A single Hyperspectral image for predicting Soil Organic Matter in saline semi-arid lands: insights from Random Forests and optimal band selection models

Soufiane Hajaj[1,2,*], Ahmed Laamrani[1,2], Jamal-Eddine Ouzemou[1,2], Nadir el Bouanani[1,2], Ali El Battay[1,2], and Abdelghani Chehbouni[1,2]

[1] Center for Remote Sensing Applications (CRSA), Mohammed VI Polytechnic University (UM6P), Ben Guerir 43150, Morocco
[2] College of Agriculture and Environmental Sciences (CAES), UM6P, Ben Guerir, Morocco

Keywords: Hyperspectral, Soil organic matter (SOM), CARS, RFE, Random Forest (RF)

Abstract

Accurate estimation of soil organic matter (SOM) in saline semi-arid environments is essential for assessing soil fertility and supporting sustainable land management. This study investigates the potential of a single EnMAP hyperspectral image combined with machine learning, spectral pre-processing, and feature selection techniques for SOM prediction. Random Forest (RF) regression was employed alongside three feature selection methods: RF feature importance (RFFS), Competitive Adaptive Reweighted Sampling (CARS), and Recursive Feature Elimination (RFE). Several spectral pre-processing techniques, including first derivative (FD), second derivative, Savitzky–Golay (SG), standard normal variate (SNV), and absorbance transformation, were evaluated. Results indicate that models based on original reflectance achieved only moderate performance (PCCC = 0.547–0.617; R^2 = 0.295–0.354). In contrast, spectral transformations, particularly FD, significantly improved predictive accuracy. The best-performing model (FD-RFFS) achieved PCCC = 0.697 ($p < 0.001$), R^2 = 0.446, RMSE = 0.825, and MAE = 0.643, representing notable improvements over raw spectra. Feature selection further enhanced model robustness, with RFFS outperforming CARS and RFE. Analysis of selected wavelengths revealed consistent importance of spectral regions within the visible–NIR (418–801 nm) and SWIR-2 (2207–2445 nm) ranges, highlighting their sensitivity to SOM variability. These findings demonstrate that combining derivative-based pre-processing with RF-driven feature selection improves SOM estimation from hyperspectral data. The proposed framework provides an effective and operational approach for SOM mapping in saline semi-arid landscapes.

1. Introduction

Accurately determining soil organic matter (SOM) content in saline semi-arid environments is essential for evaluating soil fertility and supporting sustainable land management. SOM plays a fundamental role in regulating soil structure, nutrient availability, and water retention, making it a key indicator of soil health and agricultural productivity. Although conventional laboratory analysis remains the benchmark for precise SOM quantification, it is often costly, labor-intensive, and limited in spatial representativeness, particularly in large and heterogeneous landscapes.

In this context, remote sensing techniques have emerged as efficient alternatives for large-scale soil monitoring. Hyperspectral imaging (HSI), particularly with advanced spaceborne sensors such as EnMAP, offers unprecedented opportunities by providing high-dimensional spectral information capable of capturing subtle biochemical and physicochemical variations in soils (Bourriz et al., 2025). However, extracting reliable SOM estimates from hyperspectral reflectance remains challenging in saline semi-arid regions, where spectral responses are strongly influenced by complex interactions among soil moisture, salt accumulation, surface crusting, and mineralogical composition (Al-Ali et al., 2021).

Reliable SOM prediction from HSI data typically relies on robust multivariate approaches, with Random Forest (RF) regression being widely adopted due to its ability to model non-linear relationships and handle high-dimensional datasets (Dai et al., 2026; Ghasemi & Tavakoli, 2013; He et al., 2023). Nevertheless, the predictive performance of RF is highly dependent on the quality of the input spectral data. Raw reflectance (OR) spectra are often affected by baseline shifts, scattering effects, and redundant information, which can obscure diagnostically relevant absorption features associated with organic matter. Consequently, the integration of advanced spectral pre-processing and feature selection techniques is essential to enhance the signal-to-noise ratio and improve model reliability (Zhang et al., 2024).

