Journal of Environmental Management 370 (2024) 122640

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Research article

Evaluating machine learning performance in predicting sodium adsorption ratio for sustainable soil-water management in the eastern Mediterranean

Safwan Mohammed [a,b,*], Sana Arshad [c], Bashar Bashir [d], Behnam Ata [e], Main Al-Dalahmeh [f], Abdullah Alsalman [d], Haidar Ali [g], Sami Alhennawi [g], Samer Kiwan [g], Endre Harsanyi [a,b]

[a] Institute of Land Use, Technical and Precision Technology, Faculty of Agricultural and Food Sciences and Environmental Management, University of Debrecen, 4032, Debrecen, Hungary
[b] Institutes for Agricultural Research and Educational Farm, University of Debrecen, Böszörményi 138, 4032, Debrecen, Hungary
[c] Department of Geography, The Islamia University of Bahawalpur, Bahawalpur, 63100, Pakistan
[d] Department of Civil Engineering, College of Engineering, King Saud University, P.O.Box 800, Riyadh, 11421, Saudi Arabia
[e] Department of Social Geography and Regional Development Planning, University of Debrecen, H-4032, Debrecen, Hungary
[f] Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman, Jordan
[g] Department of Natural Resources Research, General Commission for Scientific Agricultural Research (GCSAR), Damascus, Syria

ARTICLE INFO

Keywords:
Sodium adsorption ratio (SAR)
Recursive feature elimination (RFE)
Nu support vector regression (NuSVR)
SHapley additive exPlanation (SHAP)

ABSTRACT

Soil salinization is a critical global issue for sustainable agriculture, impacting crop yields and posing a threat to achieving the Sustainable Development Goal (SDG) of ensuring food security. It is necessary to monitor it in detail and uncover its underlying factors at a regional scale. In this context, the present study aimed to evaluate soil health in the eastern Mediterranean region by using the Sodium Adsorption Ratio (SAR) as an indicator of soil salinity in three distinct soil horizons. The main objective of the research was to evaluate the performance of four machine learning (ML) models, including Random Forest (RF), Nu Support Vector Regression (NuSVR), Artificial Neural Network-Multi Layer Perceptron (ANN-MLP), and Gradient Boosting Regression (GBR), for accurate prediction of SAR following the Recursive Feature Elimination (RFE) as a feature selection method. Moreover, SHapely Additive exPlanations (SHAP) was applied as sensitivity analysis to identify the most influential covariates. Main findings of the research revealed that the average clay content in the surface horizon ($H_{10\text{-}25cm}$) was 50.5% ± 10.4, which significantly increased to 57.5% ± 8.7 ($p < 0.05$). No significant mean differences were detected between the studied horizons for SAR and Na^{+}. ML output revealed that NuSVR outperformed other algorithms in accurately predicting outcomes during both the training and testing stages. Moreover, Scenario 2 (SC2) with seven selected features from the RFE method facilitated highly accurate SAR predictions. Overall, the performance of ML models is ranked as NuSVR > GBR > ANN-MLP > RF. Lastly, SHAP sensitivity analysis identified CEC, Ca^{+2}, Mg^{+2}, and Na^{+} as the most influential variables for SAR prediction in both the training and testing stages. Hence, the research yielded valuable insights for efficient agricultural soil management at a regional level using state-of-the-art technology.

1. Introduction

Soil salinity is one of the most significant forms of land degradation, affecting the sustainability of agricultural production and food security (Sobhi Gollo et al., 2023; Wuyun et al., 2022). Salinization occurs when soil accumulates excessive soluble salts and comprises alkaline, sodic, and saline soils related to high pH, sodium (Na^{+}), and salts (Klopp and Bleam, 2021). This can occur naturally or primarily through rainfall depositing oceanic salts, fluvial and aeolian deposition, and rock weathering, which is further facilitated by the streamflow or evapotranspiration of underground water (Hassani et al., 2021). Furthermore, geological structures, karst processes, mineral composition, and topographical features also aid in defining salinity patterns (Gorji et al., 2017; Metternicht and Zinck, 2003). Additionally, soil characteristics, such as porosity, structure, texture, clay minerals, compaction and infiltration rate, water storage capacity, saturated and unsaturated

* Corresponding author. Institute of Land Use, Technical and Precision Technology, Faculty of Agricultural and Food Sciences and Environmental Management, University of Debrecen, 4032, Debrecen, Hungary.
E-mail address: safwan@agr.unideb.hu (S. Mohammed).

https://doi.org/10.1016/j.jenvman.2024.122640
Received 9 February 2024; Received in revised form 2 August 2024; Accepted 21 September 2024

hydraulic conductivity, and potential salt content also play a role in this influence (Daliakopoulos et al., 2016; Pankova et al., 2015). In secondary soil salinization, anthropogenic activities such as brackish water irrigation, poor land management practices, seawater intrusion in coastal lands, and excessive fertilization lead to the accumulation of water-soluble salts in various soil horizons (Gorji et al., 2017; Mandal, 2019; Tedeschi, 2020). Moreover, flooding, over-irrigation, seepage, silting, and rising water tables accelerate these problems (Eswar et al., 2021; Maliva, 2021; Wang et al., 2023a).

Soil salinization is a natural phenomenon, affecting approximately 930 million hectares of saline/sodic soils globally (Rengasamy, 2006). Recently, Hopmans et al. (2021) reported that the salt affected soils cover 1 billion hectares, representing 7% of the total land area on the Earth's land surface. Interestingly, 30% of the irrigated lands globally are salt-affected due to secondary human-induced salinization. On the other hand, Núñez and Finkbeiner (2020) calculated the average soil susceptibility to salt to be 0.19 dS/m per gram of salt in 1 m^3, resulting in crop loss of 5.7×10^{-2} per gram of salt in 1 m^3 of soil. Parallel to salinity, excessive concentration of Na^+ ions compared to other exchangeable and soluble cations such as Ca^{2+} and Mg^{2+} defines the concept of soil sodicity (Gharaibeh et al., 2021). The Exchangeable Sodium Percentage (ESP) and Sodium Adsorption Ratio (SAR) are the most common expressions of soil sodicity and determine the suitability of soil or water for agricultural production (Daliakopoulos et al., 2016; Gautam et al., 2023). In sodic soils, high Na^+ concentrations displace cations such as Ca^{2+} and $Mg^{2+,}$ causing significant structural degradation. As exchangeable sodium hydrolyzes, soil particles weaken and detach, leading to increased dispersibility and susceptibility to erosion (Abd El-Halim et al., 2023; de la Paix et al., 2013). Reports on the global extent of soil salinization vary, with estimates ranging from 10% arable land to one billion hectares covered with saline sodic soils. (Shahid et al., 2018). In addition, secondary salinization renders 3 ha per minute unproductive, leading to an annual loss of 10–20 Mha per year (Cuevas et al., 2019).

Several countries have been identified as hotspots of soil salinity, including China, India, and Pakistan from South Asia, the United States, and a few Western and Central Asian countries (AquaStat, 2022). The Mediterranean region of the European continent is also reported to be vulnerable to extreme salinity and land degradation (Abu Hammad and Tumeizi, 2012; Daliakopoulos et al., 2016). Moreover, salt-affected soils are mostly persistent in arid and semiarid ecosystems due to less rainfall and a high rate of evapotranspiration, causing a decline in crop production (Shaaban et al., 2023; Singh, 2021). This is because excessive salt accumulation in the soil negatively affects plant growth and development, leading to physiological abnormalities and crop loss. Several studies have reported the impact of saline and sodic soils on decreased crop production (Li et al., 2023; Wang et al., 2016, 2023a). For instance, a recent study by Kafei et al. (2023) provided a quantitative evaluation of changes in wheat grain yield associated with saline sodic soil in the Mediterranean sub-humid region. Hence, it has become a major threat to global food security. Therefore, predicting soil salinity on global and regional scales is crucial for improving soil management to ensure crop production (Hassani et al., 2021; Zhou et al., 2022).

Spatiotemporal and vertical (soil horizon) variability in sodicity levels provides a clearer understanding of the impacts of climate change and terrestrial carbon dynamics (Hartemink et al., 2020; Hassani et al., 2021; Mohammed et al., 2020b). Projected changes in climate change and altered hydrological balance increase the risk of soil problems and land degradation (Liu et al., 2022; Okur and Örçen, 2020). For instance, Tomaz et al. (2020) demonstrated the vulnerability of the Mediterranean region to climate-induced soil problems. Several salt and sodium control measures have been proven to reduce saline-sodic problems. For example, adding an appropriate amount of gypsum to the soil has been reported to reduce soil electrical conductivity (EC), pH, and SAR, thereby enhancing its fertility (Shaaban et al., 2023). Moreover, the use of halophytes through inter-and sequential cropping techniques has also been reported to control EC and SAR to improve soil performance (Jurado et al., 2024; Navarro-Torre et al., 2023). It will support to combating land degradation in a sustainable way to achieve sustainable development goals (UN-SDGs, 2030) mainly SDG-2 (zero hunger) (Mohammed et al., 2020b).

In the era of big data analytics, artificial intelligence and machine learning (ML) have been widely applied as decision-making tools to provide optimal solutions for complex environmental and agricultural problems (Arshad et al., 2023b; Mohammed et al., 2023; Singh et al., 2022). ML was used in different aspects of research such as water quality (Mohammed et al., 2024b; Sajib et al., 2024; Uddin et al., 2023), agricultural research (Arshad et al., 2023b; Mohammed et al., 2024a), human health (Uddin et al., 2024), and many other disciplines. Numerous studies have employed data-driven modeling approaches to map and predict soil sodicity and salinity (Gautam et al., 2023; Wang et al., 2023b). For example, in a recent study in northwest China, Xiao et al. (2023) predicted soil salinity using three machine-learning (ML) algorithms. Similarly, Abedi et al. (2021) employed six ML algorithms to predict and model some salinity indicators, concluding that random forest (RF) is the best algorithm for predicting SAR. A bibliographic cluster analysis presentation for the keywords '(soil), (salinity)' and 'machine learning' (Fig. S1) revealed that significant research has been conducted in this domain, where more focus in using ML was recently adopted. However, the accurate prediction of soil salinity is dependent on several factors, including sample size, appropriate feature selection, and the selection of machine learning algorithms (Wang et al., 2023b). For instance, Andrade Foronda and Colinet (2023) utilized partial least squares (PLS), Support Vector Machines (SVM), and RF to predict soil sodicity from various cations and anions. Another recent study by Abba et al. (2023) resulted in the ANFIS-PSO metaheuristic algorithm with the highest accuracy (99%) and genetic algorithm feature selection with Ca^{2+}, Mg^{2+}, Na^+, and Cl^- as the main influential variables of groundwater salinization. Several other studies on SAR prediction in Mediterranean regions, including Morrocco, Turkey, and Jordan, have evaluated the performance of machine learning algorithms (El Bilali et al., 2021; Gharaibeh et al., 2021; Sattari et al., 2020).

Syria is one of the country most vulnerable to secondary salinization processes in the eastern Mediterranean (Kamrakji et al., 2016). Salinization is one of the land degradation aspects in Syrian soil, with 532,000 ha of salt-affected soil. There is limited available information on the local drivers of salinity (Choukr-Allah et al., 2023). Furthermore, a thorough examination of studies predicting soil salinity and sodicity identified a research gap in this area. The region-specific machine learning modeling approach has proven to be indispensable for an in-depth understanding of soil problems and informed decision-making processes (Das et al., 2023; Jamei et al., 2024). Furthermore, implementation of machine learning to predict SAR in Syria will provide researchers and decision-makers with appropriate tools for land rehabilitation and salinization assessment in the study area. Addressing this gap, this research aims to: 1) explore some soil characteristics in the eastern Mediterranean (southern Syria) employing Tukey mean difference test and Principal Component Analysis (PCA); 2) explore the five iteratively selected input combinations or scenarios from SVM based Recursive Feature Elimination (RFE) for accurate SAR predictions in ML models 3) evaluate the performance of four competitive ML models including Random Forest (RF), Nu Support Vector Regression (NuSVR), Artificial Neural Network-Multi Layer Perceptron (ANN-MLP), and Gradient Boosting Regression (GBR), and 4) examine the impact of individual predictors on highly accurate SAR predictions through SHapley Additive exPlanation (SHAP) method.

2. Material and method

2.1. Study area and data collection

Soil samples were collected from 107 representative soil profiles, in

the eastern Mediterranean (southern Syria) (Mohammed et al., 2020a) (Fig. 1b). Data collection and physicochemical soil properties were analyzed and explained by Mohammed et al. (2020a). Result of filed survey, laboratory analysis, soil mineralogy, and soil classification are freely available in: https://www.sciencedirect.com/science/article/pii/S2352340920307265. The common soil orders were Aridisols, Inceptisols, Mollisols, Entisols, and Vertisols (Fig. 1c). The predominant soil textures in the study area included clay, silty, sandy loam, and clay loam. For this research, the collected data encompasses clay (%), pH (H_2O), EC (dsm^{-1}), total organic matter (TOM) (%), cation exchange capacity (CEC) ($Cmol_c\ kg^{-1}$), Ca^{2+} ($Cmol_c\ kg^{-1}$), Mg^{2+} ($Cmol_c\ kg^{-1}$), Na^{+} ($Cmol_c\ kg^{-1}$), and K^{+}($Cmol_c\ kg^{-1}$). To facilitate data handling and analysis, data were organized into three horizons: the first group included soil samples collected from to 0–25 cm, referred to as H1 $_{0\text{-}25}$; the second group from to 25–60 cm depth, referred to as (H2 $_{25\text{-}60}$); and the third group included data collected deeper than 60 cm, referred to as (H3 $_{<60}$).

2.2. Sodium-adsorption ratio (SAR)

The sodium adsorption ratio (SAR) gauges the quantity of sodium (Na^{+}) compared to calcium (Ca^{2+}) and magnesium (Mg^{2+}) in the water extract from saturated soil paste. It was calculated by dividing the Na^{+} concentration by the square root of half the sum of the Ca^{2+} and Mg^{2+} concentrations (Daliakopoulos et al., 2016). Soils with SAR values exceeding 13 might exhibit heightened dispersion of organic matter and clay particles, lowered saturated hydraulic conductivity (Ksat) and aeration, and a decline in the overall soil structure (Salvato et al., 2024).

