationships with USDA-defined soil particle size classes. The soil moisture fluctuation indices were derived from Sentek EnviroSCAN probe data, including maximum, minimum, and range values over 5-day moving windows at 10, 20, and 30 cm depths, as well as inter-depth gradients (e.g., Range SM_10–30). When combined with NDVI and environmental features within a machine learning framework, these spatiotemporal dynamics allow for accurate, high-resolution mapping of soil texture, demonstrating the viability of sensor- and satellite-based approaches for digital soil characterization in heterogeneous environments.

5. Clay Ratio prediction using Machine learning, in situ and remote sensing data

5.1. Materials and Methods

For this chapter, the study area, datasets, and preprocessing workflow are identical to those described in the previous chapter. The key distinction lies in the prediction target: while the earlier chapter focused on soil texture classes, the present analysis specifically aims to predict the Clay Ratio. Furthermore, this chapter evaluates and compares the performance of three machine learning algorithms, Random Forest Regressor (RF), Extreme Gradient Boosting (XGBoost), and Gradient Boosting Regressor (GBR).

The workflow illustrated in Figure 12 outlines the sequential steps of the study, which are structured into three main phases: data preparation, model construction, and data prediction. In the data preparation phase, field measurements, remote sensing products, and ancillary datasets are integrated, processed, and harmonized. The model construction phase involves splitting the dataset into training and testing subsets, followed by the implementation of machine learning algorithms. Finally, in the data prediction phase, the trained models are applied to estimate the target variable, Clay Ratio, with subsequent accuracy assessment and validation against independent observations.

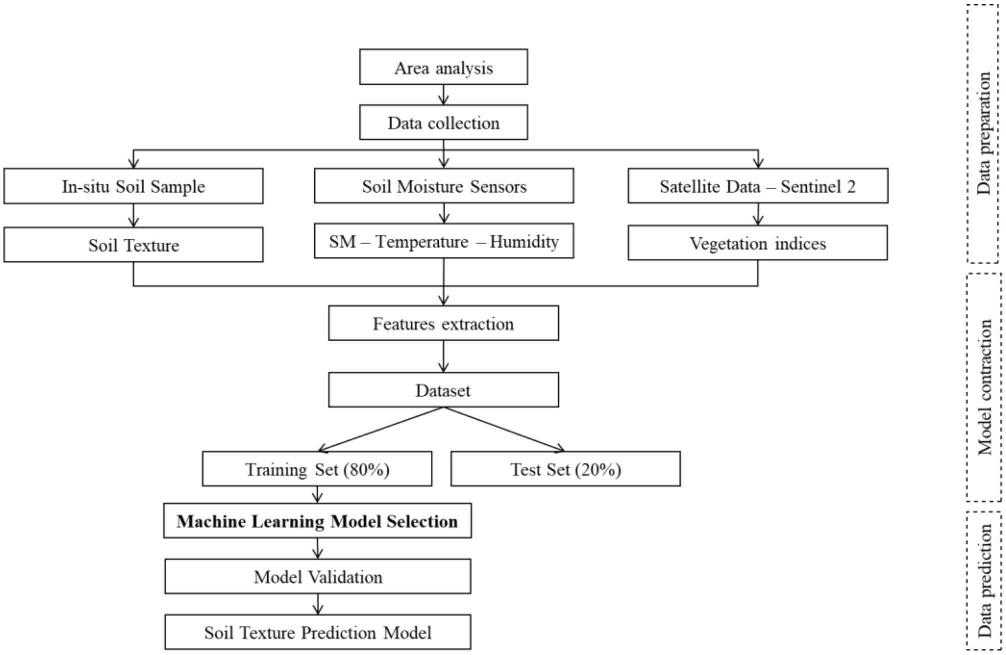


Figure 12. Clay Ration prediction workflow

The Clay Ratio was computed as the proportion of clay content to the combined content of sand and silt.