Spectral pre-processing methods such as the first derivative (FD), second derivative, standard normal variate (SNV), and Savitzky–Golay (SG) filtering are commonly applied to reduce noise, correct baseline effects, and resolve overlapping spectral features. Among these, derivative-based transformations—particularly the first derivative—have been shown to enhance subtle absorption features linked to organic functional groups, thereby improving correlations with SOM content. Complementary to pre-processing, feature selection techniques including Competitive Adaptive Reweighted Sampling (CARS), Recursive Feature Elimination (RFE), and Random Forest Feature Selection (RFFS) enable the reduction of spectral dimensionality by isolating the most informative wavelengths, leading to more parsimonious and robust predictive models.

Despite these advances, the transferability and operational implementation of hyperspectral-based SOM prediction frameworks remain limited in saline semi-arid environments. Strong spatial heterogeneity, coupled with dynamic surface conditions, often introduces non-stationarity in spectral responses, reducing model generalization beyond calibration datasets. Therefore, identifying stable and physiochemically meaningful spectral regions that consistently contribute to SOM prediction is critical for improving model robustness and scalability. Moreover, the increasing availability of EnMAP data necessitates the development of reproducible workflows that can effectively bridge experimental research and operational applications.

In this study, we evaluate the combined effects of multiple spectral pre-processing techniques and feature selection algorithms on SOM prediction using EnMAP hyperspectral imagery in saline semi-arid soils. Particular emphasis is placed on identifying spectral regions that are consistently selected across different methodological configurations, notably within the visible–NIR (418–801 nm) and SWIR-2 (2207–2445 nm) domains. By systematically assessing these approaches, this work aims to establish a reliable and operational framework for SOM estimation. The results demonstrate that integrating derivative-based pre-processing with RF-based feature importance effectively isolates key spectral features and significantly enhances predictive performance in complex soil environments.

2. Materials

2.1. Case study

The study took place in the Sehb El Mesjoune region, situated roughly 60 km north of Marrakech in the Bahira Plain of central Morocco (Fig. 1). This semi-arid to arid basin features a topographic low dominated by a central dry lake (sebkha) that briefly turns into a seasonal wetland after sporadic rainfall. Rapid water evaporation leaves behind salt crusts, accelerating soil salinization.

The surrounding landscape includes the Gantour Plateau to the north and the Jbilets Mountains to the south, both shaping local groundwater flow. The area is part of a closed hydrological basin, where saline groundwater accumulates toward the sebkha, causing high salinity in surface soils and shallow aquifers. Climatically, the region receives low mean annual rainfall (<200 mm) and experiences high potential evapotranspiration (~2,700 mm/year) (Ouzemou et al., 2025). These conditions, combined with limited drainage and the use of saline irrigation water, worsen soil salinity and contribute to agricultural decline.

Soils in the area are predominantly fine to medium in texture, with electrical conductivity values ranging from 0.5 to 235 dS/m, placing them in the saline to extremely saline categories. Halophytic plant species, including *Anabasis articulata* and *Salsola soda*, are characteristic of the landscape and serve as indicators of saline environments.

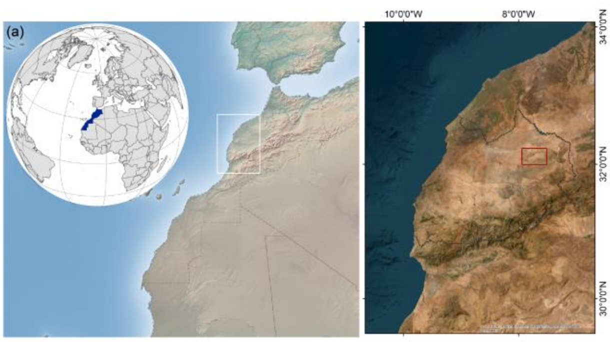

Figure 1. Study area

3. Methods

To mitigate the impact of sensor noise and light scattering in the EnMAP hyperspectral imagery, a multi-stage preprocessing pipeline was implemented. Standard Normal Variate (SNV) was applied to correct for multiplicative scatter effects caused by variations in soil surface texture. To address high-frequency noise while preserving absorption feature geometry, Savitzky–Golay (SG) smoothing was employed. Furthermore, first and second derivatives were computed using the SG derivative algorithm to resolve overlapping spectral features and eliminate baseline offsets. Raw reflectance data were also converted via absorbance transformation (log(1/R)) to linearize the relationship between spectral response and Soil Organic Matter (SOM) concentration. These methods are widely employed in spectral data pre-processing, and further methodological details can be found in (Bian, 2022). The transformed hyperspectral imaging (HSI) mean spectra obtained after preprocessing are illustrated in Figure 2.