In the next step, the SAR value was calculated using Equation (1):

$$SAR = \frac{Na^{+}}{\sqrt{\frac{Ca^{2+}+Mg^{2+}}{2}}} \quad (1)$$

2.3. Statistical analysis

To determine whether there were significant differences between the means of the three studied groups (i.e., H1$_{0\text{-}25}$, H2$_{25\text{-}60}$, H3$_{<60}$) for each soil characteristic, a post hoc Tukey test was implemented (Noguchi et al., 2020). Tukey devised an Honestly Significant Difference (HSD) test for easy pairwise comparisons. It calculates the significant difference between means using the student's q distribution, indicating the maximum difference from a set of means from the same population. All differences were measured against this distribution, making the HSD method conservative in its assessment (Richardson et al., 2021).

In addition, the multivariate statistical technique, namely Principal Component Analysis (PCA), was used to summarize soil characteristics from various groups into a few Principal Components (PCs) accounting for the majority of the variance explained (Wold et al., 1987). PCA, a multivariate statistical analysis method, is non-parametric and is extensively used in environmental science and climate studies (Islam Khan et al., 2022). PCA is commonly used to analyze complex datasets. Usually, three-dimensional visualizations are employed, or data are plotted based on their correlation with the components. However, because two or three dimensions may lose significant information, it is crucial to methodically test various component combinations when visualizing a dataset (Abdullah et al., 2019).

2.4. Machine learning application for SAR prediction

2.4.1. Feature selection method (Recursive feature elimination (RFE))

For accurate prediction modeling, two types of dimension reduction methods are commonly applied. These include feature selection and feature extraction. Feature extraction transforms the original data into a new set of uncorrelated features such as principal components. Feature selection identifies the most relevant set of features from the entire

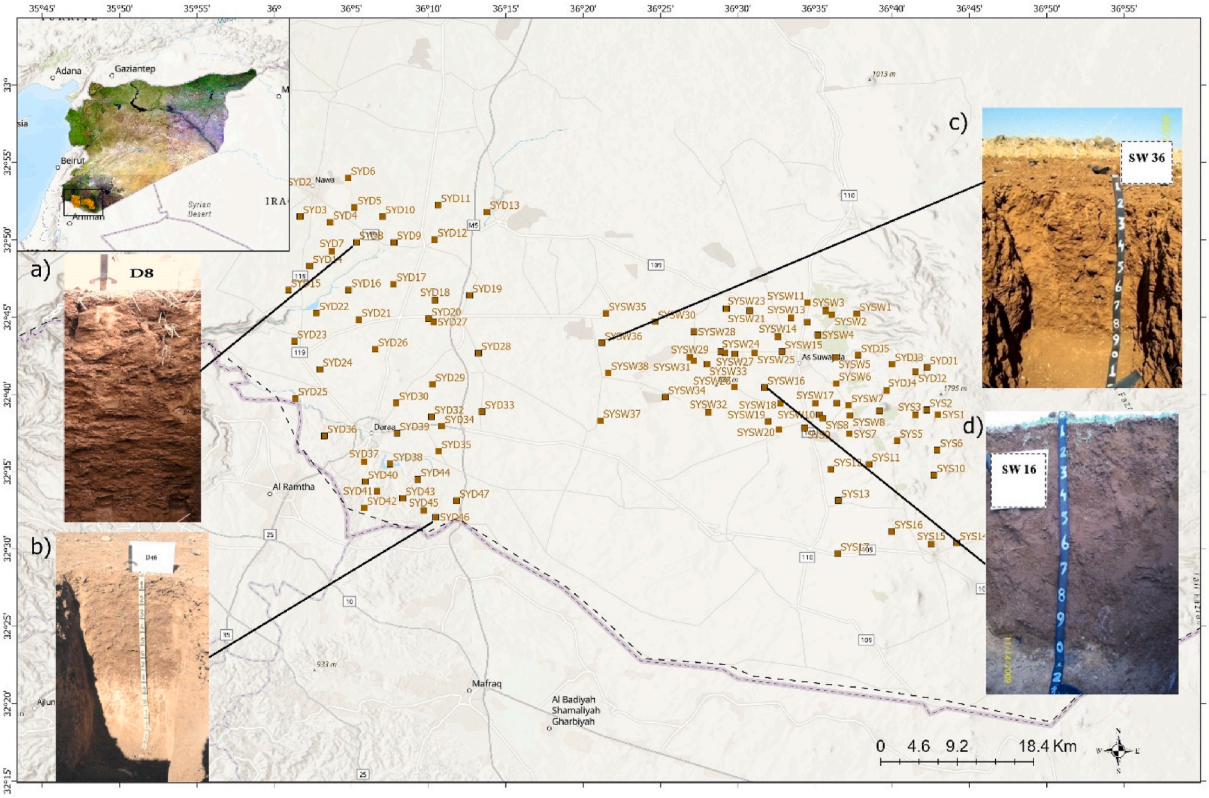

Fig. 1. Soil sampling locations (southern Syria) (Sentinel-2 RGB) with examples of soil profiles and their classification within the study area. a) D8: Vertic Haploxerepts (36.0889 _ 32.8306); b) D46: Typic Haplocalcids (36.1744 _ 32.5344), c) SW36: Chromic Haploxererts (36.3536 _ 32.7222); d) SW16: Chromic Haploxererts (36.529 _ 32.6742).

dataset that are closer to accurate predictions. These include filter, wrapper, and embedded methods (Otchere et al., 2022).

Currently, we employ wrapper-based Recursive Feature Elimination (RFE), which performs a rigorous search for more relevant variables in each iteration using a machine learning algorithm. We used a support vector machine (SVM) with REF to determine the optimum subset of predictors. The weights corresponding to the hyperplane in the SVM algorithm served as indicators of the magnitude of importance. In the RFE algorithm, the weights represent the relative significance of the predictors. A larger weight indicates a higher level of importance for the respective predictors (Khan et al., 2020). In our case 9 potential predictors (clay, pH, EC, TOM, CEC, Ca^{2+}, Mg^{2+}, Na^{+}, and K^{+}) are input in SVM-RFE process. The model was trained with a linear kernel and ran five times with n features = 8, 7, 6, 5, and 4, eliminating the less important features in each iteration.

The subset selection from SVR-RFE provided a set of input combinations, as depicted in Table 1.

2.4.2. Machine learning algorithms

The transformed feature subsets are further taken as input combinations in four competitive ML algorithms, RF, NuSVR, ANN-MLP, and GBR, for predicting SAR with a random split of 75% train and 25% test. We preferred to employ a diversified set of ML algorithms for more reliable predictions. Specifically, RF was chosen due to its ensemble nature and effective performance in high dimensional space. Flexible regularization and accurate non-linear relationship capturing capability of NuSVR made it a good choice. Furthermore, ANN-MLP provided a deep learning experience to the data for achieving high prediction accuracy. Lastly, GBR is a strengthened model for reducing variance and bias and enhance prediction accuracy using boosting techniques.

Furthermore, for more robustness and high prediction accuracy, optimal selection of hyperparameters is achieved through the grid search method and five-fold cross-validation (cv = 5) in model training. The cross-validation approach involved creating five subsets from the training data, with each subset having the same distribution of target variable. In each iteration, four subsets are used to train the model and the fifth one is used validation. Moreover, Grid search is a systematic approach that explores to identify the best hyperparameter combination for optimal model performance. We run all models in Python 3 environment, using the T4GPU in 32 GB RAM for robust model's performance. A detail of hyperparameter configuration for each applied ML algorithm is presented in Table 2. Furthermore, applied ML algorithms are briefly explained below.

2.4.2.1. Random forest. Random Forest (RF) is a widely applied non-parametric ensemble machine learning algorithm based on the principle of bagging or bootstrap aggregation. The algorithm works by constructing multiple decision trees randomly sampled from rows and features, following the bootstrap sampling approach. Hence, each tree was trained on a different subset to reduce overfitting. Finally, the weighted average of the trees is utilized to obtain the final prediction (Breiman, 2001). Moreover, the techniques combined with randomized node optimization increase the efficiency of the model in handling complex high-dimensional data and provide insight into feature importance. This improves the generalization of the algorithm and increases its accuracy for prediction modeling. Currently, n estimators are set to 10, 50, and 100 with maximum depth of trees set as none, 10, 20, and 30. The detail of hyperparameter search applied is presented in Table 2. Recently, random forests have been used to predict soil and water quality parameters (Kushwaha et al., 2023; Xiao et al., 2023).

Table 1
Scenarios (input combinations) developed from SVR-RFE method for SAR prediction.

Scenario	Input combination	ML-model	Output
SC1 (8F)	pH, EC, TOM, CEC, Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}	RF, NuSVR, ANN-MLP, GBR	SAR
SC2 (7F)	pH, EC, CEC, Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}	RF, NuSVR, ANN-MLP, GBR	SAR
SC3 (6F)	pH, EC, CEC, Ca^{2+}, Na^{+}, K^{+}	RF, NuSVR, ANN-MLP, GBR	SAR
SC4 (5F)	pH, EC, Ca^{2+}, Na^{+}, K^{+}	RF, NuSVR, ANN-MLP, GBR	SAR
SC5 (4F)	pH, EC, Na^{+}, K^{+}	RF, NuSVR, ANN-MLP, GBR	SAR

Table 2
Hyperparameter optimization in Grid search for applied ML algorithms.

ML algorithm	Hyperparameter search range
Random Forest	Number of trees 'n_estimators': [10, 50, 100], Max depth of trees 'max_depth': [None, 10, 20, 30], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4]
NuSVR	Nu SV 'nu': [0.1, 0.3], Regularization 'C': [0.1,10], 'kernel': ['linear','rbf'], 'gamma': [0.1, 0.01],
ANN-MLP	'hidden_layer_sizes': [(50, 50), (100, 100), (100, 50, 25)], 'activation': ['relu', 'tanh'], 'alpha': [0.0001, 0.001, 0.01], 'learning_rate': ['constant', 'invscaling', 'adaptive'], 'learning_rate_init': [0.001, 0.01, 0.1],
GBR	'n_estimators': [50, 100], 'learning_rate': [0.01, 0.1], 'max_depth': [3, 5], 'min_samples_split': [2, 5], 'min_samples_leaf': [1, 2], 'subsample': [0.8, 0.9]

2.4.2.2. Nu Support Vector Regression. The (SVM) is based on several kernels that transform low-dimensional input into a high-dimensional feature space separated by a hyperplane (Arshad et al., 2023a). Support Vector Regression is a variant of SVM based on the principles of statistical learning theory (Vapnik, 1997). Two primary SVM regression versions, epsilon-SVR and nu-SVR, differ in their approaches to margin control and penalty parameters. In epsilon-SVR, there is no regulation on the number of data vectors from the dataset that become support vectors. The regularization parameter C governs the errors allowed in the model. Nu versions of SVM are preferred because of their more meaningful interpretation: 'nu' signifies an upper limit on the fraction of training samples considered as errors and a lower limit on the fraction of samples classified as support vectors (Langhammer and Česák, 2016; Schölkopf et al., 1998). Mathematically, presented as

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{w}^{T}.\,\boldsymbol{\Phi}\,(\mathbf{x}) + \boldsymbol{b} \tag{2}$$

Where $f(x)$ is the predicted output of SAR, w^T is the transpose of weight vector containing weights assigned to each feature, $\Phi\,(\mathrm{x})$ is the nonlinear mapping function which transforms the original input feature to high dimensional space, and b is the bias term (Bhatt et al., 2012). Parameter tuning, which is essential for NuSVR like other SVR methods, involves selecting the best regularization constant (C) and maximum deviation (ε), which can be time consuming and computationally demanding (Schölkopf et al., 1998). Currently, optimal hyperparameter are searched by nu = 0.1, 0.3, c = 0.1, 10, with kernel choices of linear and rbf (Table 2).

2.4.2.3. ANN-MLP. The (ANN) is a soft computational technique designed for the structure of a human brain neuron network with complex input and output interactions. Multilayer perceptron (MLP) is the most commonly applied feed-forward neural network with an input, one or more hidden, and output layers (Hornik et al., 1989). The ability of MLP to capture intricate and high-dimensional relationships makes it a reliable algorithm for several prediction studies (Jamei et al., 2024;

Kan et al., 2023). The MLP architecture is characterized by back-propagation and a nonlinear activation function that provides robust solutions to complex problems that minimize the chance of overfitting (Elsherbiny et al., 2021; Shadkani et al., 2021). Every neuron within the network receives inputs from neurons in the preceding layer. It computes a weighted sum of these inputs, incorporates a bias term, and subsequently applies an activation function. In mathematical terms, the output or activation of a neuron y can be formulated as

$$y = f\left(b + \sum_{i=1}^{n} x_i w_i\right) \quad (3)$$

Where f denotes the activation function and b is the bias term, w_i is the weight associated with the input feature. The activation function is meant to introduce non-linearity which enhances the architecture capability to capture intricate relationships. Currently, in grid search optimization, we evaluated the performance of ANN-MLP using Rectified Linear Unit (ReLU) and hyperbolic tangent (tanh) activation functions, coupled with hidden layers set as (50, 10), (100, 100), and (100, 50, 25) (Table 2).

2.4.2.4. GBR. Gradient Boosting regression is another powerful ensemble ML algorithm which works by sequentially integrating weak learners, that are typically decision trees, iteratively to achieve accurate predictions (Friedman, 2002). It is based on the utilization of gradient descent optimization, which computes the gradient of loss in each iteration, whereas the gradient reflects the adjustments needed to minimize the prediction errors (Otchere et al., 2022). It is presented by.

$$F_M(x) = \sum_{m}^{M} \alpha_m h_m(x_1, x_2, \ldots\ldots x_n) \quad (4)$$

where $F_M(x)$ is the final predictive model of SAR, h_m is the weak learner and α_m is the learning rate. Currently, n estimators in grid search are set as 50, 100 with learning rate = 0.01, and 0.1 (Table 2).

