

Received February 20, 2021, accepted March 29, 2021, date of publication April 5, 2021, date of current version April 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3071015

Total Dissolved Salt Prediction Using Neurocomputing Models: Case Study of Gypsum Soil Within Iraq Region

NEERAJ DHANRAJ BOKDE¹, ZAINAB HASAN ALI², MAYSAM TH. AL-HADIDI³,
AITAZAZ AHSAN FAROOQUE⁴, MEHDI JAMEI⁵,
ALI ABDULRIDHA AL MALIKI⁶, (Member, IEEE), BESTE HAMIYE BEYAZTAS⁷,
HOSSAM FARIS^{8,9}, AND ZAHER MUNDHER YASEEN¹⁰

¹Department of Mechanical and Production Engineering, Aarhus University, 8000 Aarhus, Denmark

²Civil Engineering Department, College of Engineering, University of Diyala, Baquba 32001, Iraq

³Water Resources Engineering Department, University of Baghdad, Baghdad 64900, Iraq

⁴Faculty of Sustainable Design Engineering, University of Prince Edward Island, Charlottetown, PE C1A4P3, Canada

⁵Faculty of Engineering, Shohadaye Hoveizeh University of Technology, Dasht-e Azadegan, Susangerd 6155634899, Iran

⁶Environment and Water Directorate, Ministry of Science and Technology, Baghdad 765, Iraq

⁷Department of Statistics, Istanbul Medeniyet University, 34000 Istanbul, Turkey

⁸King Abdullah II School for Information Technology, The University of Jordan, Amman 11942, Jordan

⁹School of Computing and Informatics, Al Hussein Technical University, Amman 11831, Jordan

¹⁰New Era and Development in Civil Engineering Research Group, Scientific Research Center, Al-Ayen University, Thi-Qar 64001, Iraq

Corresponding author: Zaher Mundher Yaseen (zaheryaseen88@gmail.com)

ABSTRACT Quantification of the soil physicochemical properties is one of the essential process in the field of soil geo-science. In the current research, three types of machine learning (ML) models including support vector machine (SVM), random forest (RF), and gradient boosted decision tree (GBDT) were developed for Total Dissolved Salt (TDS) prediction over several locations in Iraq region. Various physicochemical soil properties were used as predictors for the TDS prediction. Four modeling scenarios are constructed based on the types of the associated soil input variables properties. The applied ML models were analyzed and discussed based on several statistical measures and graphical presentations. Based on the correlation analysis; Gypsum concentration, Sulfur trioxide (SO_3), Chloride (Cl), and organic matter (OR) were the essential soil properties for the TDS concentration influence. The prediction results indicated that incorporating all the types of input variables including chemical, soil consistency limits, and soil sieve analysis attained the best prediction process. In quantitative terms, the SVM model attained the maximum coefficient of determination ($R^2 = 0.849$) and minimum root mean square error (RMSE=3.882). Overall, the development of the ML models for the TDS of soil prediction provided a robust and reliable methodology that contributes to the soil geoscience field.

INDEX TERMS Soil physicochemical properties, total dissolved salt, machine learning, correlation analysis.

I. INTRODUCTION

Being a heterogeneous natural resource, soil exhibits mechanisms and processes that are difficult and complex to comprehend. Until, now, laboratory analysis is the major technique that has been used for the understanding of soil systems and assessing their quality [1]. It is important that accurate soil information should be available at both national and regional levels as it facilitates better soil management in line with the potential of the land [2], [3]. Researchers can have

a better understanding of the ecosystem dynamics via the spatial assessment of soil properties [4], [5]. In addition, the understanding of the soil properties and their influence on agriculture can support better environmental management and sustainable agricultural implementation [6], [7]. Precision agriculture requires much less level of soil information for effective crop and land management as it normally depends on proximal soil sensing for large-scale data collection [8], [9]. Soil spatial variability characterization is undoubtedly one of the most significant aspects of soil research owing to its wide range of influence on the environment and agriculture-related field-to-landscape-scale

The associate editor coordinating the review of this manuscript and approving it for publication was Zijian Zhang¹¹.

processes, such as solute transport, soil salinity distribution, intra-field crop yield variation, etc [10], [11]. The significance of soil spatial variability on the practical agricultural and environmental applications that involves soil at both field and larger spatial scales cannot be ignored because soil, in its chemical and physical composition, is spatially variable.

It has been estimated that about 23% of the cultivated land (approximately 3.5×10^8 ha) is affected by salt [12]. Unfortunately, there are no directly measured global inventories of soil salinity as all the currently known soil salinity global inventories, except a few, are mere gross approximations that are based on qualitative data [13], [14]. Field assessment of soil salinity is a difficult task due to its dynamic nature and high variability in space and time. Several techniques are available for soil salinity assessment; however, the laboratory-based approach of soil salinity determination involves 3 different approaches which are the determination of the mass of total dissolved solids (TDS, mgL^{-1}), measurement of the soil electrical conductivity (EC, dSm^{-1}), and measurement of salt species composition using spectrophotometry [15].

To achieve a precise calibration that will enable the surveying of a site via conductivity measurement requires measurement of the TDS [16]. The TDS refers to the total quantity of soluble salts and small quantities of organic matter in a soil-saturated paste extract [17]. It is normally presented in milligrams per liter (mgL^{-1}) or parts per million (ppm) [18]. Soil characterization via measuring and assessing its properties and components is a costly and time-consuming process [19]. Hence, modeling results are normally used to compensate for the lack of sampling data. Several modeling approaches (called predictive soil mapping) have been developed for the estimation of the spatial distribution of soil variables and most of these approaches rely on the statistical or numerical methodology of the relationship between the soil properties and other environmental variables. They are applied to geographic databases to establish a predictive map or to derive the values of soil properties at unmeasured sites based on data collected from the field [20], [21].

Soil characteristics measurement focuses more on direct estimation of the unknown soil variables based on the values of other known measured variables. Being that it is labor and cost-intensive to rely on laboratory tests for determining the physical and chemical soil properties [22], it becomes economically reasonable to devise other technologies for estimating the soil variables based on known variables. Various methods have been introduced for soil mapping and classification, such as statistical techniques (like principal components regression (PCR) [23]–[25] and partial least squares regression (PLSR) [25], [26]. However, the development of cheap and fast microprocessors has increased the rate of using sophisticated advanced computed aid and statistical methods in various areas of environmental and geo-sciences [27]–[29].

The use of ML models for soil properties modeling has received much attention from several researchers.

