Journal of Engineering Science and Technology
Special Issue on Development of Sustainability Systems - III, 147 - 162

EXPLAINABLE MACHINE LEARNING FOR SOIL QUALITY PREDICTION IN SEMI-ARID NORTHERN IRAQ

ARAZ R. ABRAHIM[1,*], OMAR F. MOHAMMAD[2],
AMJED M. ISAMEL[1], HIKMET GUNAL[3], MOHAMMED ODAY[3]

[1]Technical Institute of Administration, Duhok Polytechnic University, Iraq
[2]Department of Computer Science, College of Education for Pure Science, University of Al-Hamdaniya, Mosul, Iraq
[3]Department of Soil Science and Plant Nutrition, Faculty of Agriculture, Harran University, Sanliurfa, Turkey
*Corresponding Author: araz.abrahim@dpu.edu.krd

Abstract

Accurate soil quality assessment is critical for sustainable land management, especially given the growing environmental challenges to land productivity. This study introduces an interpretable machine learning framework for predicting soil quality in the Amediya subdistrict of northern Iraq, combining topography, remote sensing, and physicochemical data. The Soil Quality Index (SQI) combines key soil factors such as SOM, pH, $CaCO_3$, BD, soil type, elevation, slope, aspect, and NDVI. Principal Component Analysis (PCA) weighted variables according to their contribution to soil quality. Five models were trained: Multilayer Perceptron, Gradient Boosting Regressor (GBR), XGBoost, Extra Trees, and Decision Tree. GBR performed the best ($R^2 = 0.986$, MAE = 0.002, RMSE = 0.006). The SHAP and LIME techniques improved interpretability by identifying SOM, pH, and soil depth as primary determinants globally, with BD and aspect being relevant locally. A spatial study revealed worse soil quality on steep slopes and in areas with little SOM. The findings highlight the importance of combining predictive modelling with explainable AI for accurate, transparent soil quality evaluation, which will aid precision agriculture and land-use planning in semi-arid mountainous environments.

Keywords: Explainable artificial intelligence (XAI), Machine learning, Remote sensing, Soil quality.

1. Introduction

Soil quality is an important indication of ecosystem health and agricultural production, since it influences water retention, nutrient cycling, and plant development [1]. In semi-arid areas like northern Iraq, preserving soil health is essential for sustainable land use, especially considering unsustainable agricultural methods, climatic unpredictability, and land degradation. However, traditional methods relying on field sampling and laboratory analysis are costly, time-consuming, and spatially limited, hindering scalable monitoring [2].

Digital Soil Mapping (DSM), which combines georeferenced soil data with environmental variables and spatial prediction models, has become a viable way to overcome these constraints [3]. Among machine learning (ML) approaches, algorithms such as Gradient Boosting Decision Tree (GBDT) and Extreme Gradient Boosting (XGBoost) excel at modelling nonlinear relationships between soil properties and environmental covariates (e.g., topography and remote sensing data) [4].

Although these models are accurate, they frequently operate as "black boxes" with little interpretability, which makes it difficult to use them in evidence-based land management and policy [5, 6], Explainable Artificial Intelligence (XAI) techniques like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) have been introduced recently to interpret and visualize model predictions by quantifying the contribution of each input variable [7].

Therefore, by combining XAI with ML, model transparency may be improved while maintaining predictive capability, allowing stakeholders to understand and apply model insights more effectively [8]. The effectiveness of such integration has previously been demonstrated in environmental science. For example, SHAP analysis shows that NDVI and topographic characteristics are the most important factors in forecasting soil organic carbon. LIME has been used to identify location-specific parameters that affect soil fertility and salinity [9].

To address these gaps, these methods raise key questions: To what extent can various machine learning models forecast soil quality in contexts that are diverse and data-poor contexts? Which environmental factors have the most effect on these forecasts? Furthermore, how can model transparency be achieved without compromising accuracy?

Environmental factors included remote sensing indices, e.g., Normalized Difference Vegetation Index, NDVI, a measure of live green vegetation, and topography parameters (elevation, slope, aspect). SHAP and LIME were used to analyse the machine learning models and identify the most important covariates. The models, RF, SVR, and XGBoost, were trained and assessed using cross-validation and standard metrics (R^2, RMSE, and MAE).

To evaluate soil conditions in the semi-arid Amediya sub-district in northern Iraq, this study aims to develop an integrated, interpretable machine learning framework. In particular, the study (1) used a variety of environmental and soil variables to predict a spatially explicit Soil Quality Index (SQI); (2) used Explainable Artificial Intelligence (XAI) techniques like SHAP and LIME to identify the most influential predictors; and (3) offers an open and scientifically supported foundation for soil conservation and land-use planning.

To close the gap between sophisticated data-driven modelling and its real-world implementation in sustainable land management, this study addresses both prediction performance and model interpretability. This research examines these issues by focusing on the Amediya sub-district, located in northern Iraq, 70 km north of Duhok Province.

The area is around 20.06 km² in size, sits at an elevation of about 1400 m, and has a dry Mediterranean climate with chilly winters, dry summers, and seasonal precipitation that primarily takes the form of rain or snow. The topography is hilly, with scant vegetation consisting primarily of grasses and shrubs. Key soil quality characteristics, including pH, organic matter, calcium carbonate, and bulk density, were examined in 80 soil samples collected using a 500 × 500 m grid.