3.1. Feature Selection

To address the "curse of dimensionality" inherent in hyperspectral datasets, three supervised feature selection strategies were evaluated to isolate the most informative bands for SOM estimation: Random Forest-Based Feature Selection (RFFS): Utilized Variable Importance (VI) scores to rank spectral bands based on their contribution to the reduction in Mean Squared Error (MSE) "Embedded method" (Breiman, 2001); Recursive Feature Elimination (RFE): A widely used FS algorithm in soil attributes prediction (Dai et al., 2025), is an iterative backward elimination process that ranks variables by importance and systematically removes the least contributory bands until an optimal subset is reached; Competitive Adaptive Reweighted Sampling (CARS): Employed a Darwinian "survival of the fittest" strategy, using Monte Carlo sampling and exponentially decreasing functions to select wavelength points with the highest absolute regression coefficients (Li et al., 2009).

3.2. Predictive Modeling and Optimization

Random Forest (RF) regression was selected as the primary estimator due to its robustness against non-linear spectral responses. To ensure model generalizability, hyperparameters, including the number of trees (ntree) and the number of variables per split (mtry), were optimized by a Grid Search strategy. Model stability was assessed through an independent test set, providing a systematic assessment of the spectral regions most critical for high-accuracy SOM mapping.

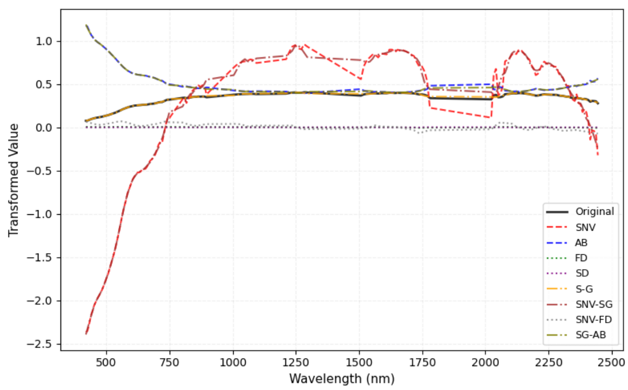

Figure 2. HSI transformed spectra

3.3. Accuracy Assessment

Model performance was evaluated using several statistical metrics in order to quantify the agreement between observed and predicted SOM values. The prediction correlation coefficient (PCCC) and the coefficient of determination (R^2) were used to assess the strength of the linear relationship and the proportion of variance explained by the model. Prediction errors were quantified through the root mean square error (RMSE) and the mean absolute error (MAE), which respectively measure the magnitude and average deviation of residuals. In addition, the residual prediction deviation (RPD) was computed to evaluate model robustness relative to the variability of the reference data, while model bias was examined to detect systematic over- or under-estimation. All metrics were calculated using the train and independent test dataset to ensure an unbiased evaluation of predictive performance.

4. Results

4.1. Assessing OR bands

Evaluation of the original reflectance (OR) spectra showed that predictive performance varied across feature selection methods, with a ranking of OR-RFE > OR-RFFS > OR-CARS. The highest accuracy was obtained with RFE (PCCC = 0.617, R^2 = 0.354), while CARS gave the lowest (PCCC = 0.547). Overall, the OR configurations yielded only moderate predictive skill: PCCC values ranged from 0.547 to 0.617, R^2 from 0.295 to 0.354, and RMSE remained close to 0.90 across models. These results indicate that raw EnMAP spectra capture soil organic matter (SOM) variability only weakly in this saline, semi-arid context.