The capability of SVM in soil moisture estimation based on remote sensing data was reported by Ahmad *et al.* [30]. The study confirmed the ability of the SVM model to capture the variability in the measured soil moisture. Rossel and Behrens [31] assessed the performance of different ML models in estimating soil organic carbon (SOC), clay content (CC), and pH. The evaluated models including multiple linear regression (MLR), multivariate adaptive regression splines (MARS), PLSR, RF, artificial neural network (ANN), SVM, and boosted trees (BT) models. Based on the research findings, the SVM model performed the best in estimating the three soil properties as it produced the smallest error values in all its VIS–NIR wavelengths, then followed by MARS and PLSR. The performance of the PCR, PLSR, and ANN models was investigated for analyzing the prediction accuracy of organic carbon (OC) and extractable forms of potassium (K), phosphorus (P), sodium (Na), and magnesium (Mg) [25]. The ANN model had the first 5 principal components (PCs) that resulted from the principal component analysis (PCA) and the optimal number of latent variables (LVs) from the PLSR as its input variable. From the results, all the ANN-LV models performed better than the PCR, PLSR, and ANN-PCs models. The development of an SVM model for dynamic soil electrical conductivity (EC) prediction has been reported by Guan *et al.* [32]. The performance of the developed model was validated against the ANN model and the results showed that the SVM model gave a better prediction of the soil EC values compared to the ANN model. ANN and PLSR models were compared by Kuang *et al.* [33] for Vis-NIR spectrophotometer calibration. The calibrated equipment is intended for on-line soil OC, pH, and CC measurements in two fields on a Danish farm. From the outcome of the on-line independent validation, the performance of the ANN model was better than that of PLSR in both fields. The study by Brungard *et al.* [34] focused on the validation of several ML models for soil taxonomic class prediction at three different geographical areas in the semi-arid area of the western United States of America. The validated models included clustering algorithms, multinomial logistic regression, discriminant analysis, neural networks, SVM, and tree-based classifiers. The results pointed towards better accuracy of the complex models compared to the simple or moderately complex models. The feasibility of using digital elevation models (DEMs) derivatives and ML models (k-nearest neighbor, SVM, decision tree (DT), and RF) for the prediction of the location and extent of salt-affected areas within the Vaalharts and Breede River irrigation schemes of South Africa was evaluated by Vermeulen and Van Niekerk [35]. The outcome of the study showed that relying on the elevation data and its derivatives as input to ML and geostatistics serves better in monitoring the level of salt accumulation in the irrigated areas, especially for simulation of sub-surface conditions. Wu *et al.* [36] reported the use of SVM and RF models for salinity prediction based on a combined dataset. The outcome of the study showed better performance of the RF model compared to SVM in terms of accuracy and less normalized root mean square error upon

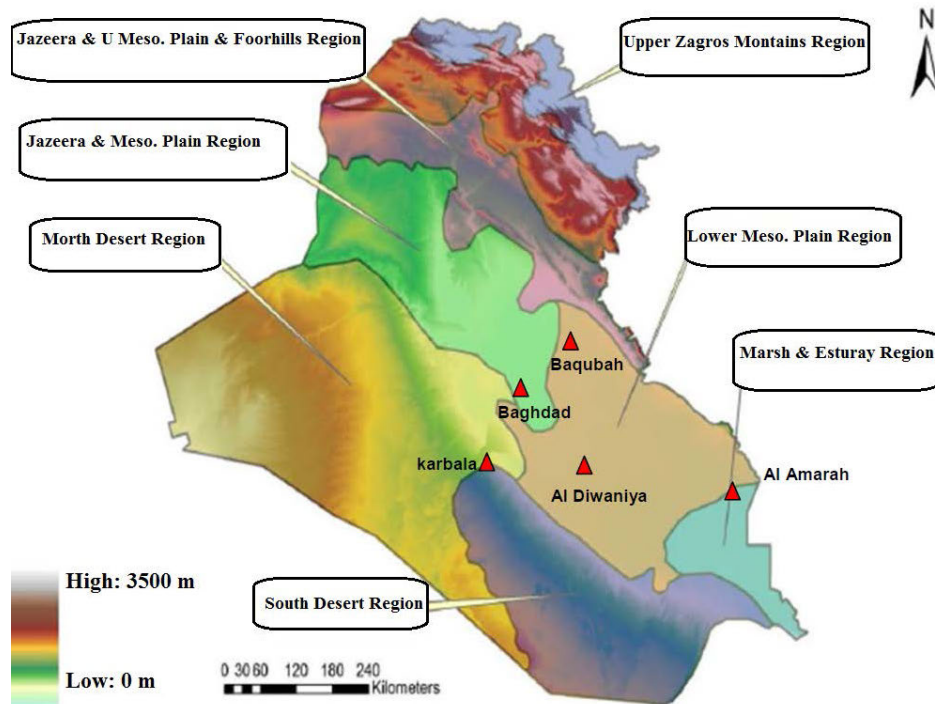


FIGURE 1. Map of Iraq showing the Physiographic regions of Iraq and study locations modified from [44].

calibration with the training and validation datasets. Another study by Wu *et al.* [37] reported the use of regressions with RF and XGBoost models for establishing the relationships between environmental variables and soil properties. The proposed procedure was shown to perform an accurate mapping of the other soil properties. Other numerous studies have demonstrated the feasibility of using ML models in the field of geosciences [38]–[43]. Although there have been several researches conducted on soil properties determination, there are very few researches adopted on the soil TDS prediction. Proposing a computer aid model methodologies for soil physicochemical properties determination can contribute to diverse soil and geoscience engineering aspects.

To the best knowledge of the current research, three different ML models (e.g., SVM, RF, and GBDT) were developed for soil total dissolved salt prediction in the middle region of Iraq. Datasets of several physicochemical soil properties were used to construct the proposed predictive models in the form of four different scenarios. The modeling results were assessed and evaluated statistically and graphically and numerous aspects were concluded based on the attained modeling results. A list of abbreviations used in this paper is tabulated in Table 1.

II. CASE STUDY LOCATION AND DATASET

The used dataset in the current research was collected from five locations in the Iraq region. The locations are represented five provinces within the Iraq country included Baghdad,

Baqubah, Karbala, Diwaniya, and Al Amarah (Figure 1) that are located in central and southern parts of the Mesopotamian plain. The climate of this region is warm and dry with annual rainfall between 100 and 500 mm, while the annual temperature is more than 22 °C.

The topographical map of the Iraq region is usually divided into four physiographic regions. The Mesopotamian plain is one of the main parts of it in addition to the Mountain Region, the Undulating Region, and the Desert Region, each region has specific climatological conditions and different geological and hydrological properties Figure 1.

Except for the mountain region, the properties of Iraq soil are characterized by dry climatic conditions which caused a weak level of soil development. It is represented by the type of diagnostic horizons and differences in morphological, physical, chemical, and mineralogical properties with a decrease of soil development from northern to southern Iraq [44]. Mesopotamian, in the modern period, refers to all the lands between the Euphrates and the Tigris rivers and trends in the same direction of those rivers from northwest to southeast orientation [45]. Geologically, the Mesopotamian zone is almost flat and vast lowland located at the southern part of the extensive geosynclines, more details for geological units of the Plain are described in [46].