2. Materials and Methods

2.1. Study area

The study was carried out at an average height of 1400 m in the Amedyia sub-district, which is situated in Northern Iraq, about 70 km north of Duhok Province, shown in Fig. 1. It is located between 37°40' and 37°60' N and 43°27' and 43°30' E, in an area of 20.06 km². The area has a transitional Mediterranean climate, characterized by dry summers, low humidity, and cold winters, with most precipitation falling as rain and snow.

Most of the vegetation is made up of grasslands and shrublands, making it somewhat sparse when compared to the rest of Duhok. The predominant vegetation types are high forest (7.7%), medium forest (8.27%), grasslands (16.55%), and shrub formations (67.44%) [10]. The region is mostly utilized for natural grazing, and dry oak is common at higher elevations.

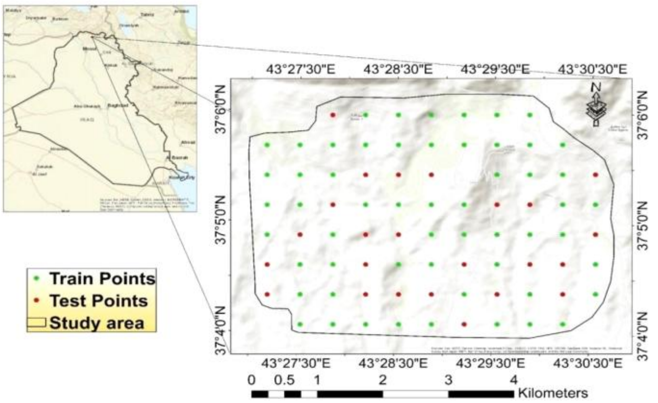

Fig. 1. Geographical locations of the study area, training, and test point.

2.2. Soil sampling

The research area was arranged in 500 × 500 m grid cells (Fig. 1), and surface soil (0-20 cm depth) samples were collected near each cell's corners. GPS coordinates were recorded, and alternate locations were substituted for areas that were inaccessible due to buildings, water bodies, or topography. The samples were crushed, air-dried, sieved through a 2 mm mesh screen, and were free of roots and stones. Using the Walkley-Black dichromate oxidation technique, the amount of soil organic matter (SOM) was measured. Using sodium hexametaphosphate as a dispersion agent [11], the hydrometer technique was used to examine soil texture [12], and the Scheibler calcimeter method was used to quantify calcium carbonate ($CaCO_3$) [13]. A steel cylinder of 100 cm^3 (5 cm in diameter and 5.1 cm in height) was used to test bulk density, as recommended by McLean [14], and soil pH was determined following the procedure proposed by Zhou et al. [15].

2.3. Remote sensing and topographic data processing

The spatial prediction of soil quality indicators in the Amediya sub-district was improved using remote sensing and topography data. The conventional formula for NDVI, a critical indication of vegetative health and biomass, was applied to Sentinel-2A Level-1C images at 10 m resolution. The NDVI is strongly related to soil organic carbon and erosion potential [16].

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{1}$$

The topographic variables height, slope, and aspect were derived from a 30 m SRTM DEM using QGIS 3.28. These variables affect plant distribution, soil formation, erosion risk, and water quality [17]. All spatial layers were resampled to a consistent 30 m resolution and matched to the soil sampling grid. The machine learning feature matrix was built by extracting raster data from sample locations. NDVI, slope, elevation, and aspect were included as continuous predictors.

2.4. Principal component analysis (PCA) and SQI construction

A multivariate statistical method, Principal Component Analysis (PCA), was used to assign objective weights to environmental and soil indices [18], ensuring a proportional contribution to the evaluation of soil quality (Table 1). PCA preserves most of the data variability while reducing correlated variables to uncorrelated components. NDVI, bulk density, SOM, $CaCO_3$, slope, aspect, elevation, soil depth, soil type, and pH were among the ten indicators. Min-max normalization was used to standardize all variables to a 0-1 scale.

Table 1. Relative weights of soil quality indicators in the composite Soil Quality Index (SQI).

Soil indicators	Slope	SOM	Soil type	Ph	BD (g/cm^3)	NDVI	Aspect	Soil depth	$CaCO_3$	Elevation
Weight	0.285	0.265	0.104	0.083	0.077	0.076	0.053	0.026	0.023	0.009
(%)	28.5	26.6	10.4	8.3	7.8	7.5	5.3	2.5	2.3	0.9

The components with eigenvalues greater than one were kept. By squaring factor loadings and multiplying by the associated eigenvalues, indicator weights, Wi, were determined by Eq. (2).

$$Wi = \frac{\sum_{j=1}^{m} Lij2.EVj}{\sum_{j=1}^{m} Lij2.EVj} \tag{2}$$

For variables that contribute more to total variation, this approach guarantees a larger effect, shown in Fig. 2.

A Soil Quality Index (SQI) was calculated after weighing. Every variable had a score between 0 and 1. While negatively correlated variables (bulk density, slope, and $CaCO_3$) utilized Eq. (4), positively correlated variables (NDVI, SOM, soil depth, and elevation) used Eq. (3).

$$Si = \frac{\mathrm{Xi - Xmin}}{\mathrm{Xmax - Xmin}} \tag{3}$$

The final SQI was generated using a weighted additive model [19] Eq. (5), which summed the products of each indicator's standardized score and PCA-derived weight. Spatial interpolation algorithms were used to create SQI maps highlighting areas with changing soil quality. These results are useful for sustainable land-use planning in the semi-arid, mountainous Amedyia subdistrict.

$$Si = \frac{\mathrm{Xmax{-}Xi}}{\mathrm{Xmax - Xmin}} \tag{4}$$

$$SQI = \sum_{I=1}^{n} \mathrm{W.Si} \tag{5}$$

where SOI indicates soil quality index, W indicates the weight assigned by PCA, and SI is the value of the indicators.