This derived variable serves as a normalized indicator to assess the relative abundance of fine particles within the soil matrix. All predictive features, comprising in situ soil moisture measurements at multiple depths, meteorological data, and remote sensing indices, were retained to ensure methodological consistency. The Random Forest regression model was again employed due to its robustness in handling non-linear relationships and heterogeneous data sources. This consistent approach facilitates comparative analysis across different soil texture attributes within the same modeling framework.

Data Description

Figure 13 presents the set of input variables used in the machine learning framework together with the target feature, Clay Ratio.

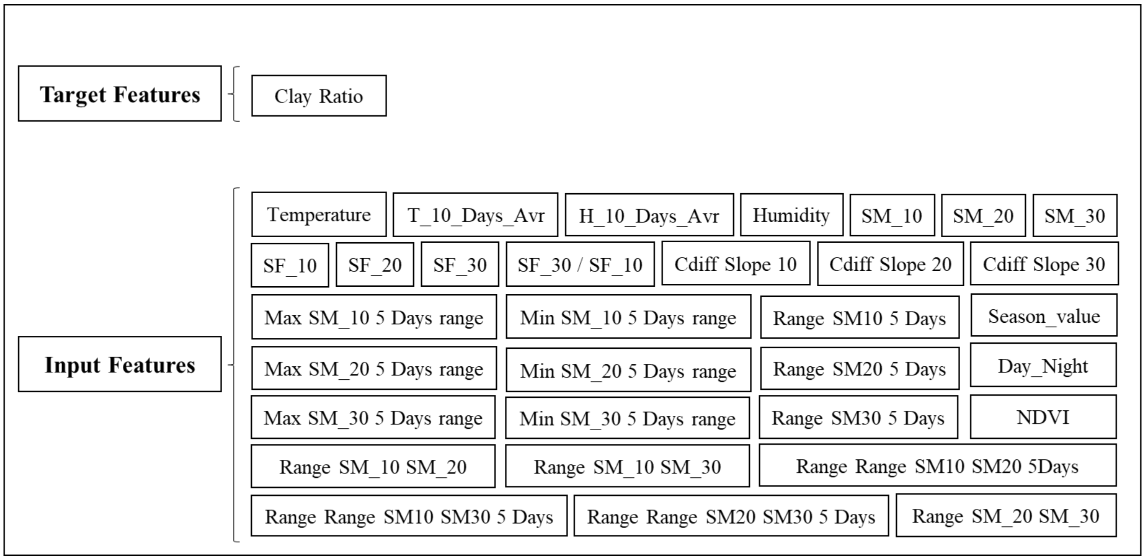


Figure 13. Input Features and Target Variable for the Clay Ratio Prediction

5.2. Results and Discussion

Model Performance Analysis

The comparative evaluation of the three ensemble models for Clay Ratio prediction, Random Forest, Gradient Boosting Regressor (GBR), and XGBoost Regressor (XGB), is presented in the table 3 below, based on three standard regression performance metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2 Score).

Table 3. Comparison of Model Performance Metrics for Clay Ratio Prediction

Target Variable	**MSE**	**RMSE**	**R^2 Score**
Random Forest	0.003	0.054	0.936
Gradient Boosting	0.0015	0.038	0.9747
XGBoost	**0.0015**	**0.038**	**0.9748**

The Random Forest model achieved an R^2 of 0.936, indicating that it explains 93.6% of the variance in clay ratio. However, compared to the other two boosting-based models, it yielded a higher error (MSE = 0.003, RMSE = 0.054), which suggests lower precision in capturing more complex relationships in the data.

In contrast, Gradient Boosting and XGBoost outperformed Random Forest on all metrics, both achieving an R^2 of approximately 0.975, which signifies that they explain over 97% of the variance in the clay ratio. The RMSE of 0.038 in both models reflects a substantial improvement in predictive accuracy over the Random Forest.

The slightly better R^2 score for XGBoost (0.9748) compared to Gradient Boosting (0.9747) highlights the marginal edge of XGBoost, which can be attributed to its regularization techniques and scalability. XGBoost also tends to be more robust to overfitting, particularly in high-dimensional datasets, due to its built-in shrinkage and column subsampling strategies.