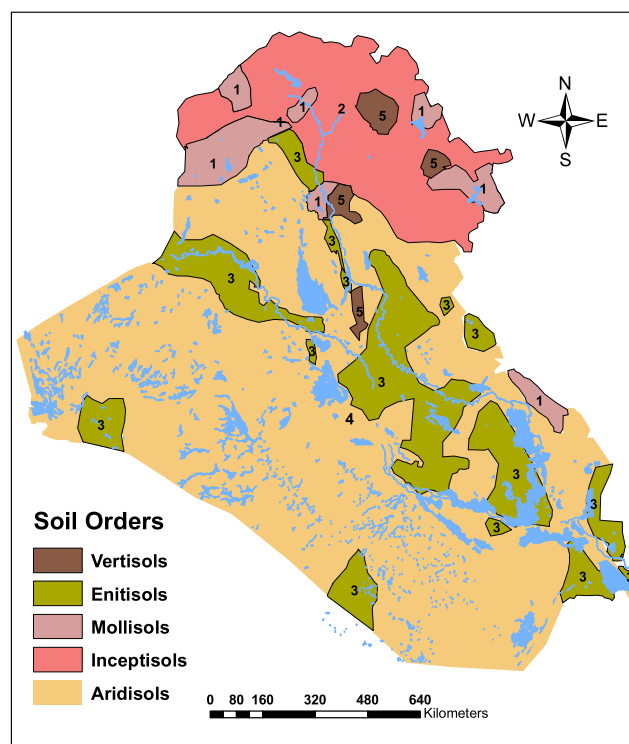
Several stratigraphic studies confirm that the Mesopotamian Plain covered by the floodplain sediments during the Quaternary and Recent geological periods [44], [47]. According to the literature, five soil

TABLE 1. A list of abbreviations used in this paper.

Abbreviations			
ANN	Artificial Neural Network	OC	Organic Carbon
BT	Boosted Trees	OOB	Out of bag
CC	Clay Content	OR	Organic matter
Cl	Chloride	P	Phosphorus
DEM	Digital Elevation Model	PC	Principal Component
DT	Decision Tree	PCA	Principal Component Analysis
EC	Electrical Conductivity	PCR	Principal Component Regression
ERM	Empirical Risk Minimization	PI	Plasticity index
GBDT	Gradient Boosted Decision Tree	PL	Plastic Limit
G _s	Specific gravity	PLSR	Partial Least Squares Regression
K	Potassium	ppm	parts per million
LL	Liquid Limit	R ²	R-squared
LV	Latent Variable	RF	Random Forest
MAE	Mean Absolute Error	RMSE	Root Mean Square Error
MAPE	Mean Absolute Percentage Error	SO ₃	Sulfur-trioxide
MARS	Multivariate Adaptive Regression Splines	SOC	Soil Organic Carbon
ME	Mean Error	SRM	Structural Risk Minimization
Mg	Magnesium	SVM	Support Vector Machine
ML	Machine Learning	TDS	Total Dissolved Salt
MLR	Multiple Linear Regression	VC	Vapnik–Chervonenkis
MPE	Mean Percentage Error	W _n	Water Content
Na	Sodium		

orders were recognized in Iraq: Aridisols Entisols Inceptisols Mollisols, and Vertisols. These orders arranged based on their percentage (62.2 %), (16.2%), (12.6%), (3.8%), and (1.2%), respectively [44] Figure 2. In addition, the most common accumulation within the Mesopotamia Plain that formed of flood plains from the fluvial origin, Aeolian sediments were also deposited which is accumulated and mixed with fluvial deposits [48]. These accumulation sediments are composed of an alternation of clay, silty clay, clayey silt, silt, sand, and gravel [44].

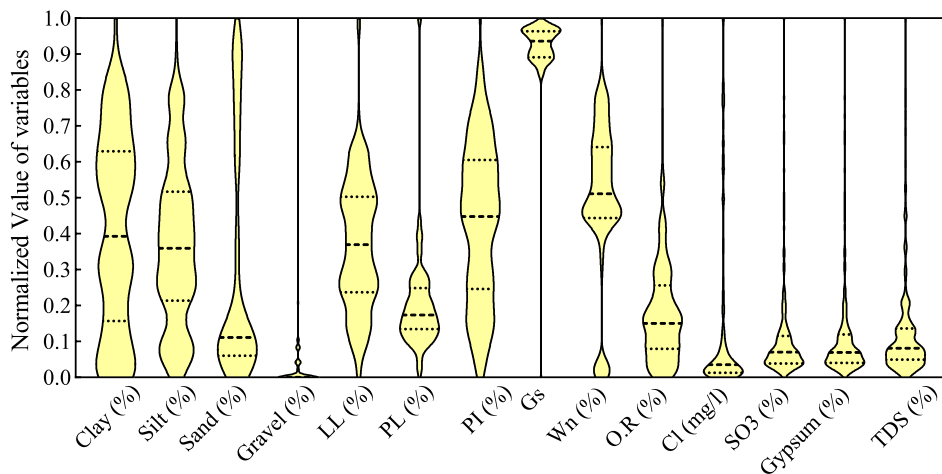
As the arid regions of Asia, the soil types of the Mesopotamian plain are dominated by Aridisols and Entisols. Because of an extreme imbalance between the rainfall and potential evapotranspiration patterns, arid soils show variations with the common soil properties with different diagnostic horizons. The common subsurface horizons are related to the accumulation of different types of salts (calcium carbonates, gypsum, as well as sodium). The accumulation of salts is represented as a serious issue to agricultural land use. Although Iraq's soil salinity problem has existed since ancient days, the floods cleared the salts and deposited a new layer of silt and clay and thus natural processes have assisted keep agriculture growing. However, the disruption of floods in recent times and old irrigation processes in Iraq helped increase salinity again [49]. In this study, different types of datasets were used to determine the TDS magnitude including Chemical properties such as Gypsum, SO₃, Cl, OR, specific gravity (G_s), and water content (W_n). In addition, soil consistency limits (shrinkage, plastic, and liquid), plastic limit (PL), liquid limit (LL), plasticity index (PI). Furthermore, soil sieve analysis including gravel, sand, silt, and clay. A total of 211 observations tests were used to build the applied predictive models. The statistical characteristics of the dataset used in the study and corresponding normalized

**FIGURE 2.** Map of Iraq showing the distribution of soil orders in Iraq with study locations modified from [44].

values are shown in Table 2 and Figure 3, respectively. Figure (Figure 4) demonstrates the histogram and distribution density function for all implemented variables. It can be seen that the gravel, G_s , PI, SO₃, have the high kurtosis (>2) values and their distribution is far from the normal distribution.