4.2. The pre-processed data benefits

Applying spectral transformations, particularly the first derivative (FD), led to a notable improvement in model performance and altered the ranking of feature selection methods. The order shifted to FD-RFFS > FD-RFE > SG-FD-CARS, with FD-RFFS achieving the best results (PCCC = 0.697, R^2 = 0.446, RMSE = 0.825). The top-performing FD-based models (FD-RFFS and FD-RFE) raised PCCC to 0.673–0.697 and R^2 to 0.420–0.446, while reducing RMSE to 0.825–0.845. Figure 3 presents the prediction results for the training and testing datasets, highlighting the performance obtained using the original and first-derivative (FD) data. Compared with the original reflectance, this corresponds to an improvement of approximately 10–15% in correlation, 20–30% in explained variance, and a 5–8% reduction in prediction error (Table 1).

Table 1. Comparison of RF-based SOM prediction using original EnMAP bands and selected transformed bands.

Method	PCCC	R_sq	RMSE	MAE	RPD	Bias
OR-RFE	0.617***	0.354	0.891	0.704	1.24	-0.049
OR-RFFS	0.610***	0.342	0.900	0.697	1.23	-0.074
OR-CARS	0.547***	0.295	0.931	0.743	1.19	-0.052
FD-RFFS	0.697***	0.446	0.825	0.643	1.34	-0.063
FD-RFE	0.673***	0.420	0.845	0.665	1.31	-0.082
SG-FD-CARS	0.652***	0.360	0.887	0.682	1.25	-0.143

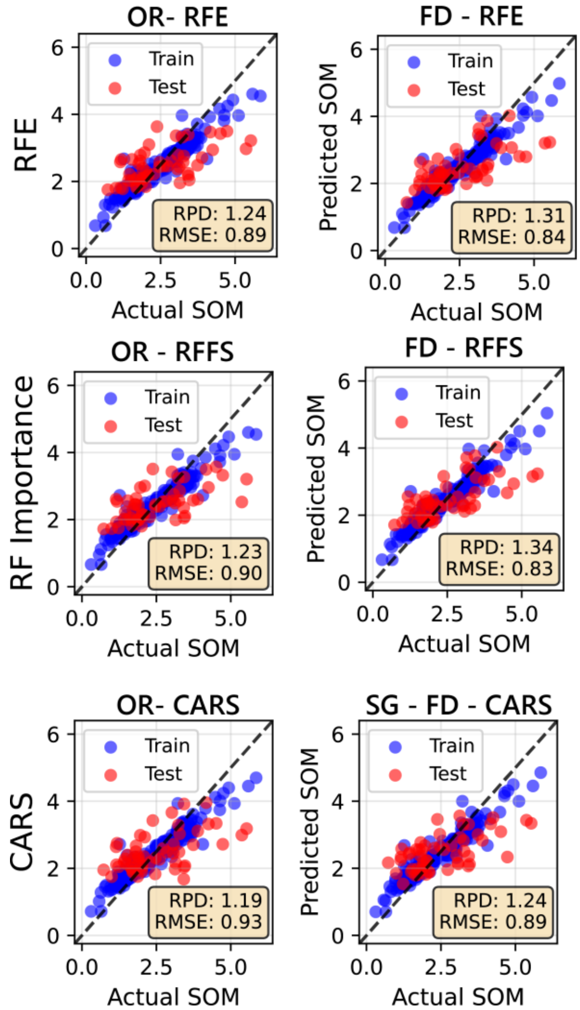

Figure 3. Prediction results for train and test datasets, highlighting the results from Original data and FD data.

4.3. Mapping results

The spatial distribution of soil organic matter (SOM) across the study area shows clear variability, consistent with the observed range of 0.98–4.52%, a mean of around 2%. The SOM map reveals distinct spatial gradients related to land cover and geomorphological conditions (Fig. 4). Higher SOM values, reaching nearly 4%, are mainly observed in the northern sector, particularly in agricultural areas where continuous cultivation and organic inputs enhance soil carbon accumulation. In contrast, the southern part of the study area, extending toward the Jbilets rocky terrain, is characterized by lower SOM levels, likely due to sparse vegetation, shallow soils, and the dominance of rocky substrates that limit organic matter accumulation. The central zone, corresponding to the dry lakebed, presents intermediate SOM values, reflecting transitional soil conditions influenced by seasonal moisture variations. No clear negative relationship between soil salinity and SOM distribution is observed according the results from (Achemrk et al., 2026), suggesting that salinity levels are either low or spatially heterogeneous. Further research is needed to better assess the potential influence of salinity on SOM dynamics.