These results demonstrate that boosting models, particularly XGBoost, are more effective than bagging methods like Random Forest in predicting soil clay content from integrated datasets. The enhanced performance is likely due to the ability of boosting algorithms to sequentially correct prediction errors and focus on difficult-to-predict observations.

Feature Importance Relates to Clay Ratio Prediction

The Random Forest model highlights Range SM_10 SM_30 as the most influential feature (Figure 14), indicating that vertical moisture gradients between shallow and deeper soil layers are critical for predicting clay content. Vegetation index (NDVI), Humidity, and 10-day averaged temperature metrics (T_10_Days_Aver, T_10_Days_Aver.1) also show high importance, reflecting their role in controlling soil moisture dynamics and vegetation stress responses associated with clay-rich soils. Features related to frequency-scaled moisture (e.g., SF_30 / SF_10) and seasonal effects (Season_value) follow closely.

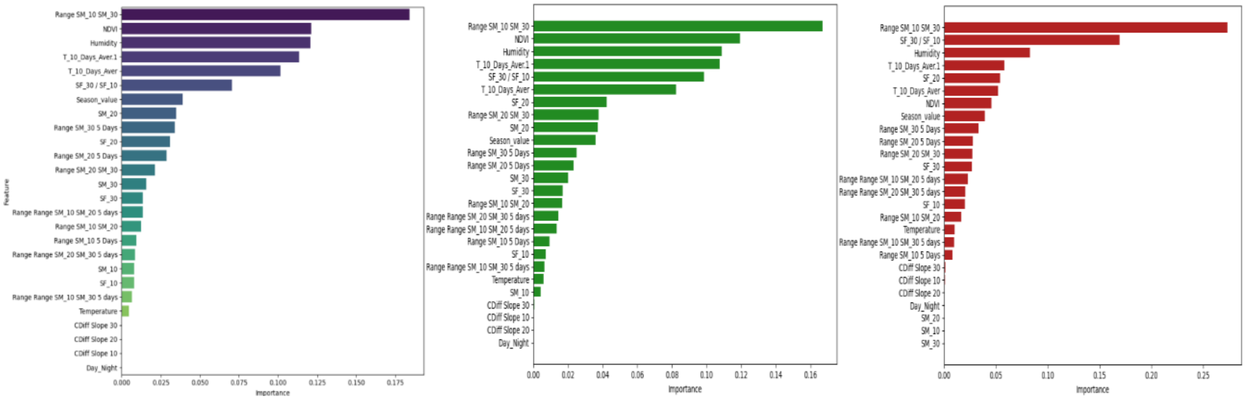


Figure 14. Feature Importance for Predicting Clay Ratio - Random Forest Regressor

Gradient Boosting confirms the dominance of Range SM_10 SM_30 as the top feature, followed by NDVI, Humidity, and T_10_Days_Aver.1, consistent with Random Forest. However, this model attributes relatively higher importance to frequency-scaled features such as SF_30 / SF_10 and SF_20, as well as moisture ranges at various depths. The distribution of importance is more gradual, showing the model's sensitivity to a broader set of hydrometeorological variables. Low influence remains with slope change metrics (CDiff Slope 10/20/30) and Day_Night.

XGBoost places the strongest emphasis on Range SM_10 SM_30, assigning it the highest individual score among all models, followed by SF_30 / SF_10. This underlines the significance of vertical moisture distribution and signal frequency as key indicators of clay content. While Humidity, T_10_Days_Aver.1, and NDVI remain important, the sharp drop in importance among mid-ranked features suggests that the model relies more heavily on a few dominant variables.

Model selection

The evaluation of model performance demonstrates that XGBoost is the most accurate and robust algorithm for predicting the Clay Ratio, outperforming both Gradient Boosting Regressor (GBR) and Random Forest (RF). While GBR and XGBoost achieved very similar results (R² ≈ 0.975, RMSE = 0.038), both were substantially better than RF (R^2 = 0.936, RMSE = 0.054), confirming the superiority of gradient-based ensemble methods in capturing non-linear soil–environment interactions. Scatter plots further validate this, showing XGBoost predictions closely aligned with the 1:1 line and exhibiting minimal bias, while RF displayed larger deviations, particularly at extremes. Overall, XGBoost emerges as the most reliable model for Clay Ratio prediction, offering high precision, strong generalization, and suitability for soil texture modeling in heterogeneous agro-environmental contexts.