TABLE 2. Statistical characteristics of the dataset used in the study.

	Minimum	Maximum	Mean	Std. Deviation	Skewness	Kurtosis
Clay	0	89	36.24	24.28	0.01879	-1.208
Silt	0	89	33.01	20.02	0.3286	-0.6196
Sand	0	98.75	30.28	32.47	0.9188	-0.8424
Gravel	0	24	0.757	2.174	7.117	66.82
LL	10	85.8	37.99	13.91	0.4253	0.3508
PL	13	64.9	23.33	7.058	3.565	18.51
PI	2	40	18.48	8.325	-0.03103	-0.8318
Gs	1.71	2.81	2.727	0.08305	-8.84	107.4
Wn (%)	0.52	46.68	23.96	9.665	-0.9962	0.89
O.R (%)	0.2	11.5	2.258	1.686	1.685	4.789
Cl (mg/l)	30	9110	1409	2295	1.938	2.466
SO ₃ (%)	0.11	18	2.109	2.501	3.404	13.68
Gypsum (%)	0.2	38.7	4.51	5.378	3.411	13.72
TDS	1.12	39.1	5.864	5.554	3.113	11.51

**FIGURE 3.** Normalized values of the datasets used in the study.

III. MACHINE LEARNING MODELS DESCRIPTION

A. SUPPORT VECTOR MACHINE

Boser *et al.* [50] pioneered the SVMs based on the statistical learning theory; it was proposed to rely on the structural risk minimization (SRM) principle which provides that the capacity concept of learning machines impacts their generalization ability more than just the feature space dimensionality of the number of free variables of the loss function. Reliance on the SRM has consistently offered better solutions compared to the principle of empirical risk minimization (ERM) that was presented by Vapnik *et al.* [51]. This process of constructing SVMs for regression tasks is briefly introduced in this section. The block diagram of the SVM method is shown in Figure 5 when using SVMs for estimating regression functions, 3 distinct characteristics are to be considered; the first one is that the estimation of regression functions based on SVMs rely on a set of linear functions which are normally defined in a high dimensional space. The second characteristic is that SVMs rely on risk minimization for estimating regression and the measurement of the risk is based on the insensitive loss function proposed by Vapnik. The third one is that the risk function used in SVMs consists of a

regularization term and the empirical error; the regularization term is derived based on the SRM concept. The generalization capability of SVM can be improved by ensuring the minimization of the sum of the errors of the training set and the term that depends on the Vapnik–Chervonenkis (VC) dimension. This work utilized SVM as a regression method with the introduction of an ϵ -insensitive loss function ($L_\epsilon(y)$) which is described as follows:

$$L_\epsilon(y) = \begin{cases} 0, & |f(x) - y| < \epsilon \\ |f(x) - y| - \epsilon, & \text{else} \end{cases} \quad (1)$$

This reflects an ϵ tube wherein the loss is zero if the predicted value lies within the tube, while the loss becomes the magnitude of the variation of the predicted value from the radius, ϵ , of the tube if the predicted point lies outside the tube. Consider a training dataset that is comprised of one training sample $(x_1, y_1), \dots, (x_1, y_1)$, where x and y are the input and output; the learning problem will then become the selection of a function that will achieve an optimal prediction of the actual response y in a manner that it is as close as possible to the precision of ϵ . Consider the linear

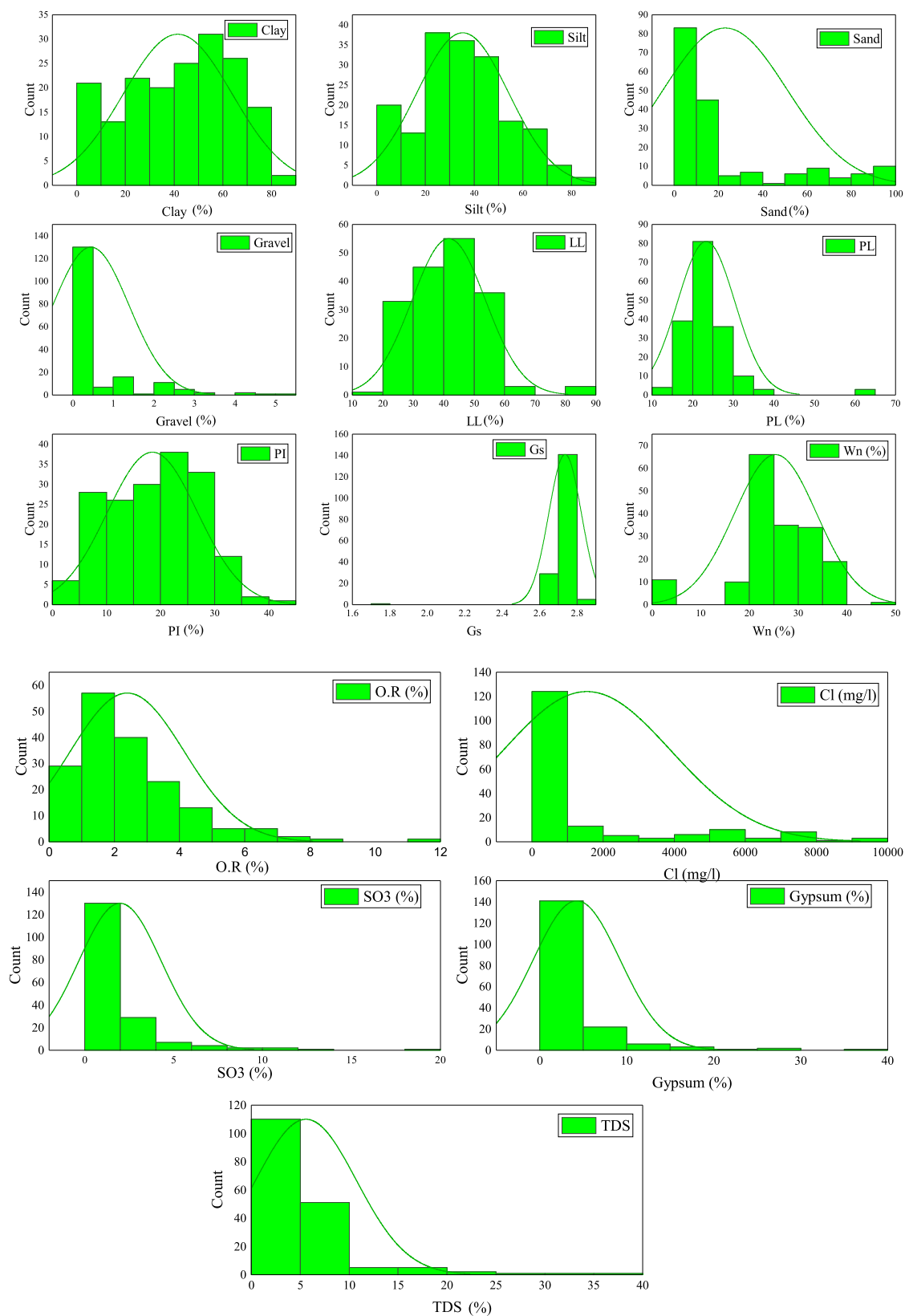


FIGURE 4. Histograms of the datasets used in the study.