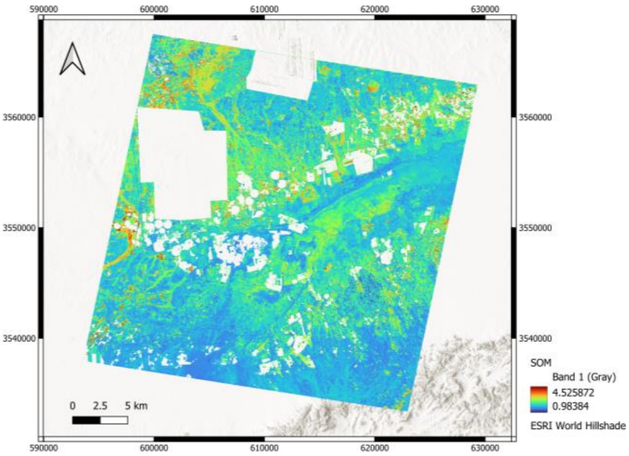

Figure 4. Map of prediction SOM (%) in the study area.

4.4. Ablation

Among the pre-processing techniques evaluated, the first derivative (FD) consistently outperformed SNV, second derivative, Savitzky–Golay, and absorbance transformations. FD yielded the highest prediction correlation coefficient (PCCC = 0.697, *p* < 0.001), an improved R^2 of 0.446, and lower error metrics (RMSE = 0.825, MAE = 0.643). Regarding feature selection, random forest-based feature selection (RFFS) proved superior to methods such as CARS and conventional RFE, leading to greater model accuracy and robustness. Figure 5 presents the test-based RMSE, R^2, and PCCC metrics for SOM prediction.

Compared with the original reflectance combined with RFFS (OR-RFFS), the FD-RFFS approach increased PCCC from 0.610 to 0.697 and reduced RMSE from 0.900 to 0.825. The rise in residual prediction deviation (RPD) from 1.23 to 1.34 further underscores the enhanced predictive capability achieved by coupling FD pre-processing with RFFS. Overall, applying the first derivative together with RF importance and RFE substantially improved prediction accuracy, and the FD-RFFS combination delivered stronger performance than other pre-processing and feature selection pairings.

A comprehensive sensitivity analysis across all tested combinations identified the SWIR-2 range (2207–2445 nm, 7 bands) and the visible–NIR range (418–801 nm, 13 bands) as the most influential spectral regions for predicting soil organic matter (SOM). The section Bands sensitivity (4.5) highlights more details regarding the most frequently selected bands using all the feature selection methods.