2.4.3. Machine learning prediction performance evaluation

The machine learning performance was evaluated, and the coefficient of determination R^2, Root Mean Square Error (RMSE), Mean Squared Error (MSE), and mean absolute percentage error (MAPE) were calculated (Table 3). R^2 is one of the indicators used to illustrate the goodness of fit between observed and predicted data (Piepho, 2023). R^2 ranges between 0 and 1. The higher the R^2 value, the better the model performance. RMSE is used for normal (Gaussian) errors (Hodson, 2022), representing the square root of the average of the squared errors. Lower RMSE values are an indicator of good model performance. The essence of MSE relies on calculating the average squared difference between predicted values by machine learning (ML) and observed SAR values (Li et al., 2022). MAPE is used to calculate the absolute error divided by the observed values of SAR, and then averages them. It was reported that the MAPE is biased towards low predicted values (Tofallis, 2015), however, it remains one of the main indicators for evaluating the ML performance. The PBIAS value near to 0 indicates a good match in observed and predicted values, greater than 0 indicates underestimation and less than 0 (negative PBIAS) indicated overestimation. Moreover, model efficiency is evaluated using Nash Sutcliffe efficiency (NSE), range from $-\infty$ to 1, with value closer to 1 indicated the very good match between observed and predicted values (Sajib et al., 2024).

Table 3
Indicators used for evaluating the performance of ML in predicting SAR value.

Indicator	Equation	Equation number	Range	Best fit
R^2	$R^2 = 1 - \frac{\sum_{i=1}^{n}(SAR_{pred} - \overline{SAR}_{obs})^2}{\sum_{i=1}^{n}(SAR_{obs} - \overline{SAR}_{obs})^2}$	(5)	0_+1	+1
RMSE	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(SAR_{obs} - \widehat{SAR}_{pred})^2}$	(6)	0_+∞	0
MSE	$MSE = \frac{1}{n}\sum_{i=1}^{n}(SAR_{obs} - \widehat{SAR}_{pred})^2$	(7)	0_+∞	0
MAPE	$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left\|\frac{SAR_{obs} - SAR_{pred}}{SAR_{obs}}\right\|$	(8)	0_+∞	0
PBIAS	$PBIAS = \left(\frac{\sum_{i=1}^{n}(SAR_{obs} - \widehat{SAR}_{pred})}{SAR_{obs}}\right)$	(9)	±∞	0
NSE	$NSE = 1 - \frac{\sum_{i=1}^{n}(SAR_{obs} - \widehat{SAR}_{pred})^2}{\sum_{i=1}^{n}(SAR_{obs} - \overline{SAR}_{obs})^2}$	(10)	−∞ to 1	1

2.4.4. SHAP for ML interpretation

SHapely Additive exPlanations (SHAP) is an efficient ML method that provides an explanation for the output of the ML model (Descals et al., 2023). The background of SHAP is the utilization of game theory to determine the significant contribution of each model input and to identify the bias in the trained model (Bogdanova et al., 2023). Positive and negative SHAP values determine the influence of the input variables on the predicted output. Hence, the SHAP model provides an interpretable explanation for the selected ML algorithm in various studies (Chandra Joshi et al., 2023; Zhang et al., 2023). The output model produces a SHAP value $\hat{y}_i$ for each predicted observation and y_{base} is the baseline of model. Mathematically presented by

$$\hat{y}_i = y_{base} + f(x_{i1}) + f(x_{i2}) + \ldots f(x_{in}) \quad (5)$$

If the SHAP value is greater than 0, then the relevant feature improves the final prediction output with high impact; otherwise, it decreases the prediction. It is advantageous to use in explaining the influence or impact of each variable on the final predictions (Wang et al., 2023b). Currently, the non-tree SHAP 'kernelExplainer' is employed as a sensitivity analysis to interpret highly accurate model performance.

3. Results

3.1. An Overview of the soil characteristics in the study area

3.1.1. Statistical exploration of soil properties in the eastern mediterranean

Descriptive statistics of soil properties (Table S1) showed that in the top horizon ($H1_{0\text{-}25}$), the average clay content was 50.40% ± 10.9, which increased to 57.55% ± 8.6 in the third soil horizon ($H3_{<60}$). pH, a significant threshold of soil alkalinity, was found to have a minimum value of 6.0, and a maximum of 8.22 with a mean of 7.55 ± 0.44 $H1_{0\text{-}25}$. Furthermore, EC ranged between minimum 0.04 dsm^{-1} (non-saline) to maximum 2.13 dsm^{-1} (moderately saline) in H1, and 0.01 dsm^{-1} (non-saline) to maximum 3.1 dsm^{-1} (high saline) in H3 (Fig. S2). Similarly, the percentage age of TOM ranged (2.17%) between a minimum of 0.39% and a maximum of 2.56% in H1 and 0.11%–1.67% in H3. Descriptive analysis of SAR determining cations including Ca^{2+} is ranged between minimum 10.3 $Cmol_c\ kg^{-1}$ to maximum 39 $Cmol_c\ kg^{-1}$ in H1, 12 $Cmol_c\ kg^{-1}$ to 40 Cmolc kg^{-1} in H2 and H3 in respective representative samples.

Similarly, Mg^{2+} is found to have a minimum value of 3.2 $Cmol_c\ kg^{-1}$ and maximum 14.3 $Cmol_c\ kg^{-1}$ in H1, and 6.2 $Cmol_c\ kg^{-1}$ and 17.7 $Cmol_c\ kg^{-1}$ in H2. Na + ions reflecting the dominant sodicity characteristics of soil samples are found to range from minimum 0.30 $Cmol_c\ kg^{-1}$ to 2.63 $Cmol_c\ kg^{-1}$ in H1, 0.30 $Cmol_c\ kg^{-1}$ to 2.7 $Cmol_c\ kg^{-1}$ $Cmol_c\ kg^{-1}$ in H2 and maximum 2.69 $Cmol_c\ kg^{-1}$ in H3. CEC is found to range from minimum 20.2 $Cmol_c\ kg^{-1}$ to maximum 56.1 $Cmol_c\ kg^{-1}$ in H1, 24.5 $Cmol_c\ kg^{-1}$ to maximum 56.1 $Cmol_c\ kg^{-1}$ in H2, and 27.7 $Cmol_c\ kg^{-1}$ to 56.2 $Cmol_c\ kg^{-1}$ in H3 (Fig. S2). Furthermore, the descriptive analysis of SAR showed a minimum 0.06 in H1, and 0.07 in H2 and H3. The maximum value of

SAR was found to be 0.64 in H1, 0.65 in H2, and 0.60 in H3 (Fig. S3).

The Tukey test indicated a significant ($p < 0.05$) difference between the clay contents across the three studied horizons (Fig. 2 (a)). The average pH ranged between 7.5 ± 0.4 and 7.6 ± 0.3; however, a significant ($p < 0.05$) difference was captured between H1 0–25 and H3 <60 (Fig. 2 (b)). Within the three studied horizons, the average EC did not change significantly and remained within a range of 0.3 dsm-1 (Fig. 2 (c)). The highest mean TOM value was observed in the topsoil horizons, measuring 1.11% ± 0.42. In the subsequent layer (H2 25–60), the average TOM value decreased significantly ($p < 0.05$) to 0.63% ± 0.29, and further decreased to 0.39% ± 0.28 in H3 <60 (Fig. 2 (d)).

The average of CEC was 39.38 ± 7.89 Cmolc kg^{-1} in the first horizon ($H1_{0\text{-}25}$), which significantly increased to 44.15 ± 5.90 Cmolc kg^{-1} in the H3 (Fig. 2 (e)). Moreover, the average Ca^{2+} was 23.17 ± 5.86 in H1 and significantly increased to 26.61 ± 5.05 in H3 (Fig. 2 (f)). Similarly, mean Mg^{2+} ions also increased significantly from 9.96 ± 1.92 in H1 to 11.62 ± 2.10 in H3 (Fig. 2 (g)).In contrast, Na^{+} concentration remained relatively consistent across the soil horizons, with an average of 1.214 ± 0.54 H3 (Fig. 2 (h)). Similarly, the SAR values did not exhibit any significant differences among the three studied layers, where the average ranged between 0.29 and 0.28 (Fig. 2 (j)).

3.1.2. Principal Component Analysis of soil properties

PCA analysis of all soil properties from representative samples revealed a high dimensionality in the data distribution with three principal components (PCs) above an eigenvalue of 1 with a cumulative percentage of 74.6% (Table S2). The biplot presentation (Fig. 2 (k)) showed that PC1 accounted for 34.13% of the total variance, whereas PC2 accounted for 27.97%. Consequently, both PC1 and PC2 accounted for 62.1% of the total variance. Three clusters were identified. The first cluster included CEC, Ca^{2+}, Mg^{2+}, and clay, which were highly correlated with PC1 with high positive loadings of 0.42, 0.47, 0.24, and 0.34, respectively. The second cluster, which had a strong correlation with PC2, consisted of SAR, Na^{+}, EC, and pH with high positive loadings of 0.30, 0.38, 0.25, and 0.35 (Table S3). In contrast, TOM stands apart from these groupings with a negative correlation with PC1 and the highest coefficient of 0.69 in PC3 (Table S3).

3.2. Assessment of ML algorithms in predicting SAR values

3.2.1. ML algorithms performance in training stage

The performance of the machine learning model in the training stage for five different scenarios (input combinations) derived from RFE showed that SC1 with eight input features performed with the highest accuracy in the NuSVR algorithm with the highest $R^2 = 0.997$, lowest RMSE = 0.008, MSE = 0.00007, and MAPE = 2.10%. Followed by SC2 with seven input features ($R^2 = 0.996$, RMSE = 0.010, MSE = 0.00009, and MAPE = 2.36%). Then SC3 with six input features ($R^2 = 0.995$, RMSE = 0.010, MSE = 0.0001, and MAPE = 2.32%), SC4 with five input features ($R^2 = 0.994$, RMSE = 0.011, MSE = 0.0001, and MAPE = 3.11%), and SC5 with four input features (Fig S4, Fig S5). This was followed by GBR in SC1 (eight input features) and SC2 (seven input features), with $R^2 = 0.990$, RMSE = 0.015, MSE = 0.0002, and MAPE = 3.61% and 3.59%, respectively. GBR for SC3 with six input features was

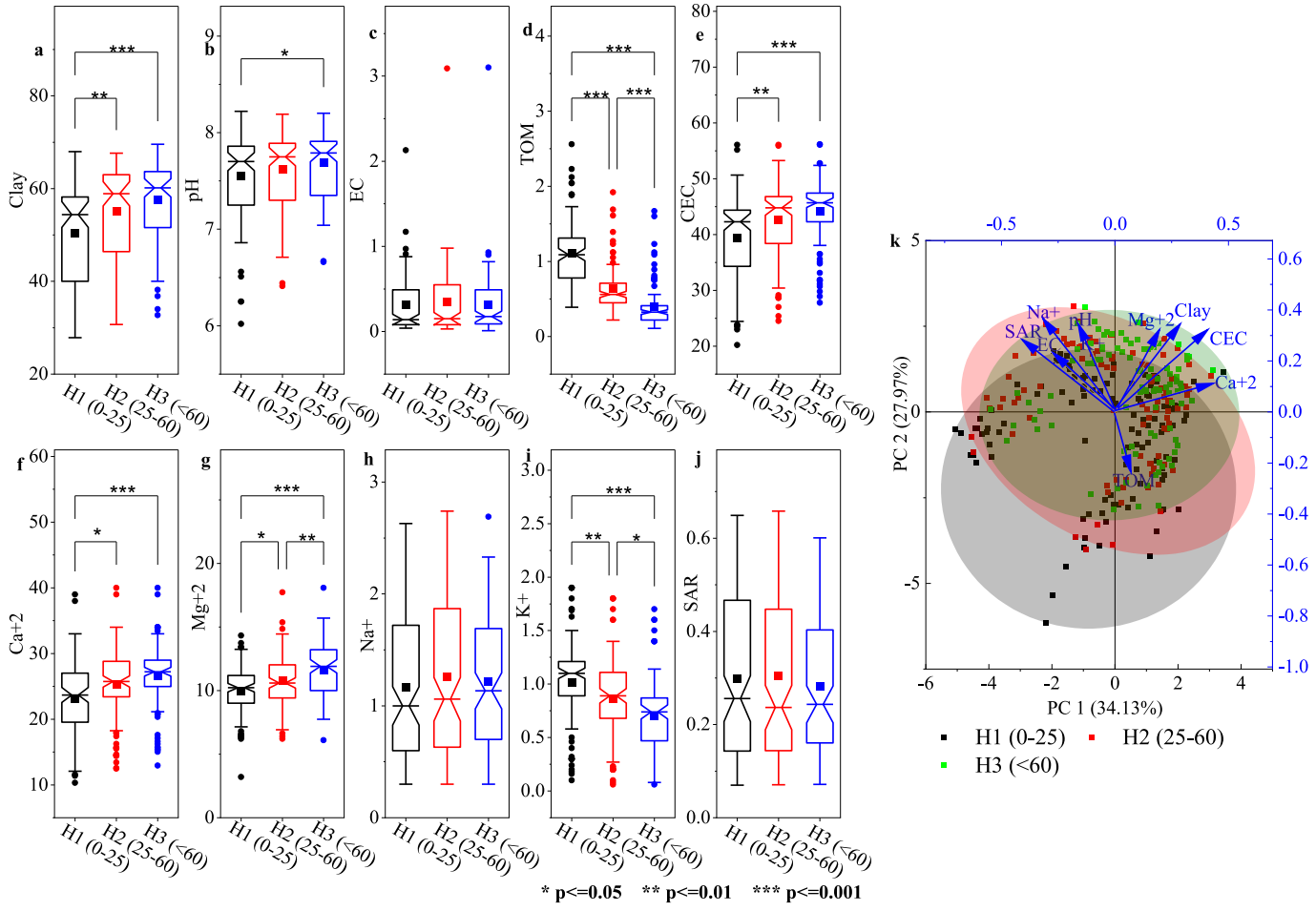

Fig. 2. Box plot and PCA analysis for studied soil properties: **a**) clay (%), **b**) pH (H_2O), **c**) EC (dsm^{-1}), **d**) total organic matter (TOM) (%), **e**) cation exchange capacity (CEC) (Cmol$_c$ kg^{-1}), **f**) Ca^{2+} (Cmol$_c$ kg^{-1}), **g**) Mg^{2+} (Cmol$_c$ kg^{-1}), **h**) Na^{+} (Cmol$_c$ kg^{-1}), **i**) K^{+}(Cmol$_c$ kg^{-1}), **j**) SAR, and **k**) PCA. (■: average, ●: outliers, —: median, $H1_{0\text{-}25}$: soil depth 0–25 cm, $H2_{25\text{-}60}$: soil depth 25–60 cm, $H3_{<60cm}$: soil depth <60 cm).

performed with $R^2 = 0.989$, RMSE = 0.016, MSE = 0.0003, and MAPE = 3.54%, and for SC4 with five input features ($R^2 = 0.988$, RMSE = 0.017, MSE = 0.003, MAPE = 3.823%) (Table 4).