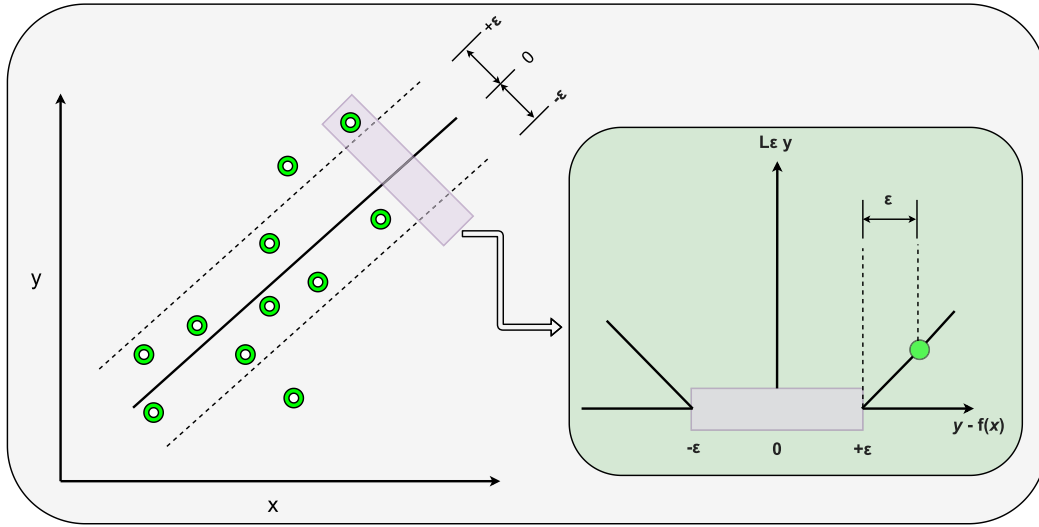


FIGURE 5. Methodology of SVM.

function:

$$f(x) = (w \cdot x) + b \quad (2)$$

where, $w \in R^n$ and $b \in r$; w is an adjustable weight vector while b is the scalar threshold; the n -dimensional vector space is represented by R^n and r is the one-dimensional vector space representation. The aim of the ERM theory is to ensure the minimization of the empirical risk function as defined by the loss function; it is given as:

$$R_{emp}(W) = \frac{1}{l} \sum_{i=1}^l (y_i - f(w \cdot x_i))^2 \quad (3)$$

The construction of the empirical risk function is reliant on the considered training samples; the constructed empirical risk merges with the actual risk upon increasing the size of the training sample. The SVMs primarily aim at finding a function $f(x)$ which will capture the deviation e from the actual output (y), and at the same time be as flat as possible. Considering Eq. (2), the flatness implies seeking a small w and this, as per Smola and Scholkopf [52], can be achieved via the minimization of the Euclidean $\frac{1}{2} \|w\|^2$. Hence, the convex optimization task involves:

$$\begin{aligned} \text{Minimize: } & \frac{1}{2} \|w\|^2 \\ \text{if : } & y_i - (< w \cdot x_i > + b) \leq \varepsilon, \quad i = 1, 2, \dots, l \\ & (< w \cdot x_i > + b) - y_i \leq \varepsilon, \quad i = 1, 2, \dots \end{aligned} \quad (4)$$

The following cost function is minimized to achieve the best regression line:

$$\begin{aligned} \text{Minimize: } & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*) \\ \text{if : } & y_i - (< w \cdot x_i > + b) \leq \varepsilon + \xi_i, \quad i = 1, 2, \dots, l \\ & (< w \cdot x_i > + b) - y_i \leq \varepsilon + \xi_i^*, \quad i = 1, 2, \dots \\ & \xi_i, \xi_i^* \geq 0, \quad i = 1, 2, \dots, l \end{aligned} \quad (5)$$

where ξ_i and ξ_i^* are the positive slack variables that capture the excess deviation in the upper and lower regions while C represents an error penalty that ensures a balance between the empirical error and the regularization term; ε is the associated loss function with the approximation accuracy of the training dataset. Based on Eq. (2), the generic function can be used for the constraints and the Lagrange optimization [53]; this could be best described as follows:

$$f(x, a_i, a_i^*) = \sum_{i=1}^n (a_i - a_i^*) K(x, x_i) + b \quad (6)$$

where $K(x, x_i)$ is the kernel function. As per Mercer's theorem [54], the kernel function is introduced to avoid explicit nonlinear mapping formation, thereby making the feature space dimension more infinite; this significantly reduces the computational load as it enables low dimensional input space operation. Some of the commonly utilized kernels for non-linear cases include polynomial (homogeneous), radial basis function, polynomial (nonhomogeneous), Gaussian function, and sigmoid function. Being that kernel representation uses linear machines in hypothesizing real-world complex problems, it serves as a powerful alternative. This study employed the radial basis function owing to its simplicity and capability to achieve optimal performance when dealing with complex problems.

B. RANDOM FOREST

Random Forest (RF) is a variant of the CART that was developed to improve the model's prediction performance [55] and its block diagram is shown in Figure 6. Its building process is similar to that of the CART model just that it differed by building many trees which gives rise to a forest of models. Building each tree involves only a subset of the predictor variables and the available data set during the building process determines the number of trees to be built (ntree) and the number of predictors to be used during the tree-building