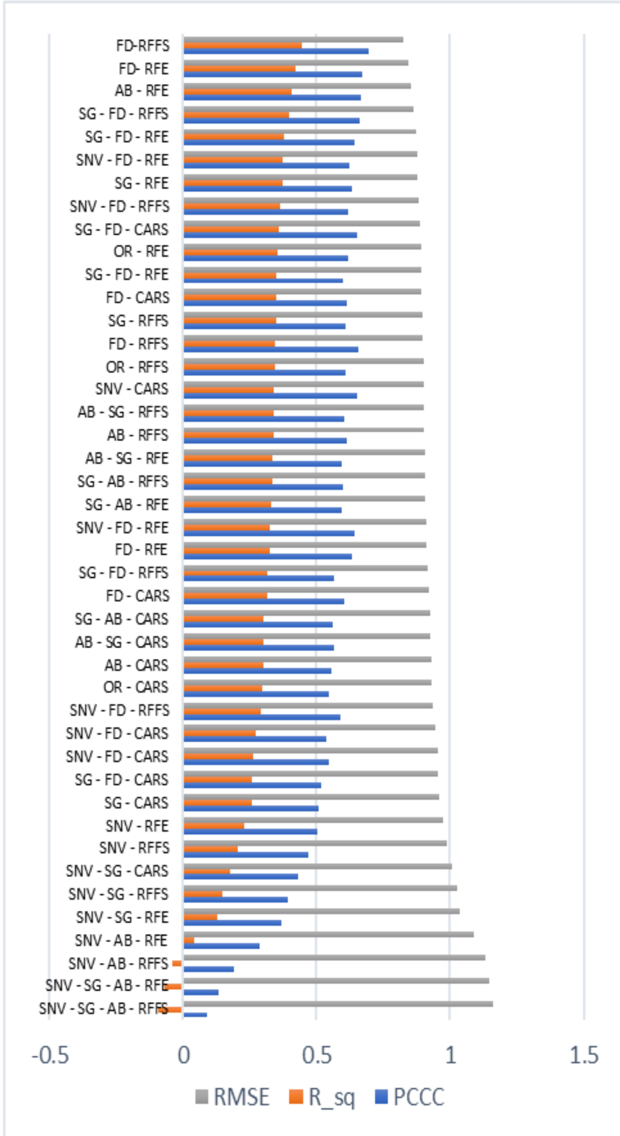

Figure 5. Test-based RMSE, R^2, and PCCC for SOM prediction. Abs: The prediction correlation coefficient (PCCC), coefficient of determination (R^2), Root mean square error (RMSE).

4.5. Bands sensitivity

In contrast, analysing the frequency of selected wavelengths across all feature selection and pre-processing combinations revealed a consistent clustering of informative spectral bands in two principal regions: the shortwave infrared-2 (SWIR-2) range (approximately 2207–2445 nm) and the visible–near infrared (visible–NIR) interval (approximately 418–801 nm). Within these zones, several bands, most notably 2207.9, 749.1, 2400.0, 734.7, and 611.0 nm (Table 2), were selected with high recurrence, highlighting their strong diagnostic relevance for predicting soil organic matter (SOM) in saline, semi-arid environments.

The prominence of SWIR-2 bands corresponds to well-known absorption features related to organic functional groups, clay-organic associations, and moisture-salt interactions, whereas the frequent selection of visible–NIR bands reflects sensitivity to soil colour, mineralogy, and surface characteristics.

When the Random Forest model was trained using these frequently recurring bands, it demonstrated improved stability and predictive performance, suggesting that band recurrence effectively indicates spectral informativeness. Overall, the agreement among multiple selection algorithms on these specific wavelength intervals provides strong evidence that reflectance patterns in the visible–NIR and SWIR-2 domains are key to the model's ability to detect subtle SOM variations. This finding underscores the advantage of combining Random Forest-based feature selection with derivative pre-processing to identify physiochemically meaningful spectral regions, thereby enhancing SOM prediction from EnMAP hyperspectral imagery.

Table 2. Frequency analysis of SOM sensitive selected bands

Frequent bands in all combinations	2207.9 (21);749.1 (20);2400.0 (20);734.7 (17); 611.0 (17);713.5 (17);599.4 (16);727.5 (16); 418.4 (15);2445.3 (15);545.6 (14);2345.8 (14); 2430.3 (14); 741.8 (13); 720.5 (13);801.2 (12); 2392.3 (12);706.6 (11);616.9 (11);2290.1 (11)

5. CONCLUSION

Accurate quantification of SOM within saline, semi-arid ecosystems is imperative for evaluating soil fertility and fostering sustainable land management practices. This investigation elucidated that a singular EnMAP hyperspectral image, when synergized with suitable spectral preprocessing and feature selection methodologies, can markedly improve SOM prediction in such complex landscapes. Employing RF regression as the predictive framework, we assessed five distinct preprocessing techniques (first derivative, second derivative, Savitzky–Golay, standard normal variate, and absorbance) alongside three feature selection algorithms (random forest-based feature selection, recursive feature elimination, and competitive adaptive reweighted sampling).