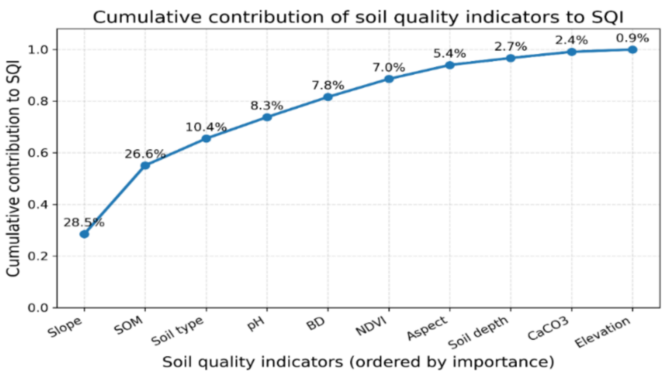

Fig. 2. Cumulative contribution of soil quality indicators to the composite SQI.

2.5. Machine learning models for soil quality prediction

Soil quality was predicted using five ML models that employed ensemble tree-based techniques and neural networks on an integrated dataset comprising physicochemical characteristics, topographic features, and remote sensing variables (Fig. 3). These models were chosen because of their shown ability to accurately depict spatial environments.

The Multilayer Perceptron (MLP) is a feedforward neural network that was trained using gradient descent and backpropagation. It has input, hidden, and output layers. It successfully depicts intricate nonlinear interactions between soil characteristics that conventional models could miss [20].

GBR uses gradient descent to create additive models sequentially, with each tree fixing the mistakes of the one before it. It can represent complicated patterns in soil properties and is appropriate for datasets with moderate to high feature complexity [7].

XGBoost uses regularization (L1/L2), parallel processing, and second-order optimization to improve on traditional boosting. Spatial forecasts for agriculture and the environment frequently utilize it because of its high accuracy and resilience [3].

Extra Trees creates unpruned decision trees by adding randomization to feature selection and splitting thresholds. This works especially well for noisy and diverse soil datasets and improves ensemble diversity while decreasing variance [21].

Decision Tree Regression divides data using recursive binary splits. It is straightforward and interpretable, but without pruning, it is prone to overfitting. However, because it is transparent, it is useful for a preliminary exploratory study. Scikit-learn XGBoost was used to implement all models in Python. Soil pH, bulk density, $CaCO_3$, SOM, slope, elevation, soil depth, aspect, and NDVI were among the input characteristics [21].

2.6. Explainable machine learning: SHAP and LIME methods

To improve the interpretability of machine learning models for predicting soil quality, two popular Explainable AI (XAI) approaches are used, SHAP and LIME [22]. These techniques aid in bridging the gap between useful, scientifically grounded land management decisions and the outputs of intricate models.

SHAP offers reliable, model-neutral insights into feature relevance by employing Shapley values from cooperative game theory to provide both global and local feature attributions [16]. The impact of important factors like pH, SOM, slope, and NDVI was examined using five ML models (MLP, XGBoost, GBR, Extra Trees, and Decision Tree), with visualizations showing the strength and direction of each variable's influence.

LIME enhances SHAP by providing localized, instance-level explanations that use basic surrogate models to approximate complex predictions [16, 23], This technique assisted in identifying region-specific feature contributions, such as the impact of slope or low pH on soil quality, hence facilitating precision agriculture and effective stakeholder engagement.

2.7. Model validation and accuracy assessment

Each machine learning model was trained using a 10-fold cross-validation strategy, which minimized overfitting and increased prediction resilience. To test model performance, R^2, MAE, and RMSE were used to assess accuracy and generality in predicting the Soil Quality Index. Visual diagnostics, such as residual and anticipated vs. actual plots, were also employed to detect bias or heteroscedasticity.

2.8. Statistical analysis

Descriptive statistics such as minimum, maximum, mean, standard deviation, skewness, kurtosis, and coefficient of variation (CV%) were used to assess data distribution and variability among the 80 soil samples. Data normality and variability were verified before modelling to guarantee robustness. All statistical analyses, data preparation, and machine learning model creation were done in Python 3.10 with the Scikit-learn, XGBoost, and SHAP libraries.

3. Results and Discussion

3.1. Descriptive statistics of soil indicators

Descriptive analysis, shown in Table 2, demonstrates significant heterogeneity in soil parameters throughout the study region. SOM ranged from 0.255% to 3.840% (mean = 1.357%), with considerable variability (CV = 65.09%), positive skew (1.015), and minor kurtosis (0.182), indicating a right-skewed distribution with several significant outliers. Soil pH (7.16-8.18; mean = 7.727) had a minimal CV of 3.57%, a little negative skew (-0.211), and a platykurtic distribution (-0.92), which is typical of calcareous soil**.**

Table 2. Descriptive statistics of soil indicators in the Amediya region.