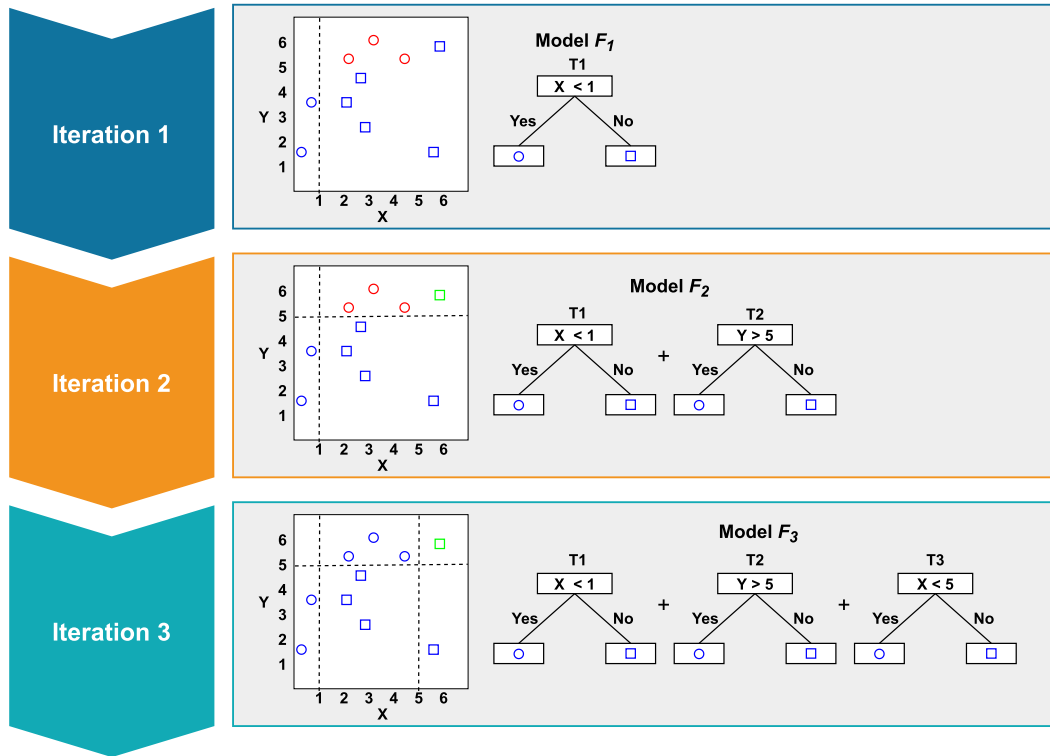


FIGURE 6. Methodology of RF.

process (mtry). A bootstrap sample of the original data set is used to build each tree; this allows the estimation of the robust error using the rest of the test set (the so-called Out-Of-Bag (OOB) sample). The prediction of the excluded OOB samples involves the use of the bootstrap samples and aggregation of the OOB predictions from all trees [56].

A random forest presents a single prediction as its outcome; it is presented as the average of all the aggregations. Some of the advantages of this procedure include achieving higher prediction performance, no issue of overfitting, low individual trees correlation as the forest's diversity is increased by using a small number of predictors, low variance/bias due to averaging over a good number of trees, and estimation of the robust error from the OOB data. However, the major problem of RF is that it has no black box feature, and this makes it difficult to interpret the relationship of the response-predictor variables as the structure of all trees in the forest cannot be possibly investigated [57]. This shortcoming can be overcome by ensuring the procedure allows the estimation of the relevance of the variables by estimating the decline in prediction accuracy before and after permuting a variable.

C. GRADIENT BOOSTED DECISION TREE

Boosting is a technique that relies on the concept of merging a set of weak learners to achieve a better and higher predictive performance [58]. It is a robust learning approach that was initially proposed for classification tasks; however,

it has found application in other fields, such as regression. Boosting is mainly considered when aiming to develop a strong and powerful approach from a set of weak approaches [59]. Regarding gradient boosting, it relies on additive models that are trained in a forward stage-wise approach as follows:

$$F_m(x) = F_{m-1}(x) + h_m(x) \quad (7)$$

where $h_m(X)$ is the basis function (known as the weak learners). Considering the GBDT (shown in Figure 7), the basis functions h_m are small regression trees (RT) of regular size; hence, the GBRT model $F_m(X)$ can be considered a summation of m small RT where the implementation of each boosting iteration m requires the addition of a new RT to the GBDT model. The aim of this procedure is to estimate the response $Y_{(i,t+k)}$ from the given training set; this implies achieving the perfect h_m that will satisfy the condition:

$$F_m(x_{i,t}) = F_{m-1}(X_{i,t}) + h_m(X_{i,t}) = Y_{i,t+k} \quad (8)$$

and this is the same as:

$$h_m(X_{i,t}) = Y_{i,t+k} - F_{m-1}(X_{i,t}) \quad (9)$$

The parameter h_m represents the fitting of the models to the current residuals $r_{(m,i,t)} = Y_{(i,t+k)} - F_{(m-1)}(X_{(i,t)})$ at iteration m . Note that the current residuals are the negative gradients of the squared error loss function (SELF).

$$-\frac{\partial \frac{1}{2}(Y_{i,t+k} - F_{m-1}(X_{i,t}))^2}{\partial F_{m-1}(X_{i,t})} = Y_{i,t+k} - F_{m-1}(X_{i,t}) \quad (10)$$

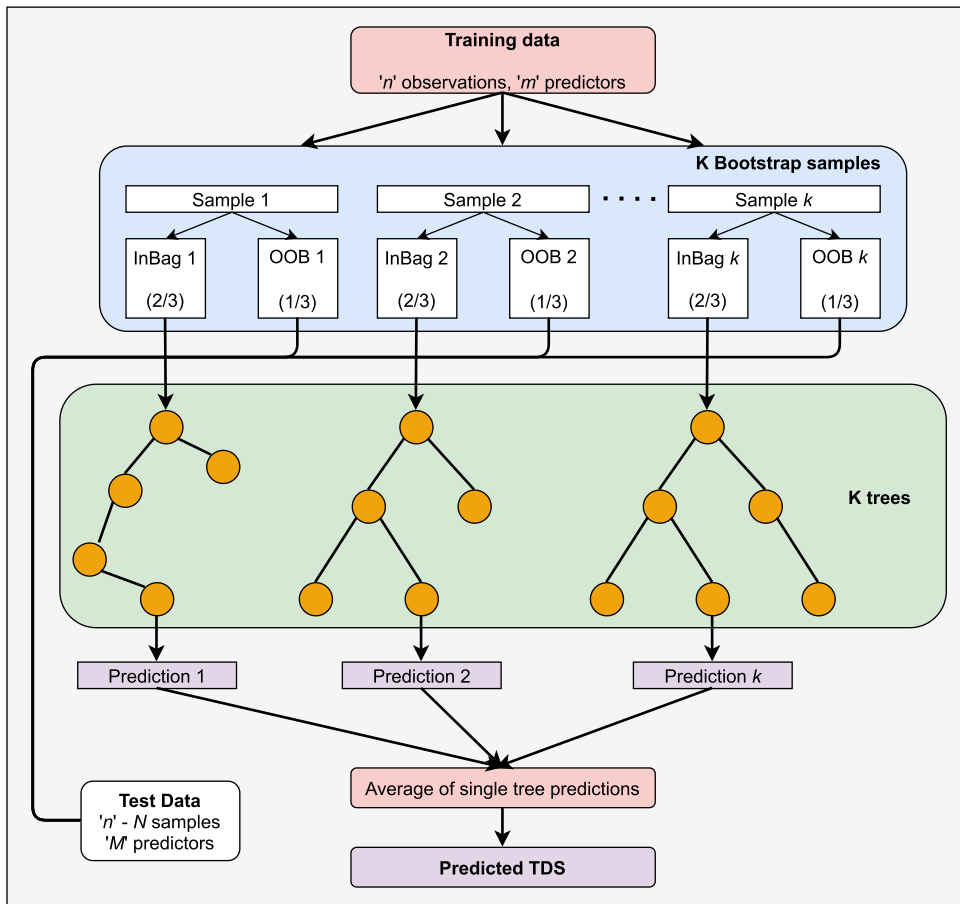


FIGURE 7. Methodology of GBDT.