Interpretation

The XGBoost feature importance analysis for clay fraction prediction underscores the dominance of vertical soil moisture gradients as key indicators of texture. The most influential variable, Range SM_10–SM_30 (0.2731), effectively captures the slower redistribution and higher retention of water in clay-rich soils compared to the rapid infiltration of sandy profiles. Similarly, the SF_30/SF_10 ratio (0.1694) highlights subsurface stratification and compaction effects that influence water movement across horizons. Secondary predictors such as humidity, temperature, and NDVI reflect evapotranspiration demand and vegetation status, providing indirect cues about soil–water interactions shaped by texture. The importance of range-based temporal and depth-specific moisture indicators further demonstrates the value of dynamic features over static absolute values, as they encapsulate wetting–drying cycles strongly governed by soil texture. In contrast, static soil moisture values and topographic or diurnal proxies contributed little to predictive power. Overall, the model highlights how integrating moisture dynamics with climatic and vegetation signals provides a robust framework for inferring soil texture, with clay prediction strongly linked to depth-dependent water redistribution and plant–soil–atmosphere interactions.

5.3. Conclusion

In conclusion, the prediction of the clay fraction using machine learning models demonstrates the strong potential of data-driven approaches for soil texture characterization. Among the tested algorithms, XGBoost achieved the highest predictive performance, with an R^2 score of 0.9748 and the lowest error metrics, confirming its robustness in capturing complex, non-linear relationships between environmental predictors and clay content. The analysis of feature importance revealed that dynamic indicators of soil moisture variability, particularly the vertical moisture gradient between 10 cm and 30 cm depth, were the most informative variables. These findings emphasize the relevance of moisture-related features over static measurements, reflecting the intrinsic link between water retention behavior and soil texture. Additionally, variables representing climatic demand and vegetative status, such as humidity, temperature, and NDVI, further enhanced model accuracy by capturing the soil–plant–atmosphere interactions.

Thesis 5

The vertical gradient of soil moisture, quantified as the differential between shallow and subsurface layers (e.g., Range SM_10–SM_30), encapsulates the moisture behavior governed by fine particle distribution and thus emerges as a key predictor of clay ratio. When integrated with spectral vegetation proxies (NDVI) and environmental variables, this dynamic moisture signal enables physically informed, non-invasive estimation of clay-enriched soils through ensemble learning frameworks. However, in coarse-textured soils, where infiltration is rapid and vertical moisture gradients are less distinct, the predictive strength of this indicator may be reduced, highlighting the need for context-specific calibration and validation.

References

Cato the Elder, & V. M. T. (1934). On agriculture (W. D. Hooper & H. B. Ash, Trans.). *Loeb Classical Library 283. Harvard University Press.*

Columella, L. J. M. (1941). On Agriculture. *Loeb Classical Library 361. Harvard University Press.*, *Volume I.*

Daniel J. Hillel. (1991). Out of the Earth: Civilization and the Life of the Soil. 1991. The Free Press Division of Macmillan, Inc., New York, New York. x + 321 pp. $22.50, cloth. *American Journal of Alternative Agriculture*, *6*(3), 114–114. https://doi.org/DOI: 10.1017/S0889189300004021

Dobos, E., & Luca Montanarella. (2010). *A quantitative procedure for building physiographic units supporting a global SOTER database Luca Montanarella European Commission*. https://www.researchgate.net/publication/288382467

Dobos, E., Micheli, E., Baumgardner, M. F., Biehl, L., & Helt, T. (2000). Use of combined digital elevation model and satellite radiometric data for regional soil mapping. *Geoderma*, *97*(3), 367–391. https://doi.org/https://doi.org/10.1016/S0016-7061(00)00046-X