Following the sequence, RF performance in SC3 with six features is ranked with $R^2 = 0.982$, RMSE = 0.021, MSE = 0.0004, and MAPE = 4.153%, and SC2 with seven features ($R^2 = 0.981$, RMSE = 0.021, MSE = 0.0005, and MAPE = 4.33%. Lastly, ANN-MLP performed the least in the ML model sequence, with high performance in SC4 with five features ($R^2 = 0.942$, RMSE = 0.0379, MSE = 0.001, and MAPE = 8.9%) (Table 4). Overall, in the training stage, the performance of the ML models is ranked as NuSVR > GBR > RF > ANN-MLP, and the scenarios are ranked as SC1(eight features-8F) > SC2 (seven features-7F) > SC3 (six features-6F) > SC4 (five features-5F) > SC5 (four features-4F) (Fig S4, Fig S5).

Furthermore, SAR predictions from all ML algorithms and input combinations presented by a half-box distribution with filter points showed that in the training set, NuSVR predictions were found to be closest to the observed ones, followed by GBR and RF in all scenarios. ANN-MLP predictions revealed a few outliers in SC2, SC3, and SC5. Interestingly, a comparison of the normal distributions of the predicted and observed data reveals that the NuSVR output closely aligns with the observed SAR in most of the studied scenarios (Fig. S6).

3.2.2. ML algorithms performance in testing stage

SAR predictions on the test data also revealed that **the performance of NuSVR was superior to that of other ML models in SC2, with seven features achieving the highest $R^2 = 0.999$, lowest RMSE = 0.004, MSE = 0.00001, and MAPE = 1.673%.** This was followed by SC1 with eight features, achieving $R^2 = 0.998$, RMSE = 0.004, MSE = 0.00001, and MAPE = 1.77%, and SC3 with six features ($R^2 = 0.998$, RMSE = 0.005, MSE = 0.00003, and MAPE = 1.92%) (Table 4). Furthermore, in NuSVR, SC4 with five features performed with $R^2 = 0.996$, RMSE = 0.009, MSE = 0.00009, and MAPE = 2.93% (Fig. 3).

Following the sequence of accurate predictions by the ML algorithms, GBR in SC2 with seven features was performed with $R^2 = 0.991$, RMSE = 0.016, MSE = 0.0003, and MAPE = 3.60%. Moreover, SC1, SC3, and SC4 achieved an R^2 of 0.990 and RMSE = 0.0.16 for SAR predictions, but with MAPE = 3.63%, 3.67%, and 3.78%, respectively. Furthermore, ANN-MLP proved for SAR prediction in SC1 and SC2 with eight and seven features, with $R^2 = 0.989$ and 0.988, RMSE = 0.017 and 0.018, MSE = 0.0003, and MAPE = 5.99% and 5.82%, respectively (Table 4).

Subsequently, RF in SC3 with six features was performed with $R^2 = 0.988$, RMSE = 0.018, MSE = 0.0003, and MAPE = 3.86%, followed by SC2 with seven features ($R^2 = 0.987$, RMSE = 0.019, MSE = 0.0004, and MAPE = 4.169 (Fig. 3). Hence, scatterplots of SAR predictions on the test data revealed that SC2 with seven features outperformed other input combinations from SVR-RFE and was ranked as SC2 (seven features) > SC1 (eight features) > SC3 (six features) > SC4 (five features) > SC5 (four features). Machine learning models, according to their performance accuracy, are ranked as NuSVR > GBR > ANN-MLP > RF (Fig. 4).

3.2.3. ML model's prediction error and efficiency (training and testing stage)

After performance assessment, prediction error evaluation based on PBIAS showed the best observed and predicted match with negligible error in NuSVR algorithm in SC1 with eight selected features (PBIAS = 0.09%) and SC2 with seven features (PBIAS = −0.01%) in testing stage. In training stage also, lowest PBIAS of 2.1% and 2.3% is shown by NuSVR in SC1 and SC2. The highest underestimation in training stage is shown by ANN-MLP in SC4 with PBIAS = 2.46% followed by RF in SC5 with PBIAS = 1.64%. At testing stage highest underestimation is revealed by ANN-MLP in SC2 with PBIAS = 2.35% and highest overestimation is also shown by ANN-MLP in SC5 and SC3. Overall, all ML models can be referred to perform good with less than 10% PBIAS. Afterwards, the highest NSE = 0.999 is also achieved by NuSVR in SC2 and SC1 (Table 4).

Like training stage, SAR predictions from all ML algorithms presented by half box distribution with fitter points showed that in test set, NuSVR predictions (blue color) are found to be closest to the observed ones (black color). The SAR predicted average from NuSVR was much closer to the observed average of SAR, followed by GBR and RF in all scenarios (Fig. 5).

3.2.4. SHAP analysis of best performed model

The SHAP interpretation of the best-performing model, specifically the NuSVR for SC2 (7F), displayed the rankings and values of the input variables for both the training and test datasets. (Fig. 6). It efficiently portrays the significant impact or contribution of each variable to model prediction. The rank of each individual variable is presented on the Y-axis with SHAP values on the X-axis, and each dot represents the corresponding SHAP value for each instance. Hence, the data distribution along the X-axis signifies its influence on the output of the model, with positive (high) values indicating a higher influence on the prediction and negative (low) values indicating a lower contribution to the predictions. In this sense, CEC, Ca^{2+}, Mg^{2+}, and Na^{+} were found to have a

Table 4
ML model's performance evaluation, prediction error, and model's efficiency evaluation at training and testing stages.

Algorithm	Scenario	Training						Testing					
		R^2	RMSE	MSE	MAPE	PBIAS	NSE	R^2	RMSE	MSE	MAPE	PBIAS	NSE
RF	**SC1(8F)**	0.975	0.025	0.0006	5.097	1.003	0.975	0.982	0.023	0.0005	4.66	0.666	0.988
	SC2(7F)	0.982	0.021	0.0005	4.335	0.87	0.98	0.988	0.019	0.0004	4.17	0.996	0.988
	SC3(6F)	0.982	0.021	0.0004	4.154	0.739	0.976	0.988	0.018	0.0003	3.866	1.675	0.986
	SC4(5F)	0.976	0.024	0.0006	4.668	0.786	0.971	0.986	0.02	0.0004	4.006	2.123	0.962
	SC5(4F)	0.971	0.027	0.0007	5.295	1.648	0.982	0.962	0.034	0.0011	6.328	0.581	0.982
NuSVR	**SC1(8F)**	0.997	0.008	0.0001	2.101	−0.389	0.997	0.998	0.004	0	1.771	0.090	0.999
	SC2(7F)	0.996	0.01	0.0001	2.365	−0.38	1.00	0.999	0.004	0	1.673	−0.017	0.999
	SC3(6F)	0.996	0.01	0.0001	2.328	−0.508	0.994	0.999	0.006	0	1.927	0.205	0.997
	SC4(5F)	0.994	0.012	0.0001	3.119	−0.229	0.974	0.997	0.009	0.0001	2.936	0.174	0.974
	SC5(4F)	0.974	0.026	0.0007	5.921	0.561	0.998	0.974	0.028	0.0008	6.403	−0.419	0.996
ANN-MLP	**SC1(8F)**	0.936	0.04	0.0016	12.443	−0.849	0.936	0.99	0.017	0.0003	5.998	0.314	0.988
	SC2(7F)	0.86	0.059	0.0035	15.297	0.37	0.93	0.988	0.018	0.0003	5.827	2.357	0.979
	SC3(6F)	0.929	0.042	0.0018	11.99	−0.912	0.943	0.979	0.025	0.0006	7.042	−1.201	0.982
	SC4(5F)	0.943	0.038	0.0014	8.935	2.466	0.873	0.982	0.023	0.0005	7.755	1.268	0.959
	SC5(4F)	0.873	0.056	0.0032	13.612	−1.369	0.990	0.959	0.035	0.0012	10.111	−1.920	0.860
GBR	**SC1(8F)**	0.99	0.016	0.0002	3.61	0.427	0.990	0.991	0.016	0.0003	3.634	0.433	0.991
	SC2(7F)	0.99	0.015	0.0002	3.597	0.34	0.99	0.991	0.016	0.0003	3.601	0.301	0.991
	SC3(6F)	0.99	0.016	0.0003	3.542	0.557	0.988	0.991	0.016	0.0003	3.676	0.257	0.991
	SC4(5F)	0.988	0.017	0.0003	3.824	0.244	0.974	0.991	0.017	0.0003	3.783	1.248	0.969
	SC5(4F)	0.974	0.025	0.0006	5.595	0.360	0.991	0.969	0.03	0.0009	6.018	0.208	0.990

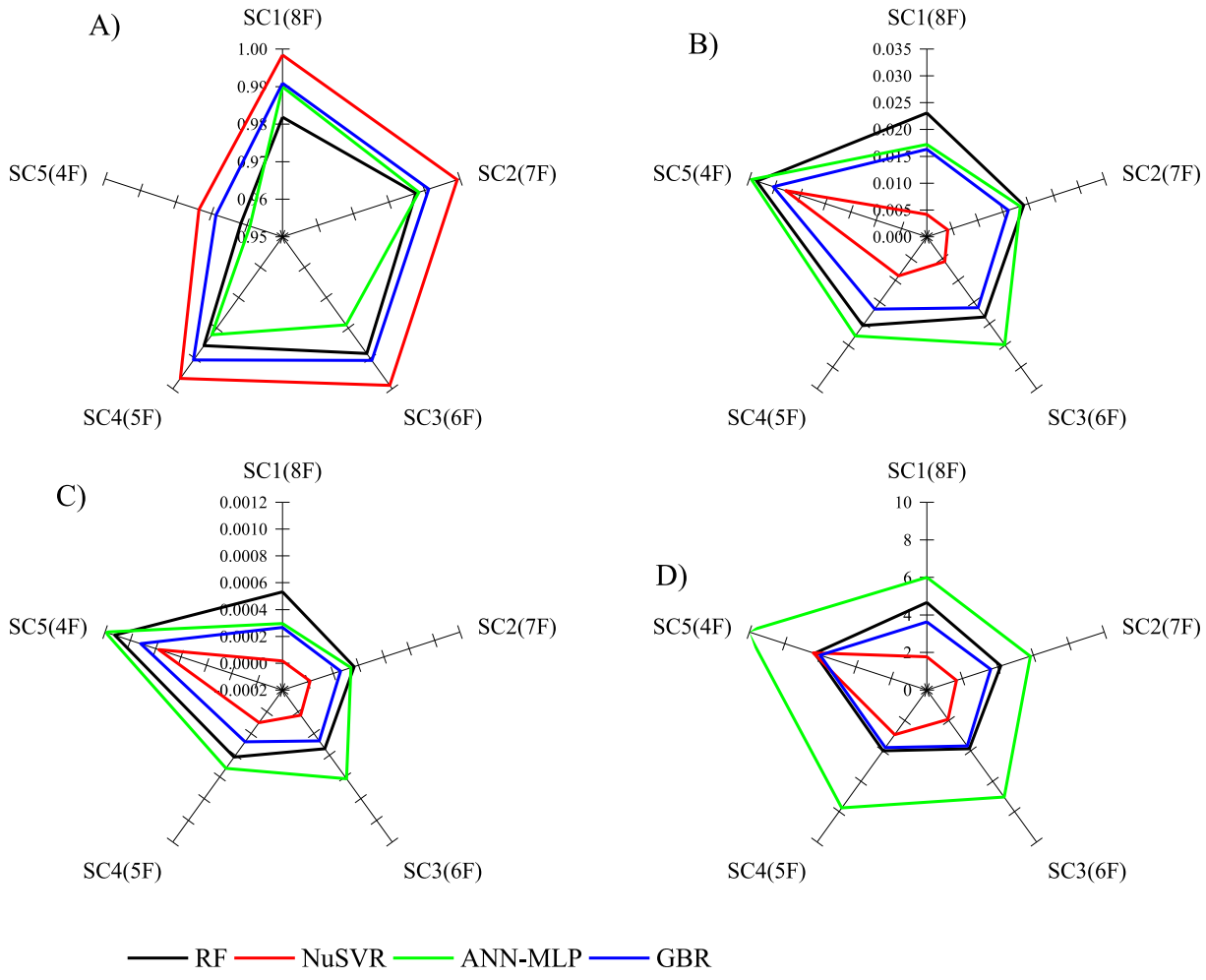

Fig. 3. Machine learning performance evaluation in all input combinations (scenarios) in test set A) R^2, B) RMSE, C) MSE, D) MAPE.

greater influence and contribution to predicting SAR in both the training and testing stages. pH, K^+, and EC were found to have minimal influence on SAR prediction.

4. Discussion

4.1. Physical and human causes of salt affected soil in the Mediterranean region

Salt accumulation in Mediterranean soils is a naturally occurring phenomenon that is facilitated by ecological factors specific to the region and is primarily regulated by the hydrological balance in the area (Aragüés et al., 2011; Herrero and Pérez-Coveta, 2005; Montazeri et al., 2023). The geochemical properties of the groundwater also define the salinity of a region. For example, historic drainage canals in the coastal Mediterranean region suggest a long-term geological impact on soil salinity (Marien et al., 2023). For instance, our research findings imply that Na^+ remained relatively consistent across the soil horizons, with an average 1.214 ± 0.54 in the H3 (Fig. 2 (h)) and mean SAR values in H1 were found to be 0.299 ± 0.17 and 0.281 ± 0.13 in H3 reflecting less sensitivity to sodicity. This could be attributed to the inherited salinity from the parent material or sufficient soil moisture, which prevented any significant salt accumulation. Overall, soil structure, climatic regime, and irrigation pattern are key factors driving soil salinity and sodicity (Lagacherie et al., 2018). Nevertheless, inappropriate irrigation methods are a common cause of soil salinity in southern Syria (Kamrakji et al., 2016).