Soil variables	Minimum	Maximum	Mean	SD	Skewness	Kurtosis	CV (%)
SOM (%)	0.255	3.84	1.357	0.88	1.015	0.182	65.09
pH	7.16	8.18	7.727	0.27	-0.211	-0.92	3.57
BD (%)	0.801	1.789	1.291	0.23	-0.15	-0.532	18.48
$CaCO_3$	4.874	19.498	13.62	2.74	-0.49	1.201	20.15
Soil depth	23	120	45.13	19.68	1.601	2.698	43.61
Slope value	2.3	27.7	13.43	6.86	0.292	-1.141	51.08
Aspect	0	337.16	181.6	72.3	-0.311	-0.556	39.79
NDVI	0.154	0.7	0.308	0.10	1.526	2.954	34.85
Elevation	825	1195	967.3	86.90	0.575	-0.454	8.984

Moderate variability of 18.48%, modest negative skew of –0.150, and platykurtosis of –0.532 were observed in the bulk density (0.801–1.789 g/cm³; mean = 1.291), indicating generally uniform compaction. There was significant variability (CV = 20.15%), negative skew (–0.490), and considerable leptokurtosis (1.201) in the $CaCO_3$ concentration (4.874–19.498%; mean = 13.62%).

The soil depth (23–120 cm; mean = 45.14 cm) showed leptokurtosis (2.698), negative skew (–0.490), and considerable variability (CV = 43.61%), indicating deeper profiles in certain places. Positive skew (0.292) and platykurtosis (–1.141) were present, and the slope (2.3–27.7%; mean = 13.43%) ranged greatly (CV = 51.08%). There was platykurtosis of -0.556, negative skew of -0.311, and uniform distribution of 39.79% in the aspect (0° – 337.16°; mean = 181.67°).

The NDVI (0.154 – 0.700; mean = 0.308) showed leptokurtic of 2.954, significant right-skewed (1.526), and high variability (CV = 34.85%), indicating scant vegetation in most of the regions. Moderate variability of 8.98%, positive skew (0.575), and platykurtosis (–0.454) were observed at elevation (825–1195 m; mean = 967.3 m). Overall, the dataset shows significant geographic variation in vegetative, organic, and topographic variables, although bulk density and pH are comparatively constant. These trends demonstrate the necessity of targeted soil management and well-informed modelling techniques.

3.2. Interpretable machine learning results for SQ assessment

To understand machine learning predictions and determine the proportional contribution of input factors to soil quality evaluation, explainable AI (XAI) techniques-specifically, SHAP and LIME-were utilized. By clarifying how topographic and Physico-chemical elements-such as organic matter, bulk density, slope, and NDVI-affect forecast results, these techniques improve model transparency and dependability. These methods allow the model behaviour to be aligned with well-established pedological concepts by providing both global (SHAP) and local (LIME) interpretability. By increasing stakeholder trust and model explainability, this strategy makes data-driven land-use planning and soil management easier.

3.2.1. SHAP analysis: Global and local feature importance

To measure the local and global effects of input factors on soil quality projections, SHAP values have become more popular in machine learning models. For instance, major contributors like Soil Organic Matter (SOM), soil depth, elevation, and pH have been demonstrated to significantly influence model outputs, in line with findings in Fig. 3. SOM frequently exhibits the greatest local effect (SHAP ≈ -0.07 to 0.025 in MLP models). According to Lehmann and Kleber [24], the most detrimental effects of decreased SOM are consistent with its vital function in soil biological activity, water retention, and nutrient retention.

The restricting impact that shallow profiles and higher elevations have on water and nutrient availability is shown by negative SHAP values for soil depth and elevation. This connection has been previously reported in soil science literature [14]. Furthermore, pH showed a significant beneficial impact in MLP and XGBoost models (SHAP ≈ –0.01 to 0.07), supporting the well-established idea that microbial activity and nutrient availability are maximized at near-neutral pH values [21]. For vegetation vigour and soil fertility, the Normalized Difference Vegetation Index (NDVI) demonstrated consistent positive contributions [25].

Calcium carbonate (CaCO3) and bulk density had negative effects on predictions, which is indicative of carbonate-induced nutrient imbalances and soil compaction, two well-known elements that impact soil health [26]. There have also been reports of indirect effects of slope and aspect on patterns of soil quality, which are probably mediated by runoff and microclimate effects [27]. In several situations, greater slope values indicated slight beneficial effects, presumably due to enhanced drainage; nonetheless, steep slopes usually increase erosion risk

In soil quality assessment, the application of SHAP values generally improves the interpretability and transparency of machine learning models, allowing for a more nuanced understanding of feature importance at both the global and individual prediction levels. This raises the scientific rigor and reliability of ML-based environmental models [28].

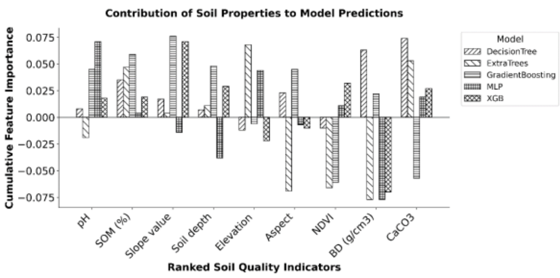

Fig. 3. SHAP explanation of local feature contributions for five ML models (Decision tree, extra trees, gradient boosting, MLP, XGBoost).

3.2.2. LIME-based local interpretability of soil quality predictions

The global findings from SHAP were supplemented with instance-specific contributions of soil characteristics to soil quality forecasts using Local Interpretable Model-Agnostic Explanations (LIME) (Fig. 4). This method clarifies how the effects of features fluctuate across value ranges and soil types. The Extra Trees model showed that the soil pH values between -0.88 and 0.12 had the greatest negative effect (~ -0.075), which is in line with the established decrease in microbial activity and nutrient availability in acidic environments. Similarly, bulk density (BD) in the -0.48 to -0.02 range had a negative effect (~ -0.04), indicating that soil compaction limits root development and aeration.