- The findings unequivocally indicate that unprocessed reflectance spectra exhibit only a minimal correlation with SOM variability, resulting in moderate predictive efficacy across all feature selection strategies. Conversely, spectral transformationsmost notably the first derivative considerably enhanced model precision. The integration of first derivative preprocessing with random forest-based feature selection (FD-RFFS) yielded the most superior performance metrics, attaining a prediction correlation coefficient of 0.697, an R^2 value of 0.446, and a root mean square error (RMSE) of 0.825. In comparison to the initial reflectance counterpart, this methodology augmented correlation by approximately 10–15%, enhanced explained variance by 20–30%, and diminished prediction error by 5–8%.

- In addition to performance metrics, a persistent observation across all preprocessing and feature selection combinations was the recurrent identification of spectral bands within two specific spectral regions: the visible–near infrared (418–801 nm) and the shortwave infrared-2 (2207–2445 nm). The convergence of multiple selection algorithms on these spectral intervals emphasizes their diagnostic significance for SOM prediction in studied saline, semi-arid soils.

- In summary, this study furnishes a rigorously validated framework for the operational monitoring of SOM utilizing EnMAP hyperspectral imagery. The results underscore that the amalgamation of derivative-based preprocessing with RF-driven feature selection proficiently isolates physiochemically relevant spectral regions, thereby addressing the complexities imposed by soil salinity, moisture content, and mineralogical diversity.
- The synergistic application of these methodologies offers a robust pathway for augmenting the predictive capability of spaceborne hyperspectral data, thereby bolstering large-scale soil monitoring initiatives in ecologically sensitive areas.

References

Achemrk, A., Ouzemou, J.-E., Laamrani, A., El Battay, A., Hajaj, S., Oussaoui, S., & Chehbouni, A. (2026). Tracking Quarter-Century Spatio-Temporal Soil Salinization Dynamics in Semi-Arid Landscapes Using Earth Observation and Machine Learning. *Remote Sensing*, *18*(5), 687.

Al-Ali, Z., Bannari, A., Rhinane, H., El-Battay, A., Shahid, S. A., & Hameid, N. (2021). Validation and comparison of physical models for soil salinity mapping over an arid landscape using spectral reflectance measurements and Landsat-OLI data. *Remote Sensing*, *13*(3), 494.

Bian, X. (2022). Spectral preprocessing methods. In *Chemometric methods in analytical spectroscopy technology* (pp. 111-168). Springer.

Bourriz, M., Hajji, H., Laamrani, A., Elbouanani, N., Abdelali, H. A., Bourzeix, F., El-Battay, A., Amazirh, A., & Chehbouni, A. (2025). Integration of hyperspectral imaging and AI techniques for crop type mapping: Present status, trends, and challenges. *Remote Sensing*, *17*(9), 1574.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Dai, J., Yang, H., Maimaitituersun, A., & Maimaitiaili, K. (2026). Development and contribution analysis of a multi-depth soil organic matter prediction model integrating VIS–NIR spectroscopy and environmental covariates. *Plant and Soil*, 1-24.

Dai, L., Wang, Z., Zhuo, Z., Ma, Y., Shi, Z., & Chen, S. (2025). Prediction of soil organic carbon fractions in tropical cropland using a regional visible and near-infrared spectral library and machine learning. *Soil and Tillage Research*, *245*, 106297.

Ghasemi, J. B., & Tavakoli, H. (2013). Application of random forest regression to spectral multivariate calibration. *Analytical Methods*, *5*(7), 1863-1871.

He, S., Tan, S., Shen, L., & Zhou, Q. (2023). Soil organic matter estimation model integrating spectral and profile features. *Sensors*, *23*(24), 9868.

Li, H., Liang, Y., Xu, Q., & Cao, D. (2009). Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Analytica Chimica Acta*, *648*(1), 77-84.

Ouzemou, J.-E., Laamrani, A., El Battay, A., & Whalen, J. K. (2025). Predicting soil salinity based on soil/water extracts in a semi-arid region of Morocco. *Soil Systems*, *9*(1), 3.

Zhang, X., Dong, C., Liu, H., Meng, X., Luo, C., Han, Y., & Ai, H. (2024). Methodology for regional soil organic matter prediction with spectroscopy: Optimal sample grouping, input variables, and prediction model. *Remote Sensing*, *16*(3), 565.

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-947-2026