Agricultural activities alter the hydrological balance, which leads to the accumulation of salts under limited drainage conditions, thereby accelerating land degradation in many parts of the semi-arid Mediterranean environment (Aragüés et al., 2011). For instance, Kattan (2020) demonstrated low electrical conductivity in rain samples from southern Syria linked to low-carbon atmospheric dust in the region. Moreover, the geological structure of the Mediterranean Basin is rich in limestone and calcareous rocks, leading to the deposition of calcium carbonate, high pH, and altered clay particles. These properties strongly affects the soil texture, water retention and drainage capacity (Lagacherie et al., 2018). Calcium rich soil is also supplemented by Mg^{2+} and affects the cation exchange capacity (CEC) (Bouajila et al., 2023). It is consistent with our findings with high coefficients loadings of CEC, Ca^{2+}, and Mg^{2+} in PC1 depicting their strong interrelationship (Fig. 2 (k)). Furthermore, the mean CEC values in the first horizon ($H1_{0\text{-}25}$) were found to be 39.38 ± 7.89 Cmolc kg^{-1} which significantly increased to 44.15 ± 5.90 Cmolc kg^{-1} in the H3 (Fig. 2 (e)). These high CEC values can be linked to the presence of clay minerals, such as montmorillonite and smectite in the soil profiles (Mohammed et al., 2020b) or due to accumulation of secondary carbonates and/or weathering of calcareous parent material (Kargas et al., 2023). Afterwards, the mean TOM significantly ($p < 0.05$) decreased from 1.11% ± 0.42 in the first horizon ($H1_{\ 0\text{-}25}$) to 0.39% ± 0.28 in $H3_{<60}$. It aligns with previous research conducted in the Mediterranean region and elsewhere, stating that the soil organic matter typically decreases with increasing soil depth (Çelik et al., 2019; Lieberman et al., 2020). Previously, Mohammed et al. (2020a) reported a decreasing of TOM with increasing soil depth in all four Syrian soil orders, due to lack of different TOM sources and rapid humification process. Hence, high dimensionality and varied behavior of cations in

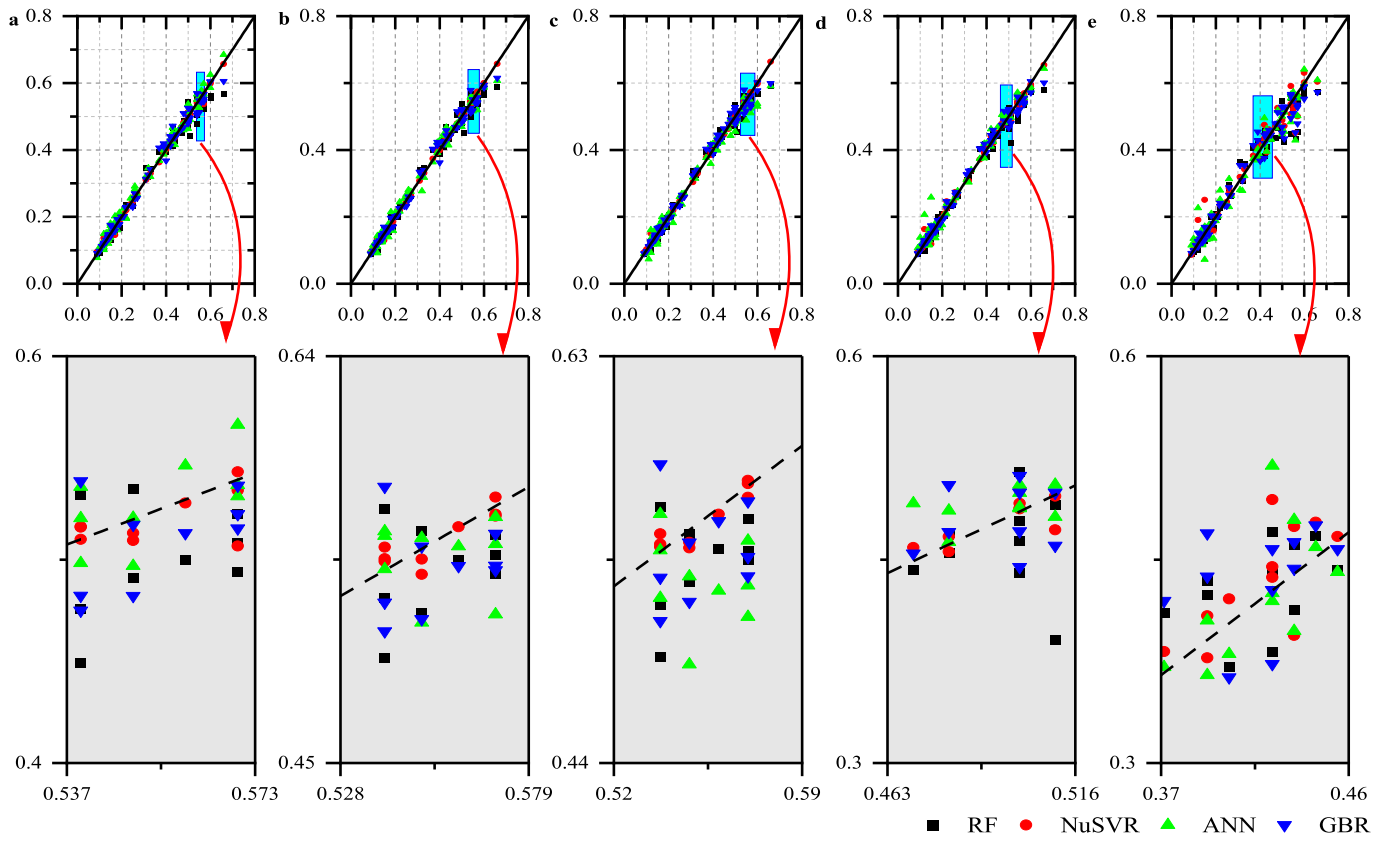

Fig. 4. Scatter plot comparing observed and ML-predicted SAR values for all input combinations (scenarios) in the testing set: a) Scenario 1 (8F), b) Scenario 2 (7F), c) Scenario 3 (6F), d) Scenario 4 (5F), e) Scenario 5 (4F).

different soil horizons lead to explaining the applicability of ML algorithms for accurate SAR prediction.

4.2. Comparative performance of ML in predicting soil-SAR

In this research, iterative selection of input combinations in SVR-RFE provided a good opportunity for ML algorithms to perform accurate SAR predictions. Previously, Khan et al. (2020) and recently, Mohamed et al. (2023) also employed the wrapper based Recursive Feature Elimination method to identify the best features for accurate drought and salinity predictions from SVM, RF, and ANN models. In our research, all scenarios of selected input combinations were trained separately for accurate prediction modeling of soil salinity from four competitive and well performed ML algorithms and revealed the highest accuracy of NuSVR over RF, ANN, and GBR on both train and test datasets. The current selection of chosen ML algorithms is based on their competitive performance in recent soil prediction studies (Andrade Foronda and Colinet, 2023; Kaplan et al., 2023; Sarkar et al., 2023).

NuSVR is the least applied version of support vector regression for prediction modeling (Bhatt et al., 2012). The ability of SVR in dealing with high dimensional data and detecting outliers makes it a valuable algorithm in predicting soil problems (Mohamed et al., 2023; Tang et al., 2023). Inclusion of regularization parameters (nu) helps in controlling the number of supports vectors and affects the balance between fitting the training data and generalizing to new data. Depending on the required level of regularization, NuSVR may outperform in specific scenarios (Onyekwena et al., 2022). Currently, following the Grid Search optimization with 5-fold CV, NuSVR outperformed other algorithms in all input combinations with the highest $R^2 = 0.999$, lowest RMSE = 0.004, MSE = 0.00001, and MAPE = 1.673% in combination of 7 variables pH, EC, CEC, Ca^{2+}, Mg^{2+}, Na^{+}, and K^{+} (Table 4). Previously, pH and EC are proven to be accurate predictors of soil salinity with neural networks (Sarani et al., 2016). However, variables interpretation from SHAP analysis in our research provided a good explanation of CEC, Ca^{2+}, and Mg^{2+} impact on SAR prediction (Fig. 6). SHAP is a comprehensive method combining game theory with local explanations providing a better feature attribution approach (Vega García and Aznarte, 2020).

Decision trees (DT) based Random Forest and Gradient Boosting Regressions also supported the SAR prediction modeling through averaging multiple DTs and sequential learning of weak DTs. A large amount of less or decorrelated DTs form ensemble Random Forest, helped in reducing variance and overfitting, increasing the generalizability of the model (Tran et al., 2021). Moreover, Decision tree algorithm of RF also computes variable or feature importance score that facilitates in appropriate feature selection from a large dataset (Yang et al., 2023). Currently, the performance of RF in model trainings of SAR prediction is ranked 3rd after NuSVR and GBR with the highest R^2 of 0.982 in SC3 (6F) (Table 4). Random Forest is robust against outliers in the data which are significantly observant in the cations of different soil horizons. It takes into consideration the majority opinion of several trees, thereby decreasing the influence of outliers on the model's performance. Moreover, training multiple trees can be parallelized, making Random Forest an efficient algorithm for large datasets, already proven in soil salinity prediction (Wang et al., 2020a, 2020b).

A boosting model creates decision trees in sequence, or one tree can be built on the experience of previous trees. The second tree focuses on cases where the first tree predicts poorly, and this process is repeated to capture the best relationship between response and explanatory variables (Mantena et al., 2023). Currently, the performance of GBR is ranked 2nd highest after NuSVR for predicting SAR in both train and test datasets (Figs. 3 and 5). This is because GBR assesses weak prediction cases to minimize the overall loss function. Only valuable decision trees are kept in an ensemble. Moreover, the learning rate of GBR tunes the

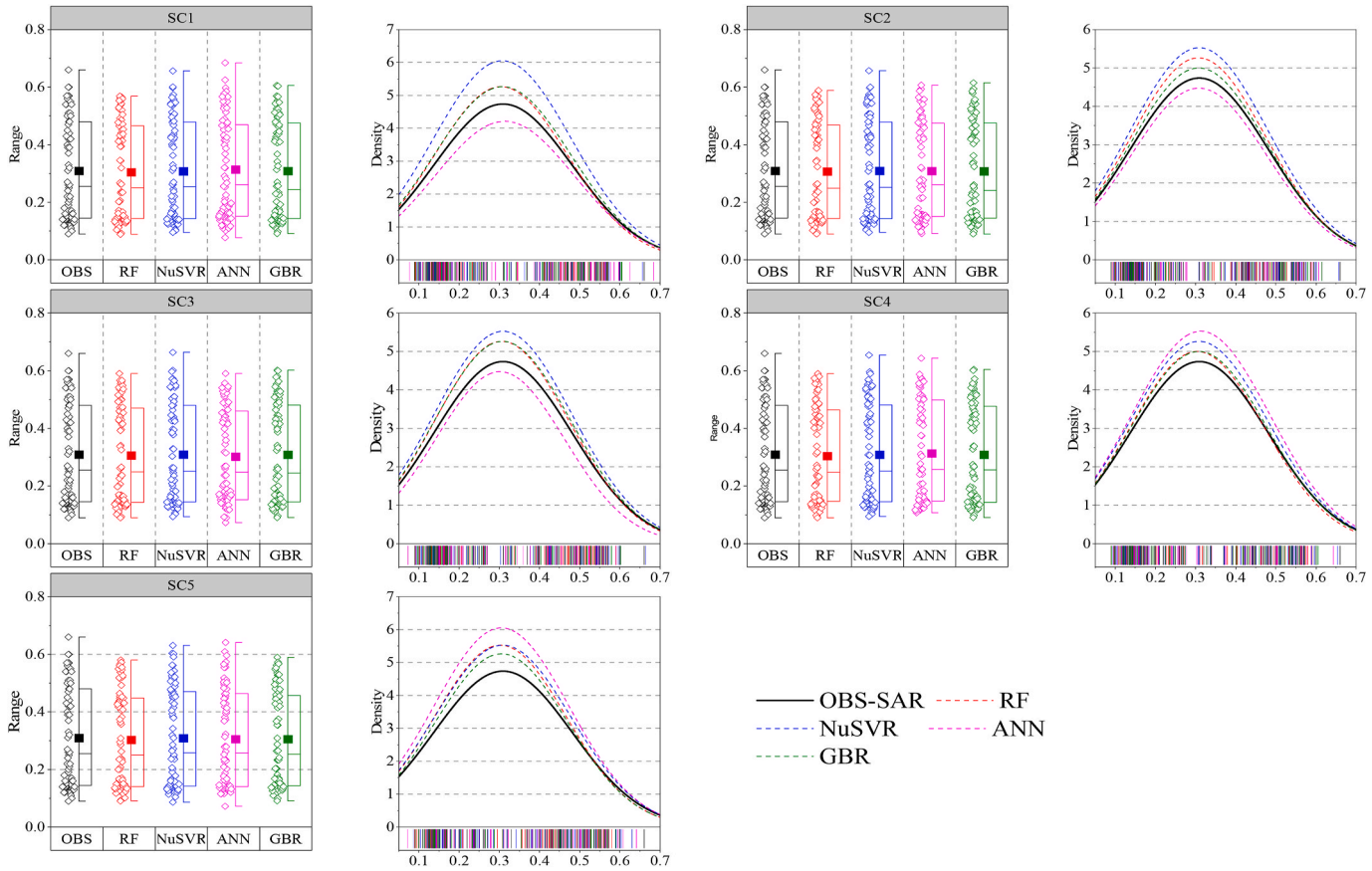

Fig. 5. Hybrid half box plot with fitter points showing the range of observed and predicted SAR in different ML algorithms along with density distribution for all input combinations (SC1 to SC5) in training stage. (Squar box ■ presents mean; central line _ presents median; half box presents interquartile range (IQR)).

model performance to improve its generalizability. The GBM is proven to be advantageous in handling complex relationships of multivariable in soil salinity prediction (Chen et al., 2022; Zarei et al., 2021) and quantifying saline concentration in groundwater (Abba et al., 2023). For instance, Salem et al. (2023) proved the efficiency of boosting and ANN regressions for predicting SAR in arid eastern Mediterranean region. Another recent study by Das et al. (2023) proved the highest efficiency of RF and deep learning ensemble for soil salinity prediction aided by hyperspectral remote sensing. The efficiency of the MLP in prediction soil characteristics was previously reported by Tizpa et al. (2015) and recently by Cherif et al. (2023). Furthermore, several scholars highlighted the potential of ANN-MLP in predicting different soil parameters (Šajn et al., 2022; Taha et al., 2018; Tizpa et al., 2015). The strength of ANN-MLP in predicting soil parameters lies in its ability to capture nonlinear relations between input and output variables (Verma and Kumar, 2021). For instance, Habibi et al. (2021) and Wang et al. (2020a) also reported the high competency of ANN-MLP and RF in soil salinity prediction. The ANN-MLP architecture with hidden layers is driven by significant activation function which determines the non-linearity of neurons and handles the complex relationships of data. By adjusting synaptic weights and bias parameters through iterative training, ANN-MLP models can learn and adapt to the data (Wang et al., 2020a). Hence, the inherent structure of adopted algorithms such as ensemble nature of RF and GBR, regularization in ANN and NuSVR set a foundation for subsequent comprehensive analysis with less chance of uncertainty. Additionally, SHAP analysis is a valuable ML interpretation method measuring variable importance, particularly in scenarios of sensitivity and uncertainty assessment (Prots et al.). It helped us to understand the significant contribution of each feature in model predictions. SHAP analysis identifies the most influential parameters in the model, which facilitates to understand the model behavior and robustness, addressing the sensitivity concerns (Herren and Hahn, 2022; Prots et al.).