The negative contribution (~ -0.03) from lower elevation values (-0.20 to 0.72) was probably caused by shallow soil formation and inadequate drainage, which supports earlier research that linked low elevation to lower soil productivity. On the other hand, a small positive effect (~+0.01) was observed at Soil Depth > 0.47, suggesting improved rooting ability and water retention. Soil Organic Matter (SOM) levels above -0.66 had a favourable impact on predictions (~ +0.04), highlighting its function in enhancing biological activity, cation exchange, and aggregation. Steep slopes may need erosion prevention. However, slope values over 0.80 also indicated a minor beneficial influence (~ +0.05), maybe because of improved drainage.

The Extra Trees model indicated that soil pH values between -0.88 and 0.12 had the most substantial negative effect (~ -0.075), aligning with established findings that acidic environments can reduce microbial activity and nutrient availability. Similarly, bulk density (BD) within the -0.48 to -0.02 range negatively impacted predictions (~-0.04), consistent with the literature suggesting that soil compaction limits root development and aeration [29]. Overall, LIME offers comprehensive local interpretability, enhancing comprehension of the ways in which certain variable ranges affect forecasts of soil quality. These revelations verify domain expertise and validate the use of explainable AI in customized regional soil management.

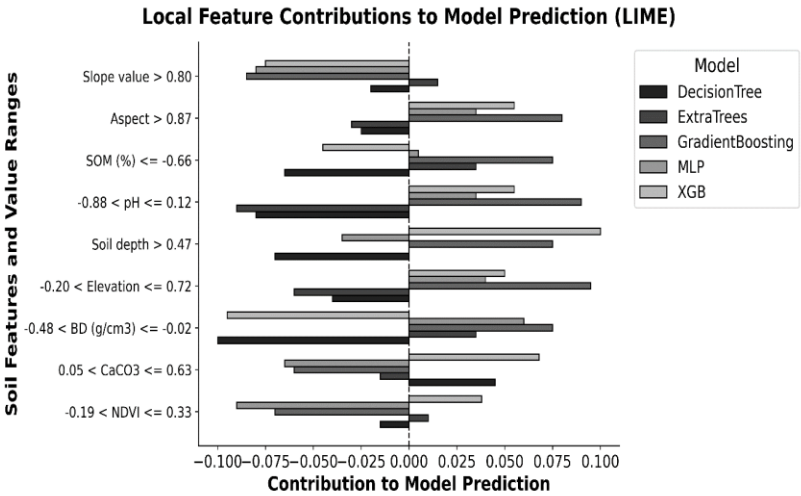

Fig. 4. LIME explanation of local feature contributions for five ML models (Decision Tree, Extra Trees, Gradient Boosting, MLP, XGBoost).

3.3. Spatial distribution of soil quality

Soil quality spatial predictions were generated using five machine learning algorithms: MLP, XGBoost, Gradient Boosting Regressor (GBR), Extra Trees, and Decision Tree (DT) (Figs. 5 and 6). Predictions were classified into very low, low, and moderate quality zones, reflecting underlying edaphic heterogeneity driven by organic matter, slope, bulk density, and vegetation.

The GBR model demonstrated the highest spatial resolution, clearly outlining large moderate-quality regions in the centre and to the east, while also revealing scattered low-quality areas in the south and northeast. This illustrates GBR's capability in modelling intricate, nonlinear interactions, supporting findings that emphasize its responsiveness to micro-variations in agricultural landscapes. XGBoost exhibited comparable fine-scale spatial variation, especially in the transitional areas between poor and moderate soil quality. Its resilience to multicollinearity and outliers improves predictive accuracy, aligning with previous studies on soil organic carbon mapping.

The MLP model delivered well-balanced predictions with seamless transitions across quality classes, likely owing to its ensemble learning framework, which mitigates overfitting. The moderate-quality hotspots located centrally and, in the southeast, correspond with advantageous elevation, deeper soil profiles, and vegetation cover. Conversely, Extra Trees produced rougher spatial patterns characterized by larger low-quality areas and diminished accuracy in reflecting micro-environmental variability. This aligns with previous findings that Extra Trees tend to generalize in intricate terrain, thereby restricting spatial detail.

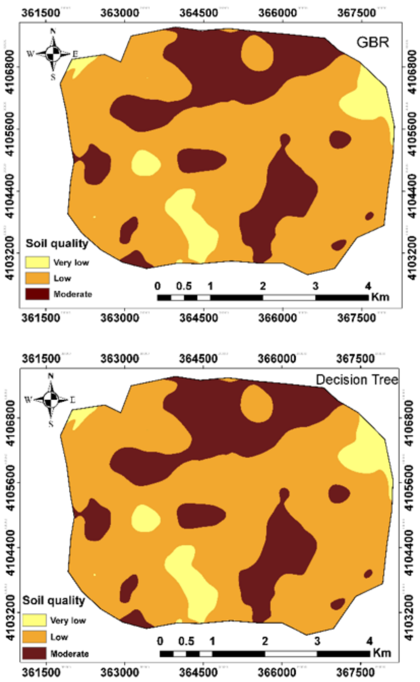

Fig. 5. Spatial distribution map of soil quality in the Amediya region produced by Extra Tree and GBR, respectively.

The DT model yielded the most basic and coarse spatial delineation, primarily indicating very low soil quality in the central and western subregions. This result is consistent with the established limitations of DT in representing spatial complexity and its diminished accuracy in heterogeneous landscapes. Regardless of the algorithmic variations, shared geographical patterns were observed, particularly the identification of low-quality soils in areas with low relief or low organic matter. These persistent trends underscore the significance of the chosen predictors and illustrate the effectiveness of machine learning in delineating continuous soil quality gradients. Models with greater spatial sensitivity (such as MLP and XGBoost) are recommended for precision agriculture and targeted soil management. In contrast, simpler models (such as DT and Extra Trees) may be suitable for quick evaluations with limited datasets.