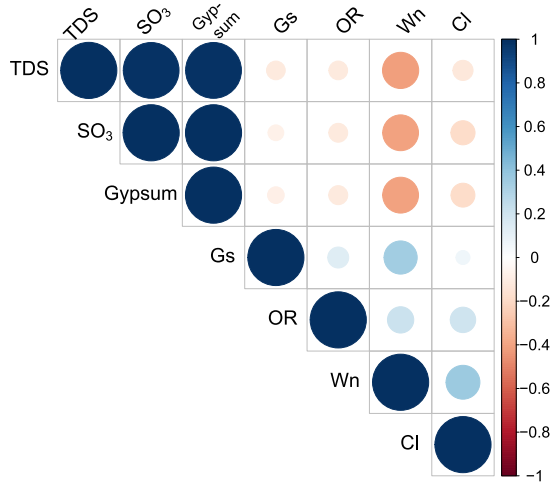
This implies that h_m is equivalent to the negative gradient of the SELF. From Eq. (11), it has shown that only gradient descent or steepest descent algorithms can minimize the SELF. This can be generalized to other loss functions simply by replacing the squared error with another form of the loss function and its gradient [59]. Each boosting iteration goes with the fitting of an RT to the current residuals and the incorporation of enough RTs to the final model will significantly reduce its training error. A simple regularization strategy towards avoiding overfitting is scaling each regression trees' contribution by a factor ν .

$$F_m(X) = F_{m-1}(X) + \nu h_m(X); \nu \in [0; 1] \quad (11)$$

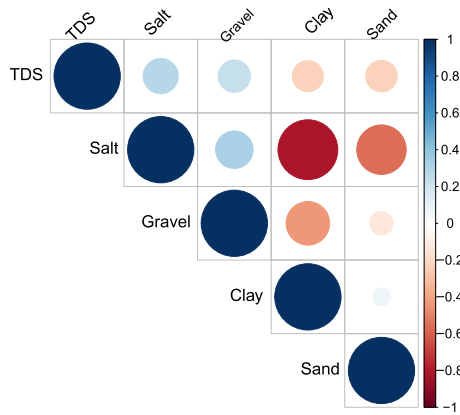
where parameter m is also termed the learning rate that regulates the gradient descent procedure in terms of its step length; ν interacts strongly with M (the number of boosting iterations). Smaller m values require more iterations, meaning that the number of basis functions should be increased to ensure convergence of the training error. It has been empirically proven that better test errors are achieved with small values of m . According to Hastie *et al.* [59], the learning rate should be set to a small constant while M should be chosen by early stopping. The interaction between ν and M has been detailed by [60].

D. MODELING DEVELOPMENT

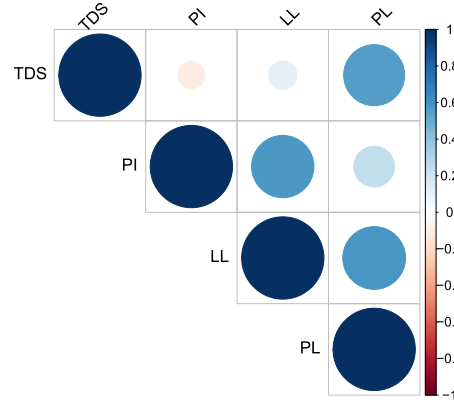
In the current research, three ML models including SVM, GBDT, and RF are applied to predict gypsum soil TDS. The models were developed using NeuroSolutions software. The experimental data are divided into two samples including training (70%) and testing sample (30%). Four modeling scenarios were adopted in this study to predict soil TDS. Scenario 1 contains six variables "the soil chemical properties" as input variables including SO_3 , gypsum, Gs, OR, W_n , and Cl. Scenario 2 includes the consistency soil limits including PL, LL, and PI. Scenario 3 incorporates the soil sieve analysis (gravel, sand, silt, and clay). Whereas, the fourth scenario includes all the input variables introduced in the first three scenarios. Here, the TDS is considered the predicted variable for all the initiated modeling scenarios. To understand the relationship between the input variables and output variable more effectively, exploratory data analysis was applied as displayed in Figure 8. Figure 8a depicts the correlation matrix between the input variables and target variable for the first modeling scenario. High correlation is presented in blue color while white color denotes the low correlation between variables. The change in correlation coefficient is illustrated by the color intensity and the size of the circle. Figure 8a shows that there is a significant correlation between



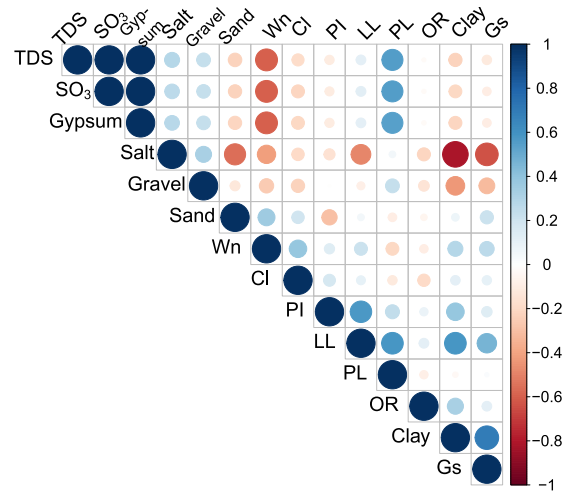
(a) Scenario 1



(c) Scenario 3



(b) Scenario 2



(d) Scenario 4

FIGURE 8. Correlation between the input variables and the target one.

the two variables SO_3 and gypsum. There is also a high correlation between gypsum, SO_3 , and the TDS. The high correlation is visualized by a big and blue circle while the red and small circles show the low correlation between other variables. The properties of the second scenario are visualized in Figure 8b. It has been noticed that there is a moderate correlation between PL and TDS, LL and PI, and between PL and LL, respectively. Figure 8c shows that there is a very small correlation between the sand, gravel, clay, silt, and TDS. The overall view of the correlation coefficient for the fourth modeling scenario between input and output variables is presented in Figure 8d. This figure shows a high correlation between a small number of variables presented in SO_3 , gypsum, and TDS. The correlation coefficient among other variables is ranged between moderate and low scale, and a large number of variables are located at a low level.