4.3. Limitation of ML in predicting soil SAR

Prediction models of soil salinity developed in one area cannot be implemented in another region due to disparities or differences in soil properties (Das et al., 2023). Hence, developing a universal soil prediction model with high accuracy is challenging. Model transferability failure is rooted in soil heterogeneity and different sensitive prediction features across sites (Das et al., 2023; Zeraatpisheh et al., 2020). However, finding sensitive and more relevant features can improve the accuracy of soil salinity prediction in varied regions. Hence, NuSVR along with SHAP kernel implementation provided a new valid approach for soil salinity prediction in our research. In this context, the effectiveness of NuSVR is also sensitive to the kernel function chosen, making the selection process challenging. Implementing NuSVR involves solving a complex quadratic programming problem, demanding substantial computational resources. Additionally, NuSVR might require a data-dependent weighting function for parameter computation, further complicating the modeling process (Hao, 2017).

The limitations of Random Forest include insensitivity to variable group size, potential fine-tuning issues, challenging interpretation, and questions about bias reduction compared to boosting algorithms (Cai et al., 2020). On the other hand, model parameters control the learning efficiency of each algorithm, and for the XGB model, they include tree-specific, boosting, and metric groups (Chen et al., 2022). Selecting these parameters is challenging and may not always result in an optimum set. We used a grid-search algorithm with cross-validation which helped to improve the accuracy of the GBR. However, the complexity of ANN-MLP models can make them difficult to comprehend, especially for

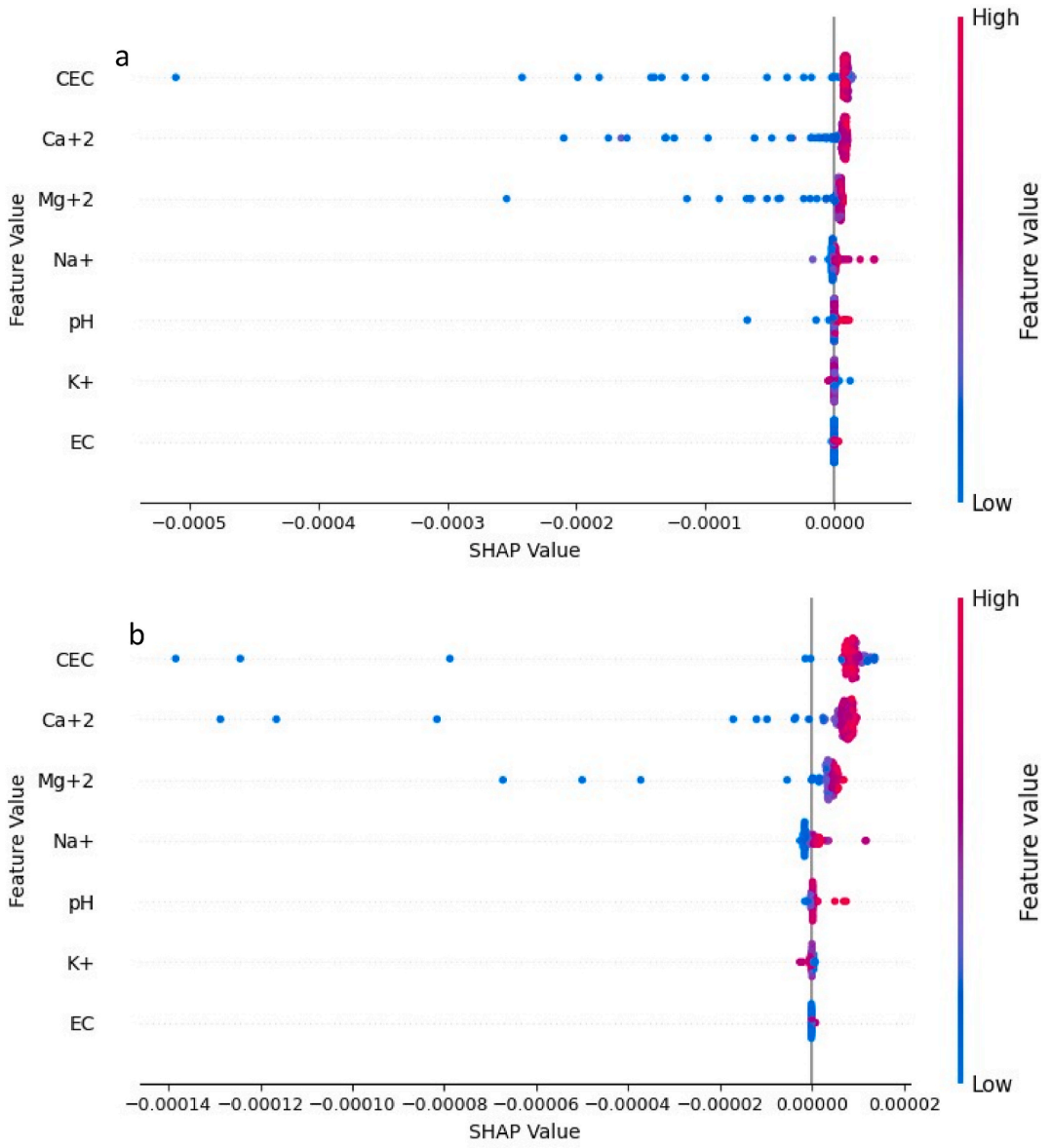

Fig. 6. SHAP score ranking the 7 features with impact on high accurate prediction model output of NuSVR in training set. The side bar presents the magnitude of impact with high positive SHAP values in red and low negative SHAP values in blue. a) train set, b) test set). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

larger or deeper networks. Inadequate sizing might result in overfitting and restricted adaptability to new data. Training Neural networks require substantial computational resources, particularly with extensive datasets and complex structures (Ghorbani et al., 2016; Mandal and Mondal, 2019). Moreover, the performance of ANN-MLP is also subjected to limiting factors such as the input features, the number of hidden layers, and soil type (Verma and Kumar, 2021).

5. Conclusion

Under the ongoing climatic change, salinization is one of the biggest challenges for agricultural sustainability, especially in the post-war phase in Syria. However, salinity and sodicity are region-specific problems primarily associated with natural physical factors. Agricultural intensification and human activities have exaggerated the problem, with severe impacts on crop production. Geographical variations in soil properties enforce the need for a region-specific soil prediction modeling approach. Hence, considering the significance of this research problem, this study explored nine soil properties (clay, pH, EC, Na^{+}, Mg^{2+}, Ca^{2+}, K^{+}, TOM, and CEC) and the sodicity indicator SAR in three soil horizons H1 $_{0\text{-}25}$, H2 $_{25\text{-}60}$, and H3 $_{<60}$, employing Tukey mean difference and Principal Component Analysis (PCA). The study also predicted the SAR from representative soil properties of southern Syria using five input combinations iteratively derived from the SVM Recursive Feature Elimination method. The major conclusions drawn from the analysis are as follows.

1. No significant ($p < 0.001$) mean difference is observed in SAR with mean value in H1 was found to be 0.299 ± 0.17 and 0.281 ± 0.13 in H3 reflecting less sensitivity to sodicity.
2. Three PCs with eigenvalues above 1 explained 74.6% of the total variance and high dimensionality of the data. The CEC, Ca^{2+}, Mg^{2+}, and clay exhibited high positive loadings of 0.42, 0.47, 0.24 and 0.34 PC1. The second cluster consisted of SAR, Na^{+}, EC, and pH, with high positive loadings of 0.30, 0.38, 0.25, and 0.35 PC2. The TOM was found to be more important in PC3.
3. The NuSVR ML model was found to be innovative and outperformed in a good competition with GBR and RF in both the training and test

datasets, with the highest accuracy achieved in SC2 (7F) with R^2 = 0.999, lowest RMSE = 0.004, MSE = 0.00001, and MAPE = 1.673% in combination with pH, EC, CEC, Ca^{2+}, Mg^{2+}, Na^{+}, and K^{+}.
4. SHAP analysis from the best prediction model, NuSVR, revealed CEC, Ca^{2+}, and Mg^{2+}, and Na^{+} had more influence in accurately predicting SAR in both training and testing sets, while pH, K^{+}, and EC were found to have minimal influence on SAR prediction.

Hence, our study succeeded in developing a highly accurate SAR prediction model (NuSVR) with the highest accuracy, which has not been tested in previous salinity prediction studies. Moreover, finding more relevant predictors was supported by SVM-RFE, and the SHAP analysis provided a more in-depth understanding of intricate relationships. The illuminating insights gained from the SHAP analysis not only enhanced the interpretability of our findings, but also shed light on the profound impact of individual variables on SAR prediction. For future research, additional studies will be conducted to demonstrate the applicability of this innovative prediction modeling approach in other regions. Also, this research will be extended to include different geographical locations in Syria to test the performance of NuSVR in accurately predicting SAR. Finally, the practical implications of this research extend to the informed decision-making process aimed at revolutionizing soil management practices to optimize crop yields and resource utilization.

Funding

This research was supported by the Researchers Supporting Project, Grant number (RSP 2024R296), King Saud University, Riyadh, Saudi Arabia. Also, Project no. TKP2021-NKTA-32 has been implemented with support from Hungary's National Research, Development, and Innovation Fund, financed under the TKP2021-NKTA funding scheme. Also, the first author would like to thank the University of Debrecen Program for Scientific Publication for its support.

CRediT authorship contribution statement

Safwan Mohammed: Conceptualization, Formal analysis, Writing – original draft, Supervision. **Sana Arshad:** Conceptualization, Writing – original draft, Formal analysis. **Bashar Bashir:** Conceptualization, Writing – review & editing. **Behnam Ata:** Writing – original draft. **Main Al-Dalahmeh:** Writing – review & editing. **Abdullah Alsalman:** Formal analysis. **Haidar Ali:** Data curation. **Sami Alhennawi:** Data curation. **Samer Kiwan:** Data curation. **Endre Harsanyi:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

https://doi.org/10.1016/j.dib.2020.105832

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jenvman.2024.122640.

References

Abba, S.I., Benaafi, M., Usman, A.G., Ozsahin, D.U., Tawabini, B., Aljundi, I.H., 2023. Mapping of groundwater salinization and modelling using meta-heuristic algorithms for the coastal aquifer of eastern Saudi Arabia. Sci. Total Environ. 858, 159697.

Abd El-Halim, A.E.-H.A., Salama, A.M., Ibrahim, M.M., Aiad, M.A., Shokr, M., 2023. Nano-gypsum in low dose improves the physicochemical properties of saline-sodic soil. Arch. Agron Soil Sci. 69, 2286–2299.

Abdullah, A.Y.M., Biswas, R.K., Chowdhury, A.I., Billah, S.M., 2019. Modeling soil salinity using direct and indirect measurement techniques: a comparative analysis. Environmental Development 29, 67–80.

Abedi, F., Amirian-Chakan, A., Faraji, M., Taghizadeh-Mehrjardi, R., Kerry, R., Razmjoue, D., Scholten, T., 2021. Salt dome related soil salinity in southern Iran: prediction and mapping with averaging machine learning models. Land Degrad. Dev. 32, 1540–1554.

Abu Hammad, A., Tumeizi, A., 2012. Land degradation: socioeconomic and environmental causes and consequences in the eastern Mediterranean. Land Degrad. Dev. 23, 216–226.

Andrade Foronda, D., Colinet, G., 2023. Prediction of soil salinity/sodicity and salt-affected soil classes from soluble salt ions using machine learning algorithms. Soil Systems 7, 47.

AquaStat, F., 2022. FAO's global information system on water and agriculture [WWW Document]. Food Agric. Organ. UN URL. https://tableau.apps.fao.org/views/ReviewDashboardv1/region_dashboard.

Aragüés, R., Urdanoz, V., Çetin, M., Kirda, C., Daghari, H., Ltifi, W., Lahlou, M., Douaik, A., 2011. Soil salinity related to physical soil characteristics and irrigation management in four Mediterranean irrigation districts. Agric. Water Manag. 98, 959–966.

Arshad, S., Kazmi, J.H., Javed, M.G., Mohammed, S., 2023a. Applicability of machine learning techniques in predicting wheat yield based on remote sensing and climate data in Pakistan, South Asia. Eur. J. Agron. 147, 126837.

Arshad, S., Kazmi, J.H., Prodhan, F.A., Mohammed, S., 2023b. Exploring dynamic response of agrometeorological droughts towards winter wheat yield loss risk using machine learning approach at a regional scale in Pakistan. Field Crops Res. 302, 109057.