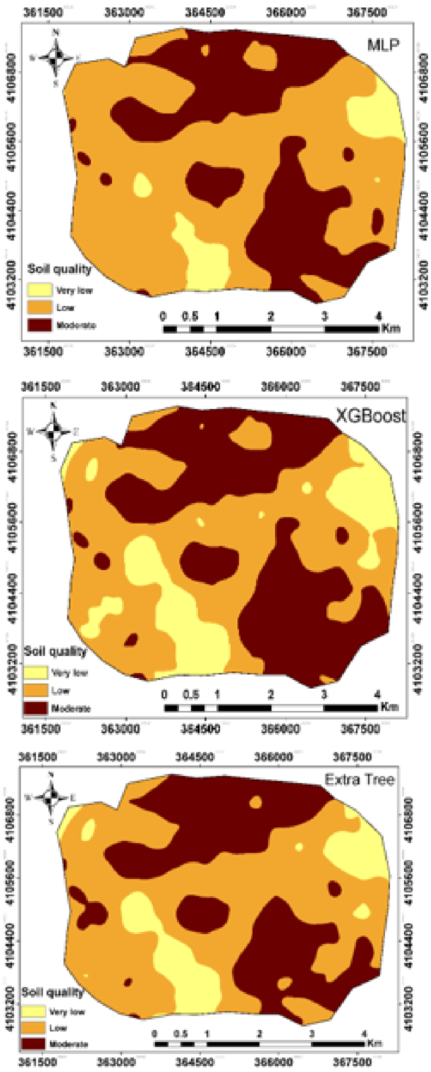

Fig. 6. Spatial distribution map of soil quality in the Amediya region produced by Decision, XGBoost, and MLP.

3.4. Accuracy assessment of machine learning models

The predictive performance of the evaluated machine learning architectures was quantified using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2), as summarized in Table 3. Statistical analysis reveals that the Gradient Boosting Regressor (GBR) attained the highest predictive precision ($R^2 = 0.986$, MAE = 0.002, RMSE = 0.006). This high fidelity in capturing complex, nonlinear soil-environmental interactions align with the findings of Rahaman and Das [2], who established the efficacy of GBR ensemble methods for high-dimensional pedological datasets. The higher performance of boosting-based models (GBR and XGBoost) compared to the Decision Tree is due to their iterative learning process, which successively reduces residuals. The single Decision Tree is more sensitive to high variation and noise in diverse soil datasets.

Table 3. Accuracy assessment metrics (MAE, RMSE, R^2) for five machine learning models used in soil quality prediction.

Models	MLP	XGBR	GBR	Extra Trees	Decision Trees
MAE	0.004	0.004	0.002	0.007	0.007
RMSE	0.009	0.007	0.006	0.012	0.017
R^2	0.974	0.983	0.986	0.954	0.899

The statistical relationship between observed and predicted SQI values for the top-performing ensemble models is visually articulated in Fig. 7. This figure illustrates the superior performance of the GBR (a) and XGBoost (b) architectures, where data points exhibit a high degree of homoscedasticity, clustering tightly along the 1:1 identity line. The minimal dispersion in Fig. 7 signifies that these boosting-based algorithms successfully minimized variance, providing the most reliable spatial estimations for the semi-arid study region.

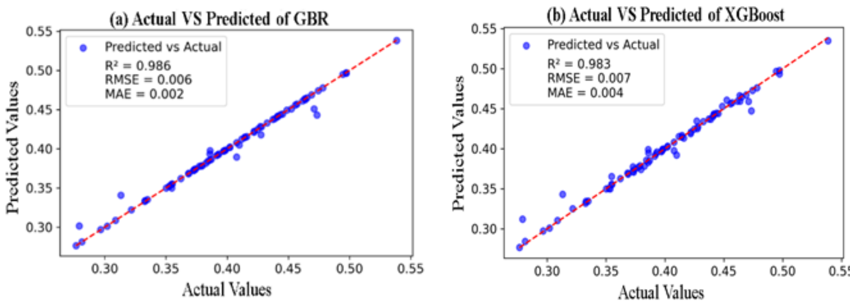

Fig. 7. Actual vs predicted value of machine learning models, (a) GBR, and (b) XGBoost.

Conversely, Fig. 8 displays the predictive performance and error distributions for the remaining three architectures: (a) MLP, (b) Extra tree, and (c) Decision tree. While the MLP model maintained a strong correlation ($R^2 = 0.974$), Fig. 8 reveals a noticeable increase in dispersion and residual fluctuations for the single-tree and alternative ensemble models. Specifically, the Decision Tree exhibited the lowest performance ($R^2 = 0.899$) and substantial heteroscedasticity, suggesting a limited ability to capture the multifaceted environmental gradients in the study area. These observations in Fig. 8 corroborate existing critiques regarding the limitations of non-ensemble tree models in complex spatial soil mapping, where they often fail to account for localized variability.

The comparative hierarchy of model performance was established as GBR > XGBoost > MLP > Extra Trees > Decision Tree. These results emphasize that while deep learning and alternative ensembles provide robust results, boosting-based architectures (shown in Fig. 7) offer the most rigorous framework for SQI prediction. The integration of these high-performance models with interpretability tools such as SHAP and LIME ensures that the resulting digital soil maps are both statistically accurate and pedologically interpretable.