Grayson, R. B., & Western, A. W. (1998). Towards areal estimation of soil water content from point measurements: time and space stability of mean response. *Journal of Hydrology*, *207*(1), 68–82. https://doi.org/https://doi.org/10.1016/S0022-1694(98)00096-1

Kerr, Y. H., Waldteufel, P., Wigneron, J.-P., Delwart, S., Cabot, F., Boutin, J., Escorihuela, M.-J., Font, J., Reul, N., Gruhier, C., Juglea, S. E., Drinkwater, M. R., Hahne, A., Martín-Neira, M., & Mecklenburg, S. (2010). The SMOS Mission: New Tool for Monitoring Key Elements ofthe Global Water Cycle. *Proceedings of the IEEE*, *98*(5), 666–687. https://doi.org/10.1109/JPROC.2010.2043032

Lacombe, G. (2007). *Evolution and uses of water resources in a semi-arid managed watershed. The case of Merguellil in Central Tunisia.* Doctoral dissertation at the Montpellier University, Science and Techniques of Languedoc.

Minasny, B., & McBratney, Alex. B. (2016). Digital soil mapping: A brief history and some lessons. *Geoderma*, *264*, 301–311. https://doi.org/https://doi.org/10.1016/j.geoderma.2015.07.017

Mohanty, B. P., Cosh, M. H., Lakshmi, V., & Montzka, C. (2017). Soil Moisture Remote Sensing: State-of-the-Science. *Vadose Zone Journal*, *16*(1), 1–9. https://doi.org/10.2136/vzj2016.10.0105

Njoku, E. G., & Entekhabi, D. (1996). Passive microwave remote sensing of soil moisture. *Journal of Hydrology*, *184*(1), 101–129. https://doi.org/https://doi.org/10.1016/0022-1694(95)02970-2

Ochsner, T. E., Cosh, M. H., Cuenca, R. H., Dorigo, W. A., Draper, C. S., Hagimoto, Y., Kerr, Y. H., Larson, K. M., Njoku, E. G., Small, E. E., & Zreda, M. (2013). State of the Art in Large-Scale Soil Moisture Monitoring. *Soil Science Society of America Journal*, *77*(6), 1888–1919. https://doi.org/https://doi.org/10.2136/sssaj2013.03.0093

Robinson, D. A., Campbell, C. S., Hopmans, J. W., Hornbuckle, B. K., Jones, S. B., Knight, R., Ogden, F., Selker, J., & Wendroth, O. (2008). Soil Moisture Measurement for Ecological and Hydrological Watershed-Scale Observatories: A Review. *Vadose Zone Journal*, *7*(1), 358–389. https://doi.org/10.2136/vzj2007.0143

Topp, G. C., Davis, J. L., & Annan, A. P. (1980). Electromagnetic determination of soil water content: Measurements in coaxial transmission lines. *Water Resources Research*, *16*(3), 574–582. https://doi.org/https://doi.org/10.1029/WR016i003p00574

Vereecken, H., Schnepf, A., Hopmans, J. W., Javaux, M., Or, D., Roose, T., Vanderborght, J., Young, M. H., Amelung, W., Aitkenhead, M., Allison, S. D., Assouline, S., Baveye, P., Berli, M., Brüggemann, N., Finke, P., Flury, M., Gaiser, T., Govers, G., … Young, I. M. (2016). Modeling Soil Processes: Review, Key Challenges, and New Perspectives. *Vadose Zone Journal*, *15*(5), vzj2015.09.0131. https://doi.org/https://doi.org/10.2136/vzj2015.09.0131

White, K. D. (1970). Roman farming. *Ithaca, N.Y., Cornell University Press*.

Zhang, Q., Liu, X., Liu, M., Zou, X., Zhu, L., & Ruan, X. (2021). Comparative analysis of edge information and polarization on sar-to-optical translation based on conditional generative adversarial networks. *Remote Sensing*, *13*(1), 1–20. https://doi.org/10.3390/rs13010128