Bhatt, D., Aggarwal, P., Bhattacharya, P., Devabhaktuni, V., 2012. An enhanced MEMS error modeling approach based on nu-support vector regression. Sensors 12, 9448–9466.

Bogdanova, A., Imakura, A., Sakurai, T., 2023. DC-SHAP method for consistent explainability in privacy-preserving distributed machine learning. Human-Centric Intelligent Systems 3, 197–210.

Bouajila, K., Hechmi, S., Mechri, M., Jeddi, F.B., Jedidi, N., 2023. Short-term effects of Sulla residues and farmyard manure amendments on soil properties: cation exchange capacity (CEC), base cations (BC), and percentage base saturation (PBS). Arabian J. Geosci. 16, 410.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Cai, J., Xu, K., Zhu, Y., Hu, F., Li, L., 2020. Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. Appl. Energy 262, 114566.

Çelik, İ., Günal, H., Acar, M., Acir, N., Bereket Barut, Z., Budak, M., 2019. Strategic tillage may sustain the benefits of long-term no-till in a Vertisol under Mediterranean climate. Soil Tillage Res. 185, 17–28.

Chandra Joshi, R., Ryu, D., Lane, P.N.J., Sheridan, G.J., 2023. Seasonal forecast of soil moisture over Mediterranean-climate forest catchments using a machine learning approach. J. Hydrol. 619, 129307.

Chen, B., Zheng, H., Luo, G., Chen, C., Bao, A., Liu, T., Chen, X., 2022. Adaptive estimation of multi-regional soil salinization using extreme gradient boosting with Bayesian TPE optimization. Int. J. Rem. Sens. 43, 778–811.

Cherif, K., Yahia, N., Bilal, B., Bilal, B., 2023. Erosion potential model-based ANN-MLP for the spatiotemporal modeling of soil erosion in wadi Saida watershed. Modeling Earth Systems and Environment 9, 3095–3117.

Choukr-Allah, R., Mouridi, Z.E., Benbessis, Y., Shahid, S.A., 2023. Salt-affected soils and their management in the Middle East and north africa (mena) region: a holistic approach. In: Choukr-Allah, R., Ragab, R. (Eds.), Biosaline Agriculture as a Climate Change Adaptation for Food Security. Springer International Publishing, Cham, pp. 13–45.

Cuevas, J., Daliakopoulos, I.N., del Moral, F., Hueso, J.J., Tsanis, I.K., 2019. A review of soil-improving cropping systems for soil salinization. Agronomy 9, 295.

Daliakopoulos, I.N., Tsanis, I.K., Koutroulis, A., Kourgialas, N.N., Varouchakis, A.E., Karatzas, G.P., Ritsema, C.J., 2016. The threat of soil salinity: a European scale review. Sci. Total Environ. 573, 727–739.

Das, A., Bhattacharya, B.K., Setia, R., Jayasree, G., Sankar Das, B., 2023. A novel method for detecting soil salinity using AVIRIS-NG imaging spectroscopy and ensemble machine learning. ISPRS J. Photogrammetry Remote Sens. 200, 191–212.

de la Paix, M.J., Lanhai, L., Xi, C., Varenyam, A., Nyongesah, M.J., Habiyaremye, G., 2013. Physicochemical properties of saline soils and aeolian dust. Land Degrad. Dev. 24, 539–547.

Descals, A., Verger, A., Yin, G., Filella, I., Peñuelas, J., 2023. Local interpretation of machine learning models in remote sensing with SHAP: the case of global climate constraints on photosynthesis phenology. Int. J. Rem. Sens. 44, 3160–3173.

El Bilali, A., Taleb, A., Nafii, A., Alabjah, B., Mazigh, N., 2021. Prediction of sodium adsorption ratio and chloride concentration in a coastal aquifer under seawater intrusion using machine learning models. Environmental Technology & Innovation 23, 101641.

Elsherbiny, O., Fan, Y., Zhou, L., Qiu, Z., 2021. Fusion of feature selection methods and regression algorithms for predicting the canopy water content of rice based on hyperspectral data. Agriculture 11, 51.

Eswar, D., Karuppusamy, R., Chellamuthu, S., 2021. Drivers of soil salinity and their correlation with climate change. Curr. Opin. Environ. Sustain. 50, 310–318.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378.

Gautam, V.K., Pande, C.B., Moharir, K.N., Varade, A.M., Rane, N.L., Egbueri, J.C., Alshehri, F., 2023. Prediction of sodium hazard of irrigation purpose using artificial neural network modelling. Sustainability 15, 7593.
Gharaibeh, M.A., Albalasmeh, A.A., Pratt, C., El Hanandeh, A., 2021. Estimation of exchangeable sodium percentage from sodium adsorption ratio of salt-affected soils using traditional and dilution extracts, saturation percentage, electrical conductivity, and generalized regression neural networks. Catena 205, 105466.
Ghorbani, M.A., Zadeh, H.A., Isazadeh, M., Terzi, O., 2016. A comparative study of artificial neural network (MLP, RBF) and support vector machine models for river flow prediction. Environ. Earth Sci. 75, 476.
Gorji, T., Sertel, E., Tanik, A., 2017. Monitoring soil salinity via remote sensing technology under data scarce conditions: a case study from Turkey. Ecol. Indicat. 74, 384–391.
Habibi, V., Ahmadi, H., Jafari, M., Moeini, A., 2021. Mapping soil salinity using a combined spectral and topographical indices with artificial neural network. PLoS One 16, e0228494.
Hao, P.Y., 2017. Pair- ${v}$ -SVR: a novel and efficient pairing nu-support vector regression algorithm. IEEE Transact. Neural Networks Learn. Syst. 28, 2503–2515.
Hartemink, A.E., Zhang, Y., Bockheim, J.G., Curi, N., Silva, S.H.G., Grauer-Gray, J., Lowe, D.J., Krasilnikov, P., 2020. Chapter Three - soil horizon variation: a review. In: Sparks, D.L. (Ed.), Advances in Agronomy. Academic Press, pp. 125–185.
Hassani, A., Azapagic, A., Shokri, N., 2021. Global predictions of primary soil salinization under changing climate in the 21st century. Nat. Commun. 12, 6663.
Herren, A., Hahn, P.R., 2022. Statistical aspects of shap: functional anova for model interpretation. arXiv preprint arXiv:2208.09970.
Herrero, J., Pérez-Coveta, O., 2005. Soil salinity changes over 24 years in a Mediterranean irrigated district. Geoderma 125, 287–308.
Hodson, T.O., 2022. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geosci. Model Dev. (GMD) 15, 5481–5487.
Hopmans, J.W., Qureshi, A.S., Kisekka, I., Munns, R., Grattan, S.R., Rengasamy, P., Ben-Gal, A., Assouline, S., Javaux, M., Minhas, P.S., Raats, P.A.C., Skaggs, T.H., Wang, G., De Jong van Lier, Q., Jiao, H., Lavado, R.S., Lazarovitch, N., Li, B., Taleisnik, E., 2021. Chapter One - critical knowledge gaps and research priorities in global soil salinity. In: Sparks, D.L. (Ed.), Advances in Agronomy. Academic Press, pp. 1–191.
Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Network. 2, 359–366.
Islam Khan, M.S., Islam, N., Uddin, J., Islam, S., Nasir, M.K., 2022. Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. Journal of King Saud University - Computer and Information Sciences 34, 4773–4781.
Jamei, M., Ali, M., Karbasi, M., Karimi, B., Jahannemaei, N., Farooque, A.A., Yaseen, Z. M., 2024. Monthly sodium adsorption ratio forecasting in rivers using a dual interpretable glass-box complementary intelligent system: hybridization of ensemble TVF-EMD-VMD, Boruta-SHAP, and eXplainable GPR. Expert Syst. Appl. 237, 121512.
Jurado, C., Díaz-Vivancos, P., Gregorio, B.-E., Acosta-Motos, J.R., Hernández, J.A., 2024. Effect of halophyte-based management in physiological and biochemical responses of tomato plants under moderately saline greenhouse conditions. Plant Physiol. Biochem. 206, 108228.
Kafei, F., Rezapour, S., Dalalian, M.R., Sabbaghtazeh, E., Rafieyan, O., 2023. Soil quality index as affected by long-time continuous cultivation in a Mediterranean sub-humid region. Rendiconti Lincei. Sci. Fis. Nat. 34, 563–575.
Kamrakji, S.S., Amer, A.-W.M., El-Didy, S.M.A., Tawfik, A.M., 2016. Salt accumulation in irrigated loamy soil; Lower Euphrates Valley, Syria. Water Science 30, 1–9.
Kan, J.-C., Ferreira, C.S.S., Destouni, G., Haozhi, P., Vieira Passos, M., Barquet, K., Kalantari, Z., 2023. Predicting agricultural drought indicators: ML approaches across wide-ranging climate and land use conditions. Ecol. Indicat. 154, 110524.
Kaplan, G., Gašparović, M., Alqasemi, A.S., Aldhaheri, A., Abuelgasim, A., Ibrahim, M., 2023. Soil salinity prediction using machine learning and sentinel – 2 remote sensing data in hyper – arid areas. Phys. Chem. Earth, Parts A/B/C 130, 103400.
Kargas, G., Londra, P.A., Koka, D., Sgoubopoulou, A., 2023. Relationships between saturated paste and 1:1 or 1:5 soil/water extract sodium adsorption ratios. Irrigat. Drain. 72, 503–514.
Kattan, Z., 2020. Factors affecting the chemical composition of precipitation in Syria. Environ. Sci. Pollut. Control Ser. 27, 28408–28428.
Khan, N., Sachindra, D.A., Shahid, S., Ahmed, K., Shiru, M.S., Nawaz, N., 2020. Prediction of droughts over Pakistan using machine learning algorithms. Adv. Water Resour. 139, 103562.
Klopp, H., Bleam, W., 2021. The effects of soil solution electrical conductivity and sodium adsorption ratio on soil liquid limit and soil strength. Commun. Soil Sci. Plant Anal. 52, 2644–2653.
Kushwaha, N.L., Rajput, J., Suna, T., Sena, D.R., Singh, D.K., Mishra, A.K., Sharma, P.K., Mani, I., 2023. Metaheuristic approaches for prediction of water quality indices with relief algorithm-based feature selection. Ecol. Inf. 75, 102122.
Lagacherie, P., Álvaro-Fuentes, J., Annabi, M., Bernoux, M., Bouarfa, S., Douaoui, A., Grünberger, O., Hammani, A., Montanarella, L., Mrabet, R., Sabir, M., Raclot, D., 2018. Managing Mediterranean soil resources under global change: expected trends and mitigation strategies. Reg. Environ. Change 18, 663–675.
Langhammer, J., Česák, J., 2016. Applicability of a nu-support vector regression model for the completion of missing data in hydrological time series. Water 8, 560.
Li, J., He, P., Chen, J., Hamad, A.A.A., Dai, X., Jin, Q., Ding, S., 2023. Tomato performance and changes in soil chemistry in response to salinity and Na/Ca ratio of irrigation water. Agric. Water Manag. 285, 108363.
Li, X., Xiao, Y.H., Huang, L., 2022. Mean Squared Error Analysis of the One-Bit Signal Power Estimator, 2022 5th International Conference on Information Communication and Signal Processing (ICICSP), pp. 651–655.
Lieberman, N.R., Izquierdo, M., Muñoz-Quirós, C., Cohen, H., Chenery, S.R., 2020. Geochemical signature of superhigh organic sulphur Raša coals and the mobility of toxic trace elements from combustion products and polluted soils near the Plomin coal-fired power station in Croatia. Appl. Geochem. 114, 104472.
Liu, X., Zhu, Y., McLean Bennett, J., Wu, L., Li, H., 2022. Effects of sodium adsorption ratio and electrolyte concentration on soil saturated hydraulic conductivity. Geoderma 414, 115772.
Maliva, R., 2021. Groundwater related impacts of climate change on infrastructure. In: Maliva, R. (Ed.), Climate Change and Groundwater: Planning and Adaptations for a Changing and Uncertain Future: WSP Methods in Water Resources Evaluation Series No. 6. Springer International Publishing, Cham, pp. 177–195.
Mandal, A.K., 2019. Modern technologies for diagnosis and prognosis of salt-affected soils and poor-quality waters. In: Dagar, J.C., Yadav, R.K., Sharma, P.C. (Eds.), Research Developments in Saline Agriculture. Springer Singapore, Singapore, pp. 95–152.
Mandal, S., Mondal, S., 2019. Artificial neural network (ANN) model and landslide susceptibility. In: Mandal, S., Mondal, S. (Eds.), Statistical Approaches for Landslide Susceptibility Assessment and Prediction. Springer International Publishing, Cham, pp. 123–133.
Mantena, S., Mahammood, V., Rao, K.N., 2023. Prediction of soil salinity in the Upputeru river estuary catchment, India, using machine learning techniques. Environ. Monit. Assess. 195, 1006.
Marien, L., Crabit, A., Dewandel, B., Ladouche, B., Fleury, P., Follain, S., Cavero, J., Berteloot, V., Colin, F., 2023. Salinity spatial patterns in Mediterranean coastal areas: the legacy of historical water infrastructures. Sci. Total Environ. 899, 165730.
Metternicht, G.I., Zinck, J.A., 2003. Remote sensing of soil salinity: potentials and constraints. Remote Sensing of Environment 85, 1–20.
Mohamed, S.A., Metwaly, M.M., Metwalli, M.R., AbdelRahman, M.A.E., Badreldin, N., 2023. Integrating active and passive remote sensing data for mapping soil salinity using machine learning and feature selection approaches in arid regions. Rem. Sens. 15, 1751.
Mohammed, S., Arshad, S., Alsilibe, F., Moazzam, M.F.U., Bashir, B., Prodhan, F.A., Alsalman, A., Vad, A., Ratonyi, T., Harsányi, E., 2024a. Utilizing machine learning and CMIP6 projections for short-term agricultural drought monitoring in central Europe (1900–2100). J. Hydrol. 633, 130968.
Mohammed, S., Arshad, S., Bashir, B., Vad, A., Alsalman, A., Harsányi, E., 2024b. Machine learning driven forecasts of agricultural water quality from rainfall ionic characteristics in Central Europe. Agric. Water Manag. 293, 108690.
Mohammed, S., Habib, H., Ali, H., AlHennaw, S., Kiwan, S., Ghanem, S., Alsafadi, K., Brevik, E.C., Sulieman, M.M., Harsányi, E., 2020a. Soils of the Southern Syria – a big database for the future land management planning. Data Brief 31, 105832.
Mohammed, S., Jouhra, A., Enaruvbe, G.O., Bashir, B., Barakat, M., Alsilibe, F., Cimusa Kulimushi, L., Alsalman, A., Szabó, S., 2023. Performance evaluation of machine learning algorithms to assess soil erosion in Mediterranean farmland: a case-study in Syria. Land Degrad. Dev. 34, 2896–2911.
Mohammed, S., Khallouf, A., Kiwan, S., Alhenawi, S., Ali, H., Harsányi, E., Kátai, J., Habib, H., 2020b. Characterization of major soil orders in Syria. Eurasian Soil Sci. 53, 420–429.
Montazeri, A., Mazaheri, M., Morid, S., Mosaddeghi, M.R., 2023. Effects of upstream activities of tigris-euphrates river basin on water and soil resources of shatt al-arab border river. Sci. Total Environ. 858, 159751.
Navarro-Torre, S., Ferrario, S., Caperta, A.D., Victorino, G., Bailly, M., Sousa, V., Viegas, W., Nogales, A., 2023. Halotolerant endophytes promote grapevine regrowth after salt-induced defoliation. J. Plant Interact. 18, 2215235.
Noguchi, K., Abel, R.S., Marmolejo-Ramos, F., Konietschke, F., 2020. Nonparametric multiple comparisons. Behav. Res. Methods 52, 489–502.
Núñez, M., Finkbeiner, M., 2020. A regionalised life cycle assessment model to globally assess the environmental implications of soil salinization in irrigated agriculture. Environmental Science & Technology 54, 3082–3090.
Okur, B., Örçen, N., 2020. Chapter 12 - soil salinization and climate change. In: Prasad, M.N.V., Pietrzykowski, M. (Eds.), Climate Change and Soil Interactions. Elsevier, pp. 331–350.
Onyekwena, C.C., Xue, Q., Li, Q., Wan, Y., Feng, S., Umeobi, H.I., Liu, H., Chen, B., 2022. Support vector machine regression to predict gas diffusion coefficient of biochar-amended soil. Appl. Soft Comput. 127, 109345.
Otchere, D.A., Ganat, T.O.A., Ojero, J.O., Tackie-Otoo, B.N., Taki, M.Y., 2022. Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. J. Petrol. Sci. Eng. 208, 109244.
Pankova, E.I., Aidarov, I.P., Golovanov, D.L., Yamnova, I.A., 2015. Salinization as the main soil-forming process in soils of natural oases in the Gobi desert. Eurasian Soil Sci. 48, 1017–1028.
Piepho, H.-P., 2023. An adjusted coefficient of determination (R2) for generalized linear mixed models in one go. Biom. J. 65, 2200290.
Prots, A., Schlüter, L., Voigt, M., Mailach, R., Meyer, M., 2023. Robust Sensitivity Analysis of Complex Simulation Models Subject to Noise. AIAA SCITECH. Forum.
Rengasamy, P., 2006. World salinization with emphasis on Australia. J. Exp. Bot. 57, 1017–1023.
Richardson, N.E., Konon, E.N., Schuster, H.S., Mitchell, A.T., Boyle, C., Rodgers, A.C., Finke, M., Haider, L.R., Yu, D., Flores, V., Pak, H.H., Ahmad, S., Ahmed, S., Radcliff, A., Wu, J., Williams, E.M., Abdi, L., Sherman, D.S., Hacker, T.A., Lamming, D.W., 2021. Lifelong restriction of dietary branched-chain amino acids has sex-specific benefits for frailty and life span in mice. Nature Aging 1, 73–86.
Sajib, A.M., Diganta, M.T.M., Moniruzzaman, M., Rahman, A., Dabrowski, T., Uddin, M. G., Olbert, A.I., 2024. Assessing water quality of an ecologically critical urban canal incorporating machine learning approaches. Ecol. Inf. 80, 102514.