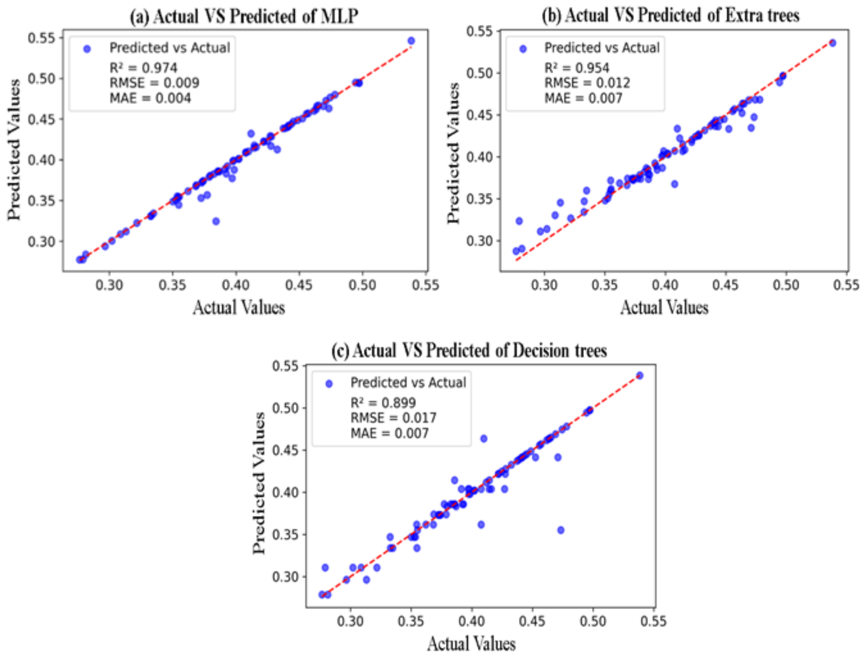

Fig. 8. Actual vs predicted value of machine learning models, (a) MLP, (b) Extra tree, and (c) Decision tree.

4. Conclusion

The study established a transparent framework for soil quality evaluation by effectively integrating five machine learning models with SHAP and LIME. The best predictive architecture for capturing intricate, nonlinear soil-environment connections is the Gradient Boosting Regressor (GBR). Slope, pH, and soil organic matter (SOM) were determined by SHAP analysis to be the main worldwide factors influencing soil quality. The "black-box" aspect of classical machine learning was successfully resolved by complementary LIME assessments, which demonstrated how local variable fluctuations affect individual site predictions.

Spatial modelling revealed important weaknesses in locations with steep topography and low SOM, which are directly linked to increased erosion risks and decreased fertility. These findings show that combining predictive modelling with Explainable AI (XAI) creates a scientifically sound framework for precision agriculture and sustainable land management. Finally, our study bridges the gap between complex computational models and stakeholder trust, establishing a repeatable approach for data-driven environmental monitoring in semi-arid locations. Future studies should use more datasets and temporal monitoring to improve model generalizability.

Nomenclatures

BD	Bulk Density
$CaCO_3$	Calcium Carbonate
DEM	Digital Elevation Model
DT	Decision Tree
Extra Trees	Extremely Randomized Trees
GBR	Gradient Boosting Regressor
LIME	Local Interpretable Model-agnostic Explanations
MAE	Mean Absolute Error
MLP	Multilayer Perceptron
NDVI	Normalized Difference Vegetation Index
NIR	Near-Infrared
PCA	Principal Component Analysis
pH	Potential of Hydrogen
R^2	Coefficient of Determination
RED	Red Band (Visible Spectrum)
RMSE	Root Mean Square Error
SHAP	Shapley Additive exPlanations
SOM	Soil Organic Matter
SQI	Soil Quality Index
XAI	Explainable Artificial Intelligence
XGBoost	eXtreme Gradient Boosting

References

1. Bünemann, E.K. et al. (2018). Soil quality-A critical review. *Soil Biology and Biochemistry*, 120, 105-125.
2. Rahman, R.; and Das, K.N. (2025). Artificial intelligence and machine learning in soil analysis for precision agriculture: A review. *Journal of Experimental Agriculture International*, 47(5), 511-524.
3. McBratney, A.B.; Santos, M.L.M.; and Minasny, B. (2003). On digital soil mapping. *Geoderma*, 117(1-2), 3-52.
4. Mahmoudzadeh, H.; Matinfar, H.R.; Taghizadeh-Mehrjardi, R.; and Kerry, R. (2020). Spatial prediction of soil organic carbon using machine learning techniques in western Iran. Geoderma Regional, 21, e00260
5. Nwamekwe, C.O.; Ewuzie, N.V.; Okpala, C.; Ezeanyim, O.C.; Nwabueze, C.V.; and Nwabunwanne, E.C. (2025). Optimizing machine learning models for soil fertility analysis: Insights from feature engineering and data localization. *Gazi University Journal of Science Part A*: *Engineering and Innovation*, 12(1), 36-60.
6. Ahmad, M.; Al Zubi, M.; Almujibah, H.; Sabri Sabri, M.M.; Mustafvi, J.B.; Haq, S.; Ouahbi, T.; and Alzlfawi, A. (2025). Improved prediction of soil shear

strength using machine learning algorithms: Interpretability analysis using SHapley Additive exPlanations. *Frontiers in Earth Science*, 13, 1542291.

7. Wang, X.; Yang, Y.; Lv, J.; and He, H. (2023). Past, present and future of the applications of machine learning in soil science and hydrology. *Soil and Water Research*, 18(2), 67-80.
8. Zayani, H.; Fouad, Y.; Michot, D.; Kassouk, Z.; Baghdadi, N.; Vaudour, E.; Lili-Chabaane, Z.; and Walter, C. (2023). Using machine-learning algorithms to predict soil organic carbon content from combined remote sensing imagery and laboratory Vis-NIR spectral datasets. *Remote Sensing*, 15(17), 4264.
9. Günal, H.; Ismael, A.M.; Kılıç, M.; and Budak, M. (2024). Optimizing soil organic carbon prediction in arid environments: A gradient boosting-kriging approach in Northern Iraq. *Research Square*, preprint.
10. Walkley, A.; and Black, I.A. (1934). An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. *Soil Science*, 37(1), 29-38.
11. Gee, G.W.; and Bauder, J.W. (1986). *Particle-size analysis*. In Klute, A. (Ed.), *Methods of soil analysis*: *Part* 1 *Physical and mineralogical methods*. American Society of Agronomy, Inc., Madison, 383-411.
12. Nelson, D.W.; and Sommers, L.E. (1982). *Total carbon*, *organic carbon*, *and organic matter*. In Page, A.L. (Ed.), *Methods of soil analysis*: *Part* 2 *Chemical and microbiological properties*. Soil Science Society of America, Inc, 539-57.
13. Black, C.A.; Evans, D.D.; White, J.L.; Ensminger, L.E.; and Clark, F.E. (1965). *Methods of soil analysis*: *Part* 2 *Chemical and microbiological properties*. American Society of Agronomy, Inc.
14. McLean, E.O. (1982). *Soil pH and lime requirement*. In Page, A.L.; Miller, R.H.; and Keeney, D.R. (Eds.), *Methods of soil analysis*: *Part* 2 *Chemical and microbiological properties*. Soil Science Society of America, Inc., 199-224.
15. Zhou, X.; Guan, H.; Xie, H.; and Wilson, J.L. (2009). Analysis and optimization of NDVI definitions and areal fraction models in remote sensing of vegetation. *International Journal of Remote Sensing*, 30(3), 721-751.
16. Gong, Z.; Kawamura, K.; Ishikawa, N.; Goto, M.; Wulan, T.; Alateng, D.; Yin, T.; and Ito, Y. (2015). MODIS normalized difference vegetation index (NDVI) and vegetation phenology dynamics in the Inner Mongolia grassland. *Solid Earth*, 6(4), 1185-1194.
17. Muhsina, A.; Joseph, B.; and Kumar, V. (2020). Principal component analysis on soil fertility parameters of vegetable growing panchayats/locations in Ernakulam district of Kerala. *International Research Journal of Pure and Applied Chemistry*, 21(10), 58-65.
18. Andrews, S.S.; Karlen, D.L.; and Cambardella, C.A. (2004). The soil management assessment framework: A quantitative soil quality evaluation method. *Soil Science Society of America Journal*, 68(6), 1945-1962.
19. Singh, S.; Bhattarai, R.; Negm, L.M.; Youssef, M.A.; and Pittelkow, C.M. (2020). Evaluation of nitrogen loss reduction strategies using DRAINMOD-DSSAT in east-central Illinois. Agricultural Water Management, 240, 106322
20. Huang, Y.; Srivastava, R.; Ngo, C.; Gao, J.; Wu, J.; and Chiao, S. (2023). Data-driven soil analysis and evaluation for smart farming using machine learning approaches. *Agriculture*, 13(9), 1777.

21. Lundberg, S.M.; and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the* 31st *International Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, California, USA, 4768-4777.
22. Lehmann, J.; and Kleber, M. (2015). The contentious nature of soil organic matter. *Nature*, 528(7580), 60-68.
23. EI Behairy, R.A.; EI Arwash, H.M.; EI Baroudy, A.A.; Ibrahim, M.M.; Mohamed; E.S.; Rebouh, N.Y.; and Shokr, M.S. (2024). An accurate approach for predicting soil quality based on machine learning in drylands. *Agronomy*, 14(4), 627.
24. Xie, J.; Shi, C.; Liu, Y.; Wang, Q.; Zhong, Z.; He, S.; and Wang, X. (2025). Soil salinization prediction through feature selection and machine learning at the irrigation district scale. *Frontiers in Earth Science*, 12, 1488504.
25. Zhao, Z.-D.; Zhao, M.-S.; Lu, H.-L.; Wang, S.-H.; and Lu, Y.-Y. (2023). Digital mapping of soil pH based on machine learning combined with feature selection methods in East China. *Sustainability*, 15(17), 12874.
26. Padarian, J.; McBratney, A.B.; and Minasny, B. (2020). Game theory interpretation of digital soil mapping convolutional neural networks. *SOIL*, 6(2), 389-397.
27. Näsi, R.; Mikkola, H.; Honkavaara, E.; Koivumäki, N.; Oliveira, R.A.; Peltonen-Sainio, P.; Keijälä, N.-S.; Änäkkälä, M.; Arkkola, L.; and Alakukku, L. (2023). Can basic soil quality indicators and topography explain the spatial variability in agricultural fields observed from drone orthomosaics? *Agronomy*, 13(3), 669.
28. Sihi, D.; Dari, B.; Kuruvila, A.P.; Jha, G.; and Basu, K. (2022). Explainable machine learning approach quantified the long-term (1981-2015) impact of climate and soil properties on yields of major agricultural crops across CONUS. *Frontiers in Sustainable Food Systems*, 6, 847892.