Šajn, R., Stafilov, T., Balabanova, B., Alijagić, J., 2022. Multi-Scale application of advanced ANN-MLP model for increasing the large-scale improvement of digital data visualisation due to anomalous lithogenic and anthropogenic elements distribution. Minerals 12, 174.
Salem, S.B.H., Gaagai, A., Ben Slimene, I., Moussa, A.B., Zouari, K., Yadav, K.K., Eid, M. H., Abukhadra, M.R., El-Sherbeeny, A.M., Gad, M., Farouk, M., Elsherbiny, O., Elsayed, S., Bellucci, S., Ibrahim, H., 2023. Applying multivariate analysis and machine learning approaches to evaluating groundwater quality on the kairouan plain, Tunisia. Water 15, 3495.
Salvato, L.A., Pittelkow, C.M., O'Geen, A.T., Linquist, B.A., 2024. A geospatial assessment of soil properties to identify the potential for crop rotation in rice systems. Agric. Ecosyst. Environ. 359, 108753.
Sarani, F., Ahangar, A.G., Shabani, A., 2016. Predicting ESP and SAR by artificial neural network and regression models using soil pH and EC data (Miankangi Region, Sistan and Baluchestan Province, Iran). Arch. Agron Soil Sci. 62, 127–138.
Sarkar, S.K., Rudra, R.R., Sohan, A.R., Das, P.C., Ekram, K.M.M., Talukdar, S., Rahman, A., Alam, E., Islam, M.K., Islam, A.R.M.T., 2023. Coupling of machine learning and remote sensing for soil salinity mapping in coastal area of Bangladesh. Sci. Rep. 13, 17056.
Sattari, M.T., Feizi, H., Colak, M.S., Ozturk, A., Apaydin, H., Ozturk, F., 2020. Estimation of sodium adsorption ratio in a river with kernel-based and decision-tree models. Environ. Monit. Assess. 192, 575.
Schölkopf, B., Bartlett, P., Smola, A., Williamson, R., 1998. Support vector regression with automatic accuracy control. In: Niklasson, L., Bodén, M., Ziemke, T. (Eds.), ICANN 98. Springer, London, London, pp. 111–116.
Shaaban, M., Wu, Y., Núñez-Delgado, A., Kuzyakov, Y., Peng, Q.-A., Lin, S., Hu, R., 2023. Enzyme activities and organic matter mineralization in response to application of gypsum, manure and rice straw in saline and sodic soils. Environ. Res. 224, 115393.
Shadkani, S., Abbaspour, A., Samadianfard, S., Hashemi, S., Mosavi, A., Band, S.S., 2021. Comparative study of multilayer perceptron-stochastic gradient descent and gradient boosted trees for predicting daily suspended sediment load: the case study of the Mississippi River, U.S. Int. J. Sediment Res. 36, 512–523.
Shahid, S.A., Zaman, M., Heng, L., 2018. Soil salinity: historical perspectives and a world overview of the problem. In: Zaman, M., Shahid, S.A., Heng, L. (Eds.), Guideline for Salinity Assessment, Mitigation and Adaptation Using Nuclear and Related Techniques. Springer International Publishing, Cham, pp. 43–53.
Singh, A., 2021. Soil salinization management for sustainable development: a review. J. Environ. Manag. 277, 111383.
Singh, A., Mehrotra, R., Rajput, V.D., Dmitriev, P., Singh, A.K., Kumar, P., Tomar, R.S., Singh, O., Singh, A.K., 2022. Geoinformatics, artificial intelligence, sensor technology, big data. Sustainable Agriculture Systems and Technologies, pp. 295–313.
Sobhi Gollo, V., González, E., Elbracht, J., Fröhle, P., Shokri, N., 2023. Impact of Soil Texture and Heterogeneity on Complex Interactions between Surface Soil Salinity and Saltwater Intrusion in Coastal Regions, pp. EGU–2285.
Taha, O.M.E., Majeed, Z.H., Ahmed, S.M., 2018. Artificial neural network prediction models for maximum dry density and optimum moisture content of stabilized soils. Transportation Infrastructure Geotechnology 5, 146–168.
Tang, Y., Wang, Z., Zhang, T., 2023. Soil salinity estimation in Shule River Basin using support vector regression model. Land Degrad. Dev. 34, 4094–4108.
Tedeschi, A., 2020. Irrigated agriculture on saline soils: a perspective. Agronomy 10, 1630.
Tizpa, P., Jamshidi Chenari, R., Karimpour Fard, M., Lemos Machado, S., 2015. ANN prediction of some geotechnical properties of soil from their index parameters. Arabian J. Geosci. 8, 2911–2920.
Tofallis, C., 2015. A better measure of relative prediction accuracy for model selection and model estimation. J. Oper. Res. Soc. 66, 1352–1362.
Tomaz, A., Palma, P., Alvarenga, P., Gonçalves, M.C., 2020. Chapter 13 - soil salinity risk in a climate change scenario and its effect on crop yield. In: Prasad, M.N.V., Pietrzykowski, M. (Eds.), Climate Change and Soil Interactions. Elsevier, pp. 351–396.
Tran, D.A., Tsujimura, M., Ha, N.T., Nguyen, V.T., Binh, D.V., Dang, T.D., Doan, Q.-V., Bui, D.T., Anh Ngoc, T., Phu, L.V., Thuc, P.T.B., Pham, T.D., 2021. Evaluating the predictive power of different machine learning algorithms for groundwater salinity prediction of multi-layer coastal aquifers in the Mekong Delta, Vietnam. Ecol. Indicat. 127, 107790.
Uddin, M.G., Imran, M.H., Sajib, A.M., Hasan, M.A., Diganta, M.T.M., Dabrowski, T., Olbert, A.I., Moniruzzaman, M., 2024. Assessment of human health risk from potentially toxic elements and predicting groundwater contamination using machine learning approaches. J. Contam. Hydrol. 261, 104307.
Uddin, M.G., Rahman, A., Nash, S., Diganta, M.T.M., Sajib, A.M., Moniruzzaman, M., Olbert, A.I., 2023. Marine waters assessment using improved water quality model incorporating machine learning approaches. J. Environ. Manag. 344, 118368.
Vapnik, V.N., 1997. The support vector method. In: Gerstner, W., Germond, A., Hasler, M., Nicoud, J.-D. (Eds.), Artificial Neural Networks — ICANN'97. Springer, Berlin Heidelberg, Berlin, Heidelberg, pp. 261–271.
Vega García, M., Aznarte, J.L., 2020. Shapley additive explanations for NO2 forecasting. Ecol. Inf. 56, 101039.
Verma, G., Kumar, B., 2021. Multi-layer perceptron (MLP) neural network for predicting the modified compaction parameters of coarse-grained and fine-grained soils. Innovative Infrastructure Solutions 7, 78.
Wang, F., Shi, Z., Biswas, A., Yang, S., Ding, J., 2020a. Multi-algorithm comparison for predicting soil salinity. Geoderma 365, 114211.
Wang, H., Yilihamu, Q., Yuan, M., Bai, H., Xu, H., Wu, J., 2020b. Prediction models of soil heavy metal(loid)s concentration for agricultural land in Dongli: a comparison of regression and random forest. Ecol. Indicat. 119, 106801.
Wang, H., Zheng, C., Ning, S., Cao, C., Li, K., Dang, H., Wu, Y., Zhang, J., 2023a. Impacts of long-term saline water irrigation on soil properties and crop yields under maize-wheat crop rotation. Agric. Water Manag. 286, 108383.
Wang, L., Hu, P., Zheng, H., Liu, Y., Cao, X., Hellwich, O., Liu, T., Luo, G., Bao, A., Chen, X., 2023b. Integrative modeling of heterogeneous soil salinity using sparse ground samples and remote sensing images. Geoderma 430, 116321.
Wang, Q., Huo, Z., Zhang, L., Wang, J., Zhao, Y., 2016. Impact of saline water irrigation on water use efficiency and soil salt accumulation for spring maize in arid regions of China. Agric. Water Manag. 163, 125–138.
Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemometr. Intell. Lab. Syst. 2, 37–52.
Wuyun, D., Bao, J., Crusiol, L.G.T., Wulan, T., Sun, L., Wu, S., Xin, Q., Sun, Z., Chen, R., Peng, J., Xu, H., Wu, N., Hou, A., Wu, L., Ren, T., 2022. Generating salt-affected irrigated cropland map in an arid and semi-arid region using multi-sensor remote sensing data. Rem. Sens. 14, 6010.
Xiao, C., Ji, Q., Chen, J., Zhang, F., Li, Y., Fan, J., Hou, X., Yan, F., Wang, H., 2023. Prediction of soil salinity parameters using machine learning models in an arid region of northwest China. Comput. Electron. Agric. 204, 107512.
Yang, H., Wang, P., Chen, A., Ye, Y., Chen, Q., Cui, R., Zhang, D., 2023. Prediction of phosphorus concentrations in shallow groundwater in intensive agricultural regions based on machine learning. Chemosphere 313, 137623.
Zarei, A., Hasanlou, M., Mahdianpari, M., 2021. A comparison of machine learning models for soil salinity estimation using multi-spectral earth observation data. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. V-3–2021, 257–263.
Zeraatpisheh, M., Jafari, A., Bagheri Bodaghabadi, M., Ayoubi, S., Taghizadeh-Mehrjardi, R., Toomanian, N., Kerry, R., Xu, M., 2020. Conventional and digital soil mapping in Iran: past, present, and future. Catena 188, 104424.
Zhang, W., Ashraf, W.M., Senadheera, S.S., Alessi, D.S., Tack, F.M.G., Ok, Y.S., 2023. Machine learning based prediction and experimental validation of arsenite and arsenate sorption on biochars. Sci. Total Environ. 904, 166678.
Zhou, Y., Chen, S., Hu, B., Ji, W., Li, S., Hong, Y., Xu, H., Wang, N., Xue, J., Zhang, X., Xiao, Y., Shi, Z., 2022. Global soil salinity prediction by open soil vis-NIR spectral library. Rem. Sens. 14, 5627.