E. PERFORMANCE METRICS

The performance of the applied predictive models for the TDS prediction used in the discussed modeling scenarios of the investigated dataset were validated using several statistical metrics including compared with Mean Error (ME), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Percentage Error (MPE), Mean Absolute Percentage Error (MAPE), and R-squared (R^2) values. These performance metrics were defined in equations 12, 13, 14, 15, 16, and 17 respectively.

$$ME = \frac{1}{N} \sum_{i=1}^N (X_i - \hat{X}_i) \quad (12)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N |X_i - \hat{X}_i|^2} \quad (13)$$

TABLE 3. Statistical metrics for training and testing dataset for scenario 1.

Methods	ME	RMSE	MAE	MPE	MAPE	R^2
Scenario 1: Training data						
SVM	-1.688	5.767	2.958	-39.007	47.630	0.810
GBDT	-0.704	3.114	1.221	-2.390	18.170	0.875
RF	-0.018	0.468	0.346	0.711	9.940	0.945
Scenario 1: Testing data						
SVM	-1.665	5.703	3.045	-38.496	50.870	0.824
GBDT	-0.872	3.650	1.570	-5.148	25.630	0.753
RF	-3.541	7.321	3.861	-39.846	25.780	0.614

$$MAE = \frac{1}{N} \sum_{i=1}^N |X_i - \hat{X}_i| \quad (14)$$

$$MPE = \frac{1}{N} \sum_{i=1}^N \frac{X_i - \hat{X}_i}{X_i} \times 100\% \quad (15)$$

$$MAPE = \frac{1}{N} \sum_{i=1}^N \frac{|X_i - \hat{X}_i|}{X_i} \times 100\% \quad (16)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (X_i - \hat{X}_i)^2}{\sum_{i=1}^N (X_i - X_{\text{mean}})^2} \quad (17)$$

where, X_i and $\hat{X}_i$ are the measured and predicted data at the time of t . X_{mean} is the mean of the measured data and N is the number of predicted values.

IV. APPLICATION RESULTS AND ANALYSIS

In this section, the modeling results of the applied predictive models were discussed and assessed based on several statistical and graphical presentations. The results of scenario 1 over the training dataset (in Table 3) showed that the RF model yielded the highest squared correlation ($R^2 = 0.945$) followed by GBDT ($R^2 = 0.875$) and SVM ($R^2 = 0.810$) models. The lowest errors were gained by the RF model with RMSE (0.468), MAE (0.346), and MAPE (9.940). Over the testing phase, the highest squared correlation was yielded by the SVM model (0.824) followed by GBDT (0.753) and RF (0.614). The lowest error was obtained by GBDT with RMSE (3.650), MAE (1.570), and MAPE (25.63). Based on the presented statistical values of the performance metrics, it is clearly indicated that the predictive models performed inconsistently over the training and testing modeling phases. This is a normal fact as the predictive models behave differently based on the learning process of the simulated dataset. However, this is emphasizing the utilization of different statistical performance metrics and graphics to have been assessed and evaluated for the predictive models.

According to the presented statistical results in Table 4, the predictive models attained: SVM model ($R^2 = 0.337$, RMSE=4.653, and MAPE=70.112) and RF ($R^2 = 0.233$, RMSE = 4.852, and MAPE=56.10) models yield promising results than GBTD ($R^2 = 0.272$, RMSE=5.289, and MAPE=76.370) over the testing phase. With regard to Table 4 results, the SVM model has a higher squared correlation (0.337) than RF (0.232) in the testing phase whereas the RF model has the lowest error metrics

TABLE 4. Statistical metrics for training and testing dataset for scenario 2.

Scenario 2: Training data						
Methods	ME	RMSE	MAE	MPE	MAPE	R^2
SVM	-0.004	4.944	3.007	-2.512	67.783	0.225
GBDT	0.000	5.258	3.209	3.621	72.958	0.402
RF	-0.264	4.362	2.544	-6.535	58.167	0.387
Scenario 2: Testing data						
SVM	0.331	4.653	3.028	4.899	70.112	0.337
GBDT	-0.044	5.289	3.307	1.746	76.370	0.272
RF	-0.454	4.852	2.666	-6.932	56.100	0.233

TABLE 5. Statistical metrics for training and testing dataset for scenario 3.

Scenario 3: Training data						
Methods	ME	RMSE	MAE	MPE	MAPE	R^2
SVM	-1.028	5.122	2.747	-14.151	50.852	0.341
GBDT	0.000	4.982	2.991	5.483	67.925	0.403
RF	0.189	2.883	1.965	6.667	44.005	0.746
Scenario 3: Testing data						
SVM	-0.813	5.221	2.775	-9.226	52.151	0.464
GBDT	0.217	5.216	3.269	8.571	75.202	0.374
RF	0.404	4.578	2.782	7.937	68.330	0.343

TABLE 6. Statistical metrics for training and testing dataset for scenario 4.

Scenario 4: Training data						
Methods	ME	RMSE	MAE	MPE	MAPE	R^2
SVM	-0.742	4.106	2.087	-2.040	39.542	0.773
GBDT	-0.779	3.534	1.372	-3.871	21.350	0.769
RF	-0.064	1.000	0.796	2.793	23.730	0.883
Scenario 4: Testing data						
SVM	-0.553	4.007	2.015	3.046	38.499	0.849
GBDT	-0.830	3.882	1.556	-3.301	22.840	0.778
RF	-4.536	8.446	4.937	-62.569	34.750	0.624

among all three models in scenario 2 for both training (RMSE=4.362, MAE=2.544, and MAPE=58.167) and testing (RMSE=4.852, MAE=2.670, and MAPE=56.10) phases.

Table 5 clearly demonstrated that SVR model performed the best owing to the highest squared correlation (0.464) and lowest error measures (RMSE=5.22, MAE=2.775, and MAPE=52.151), outperformed the GBTD ($R^2 = 0.374$, RMSE=5.215, MAE=3.269, and MAPE=75.202) and RF ($R^2 = 0.343$, RMSE=4.578, MAE=2.782, and MAPE=68.33).

Table 6 presented that the SVM model gained the highest squared correlation (0.849) during the testing phase of scenario 4 whereas the GBDT has the lowest error metrics (RMSE=3.882, MAE=1.556, and MAPE=22.84) among all ML models. Thus, it can be concluded that SVR and GBDT have better performance in predicting TDS of soil than RF in scenario 4.

Using the models' performance metrics over the testing phases, the three ML models and for the four modeling scenarios were reflected in the form of spider plots in Figure 9 for each scenario. According to Figure 9a, the GBTD ($R^2 = 0.753$, RMSE=3.650, and MAPE=25.63) by having the lowest error metrics and despite lower the squared correlation than SVM ($R^2 = 0.823$) model was identified as the superior predictive model in scenario 1. Based on the prior reported statistical results and the spider plot given

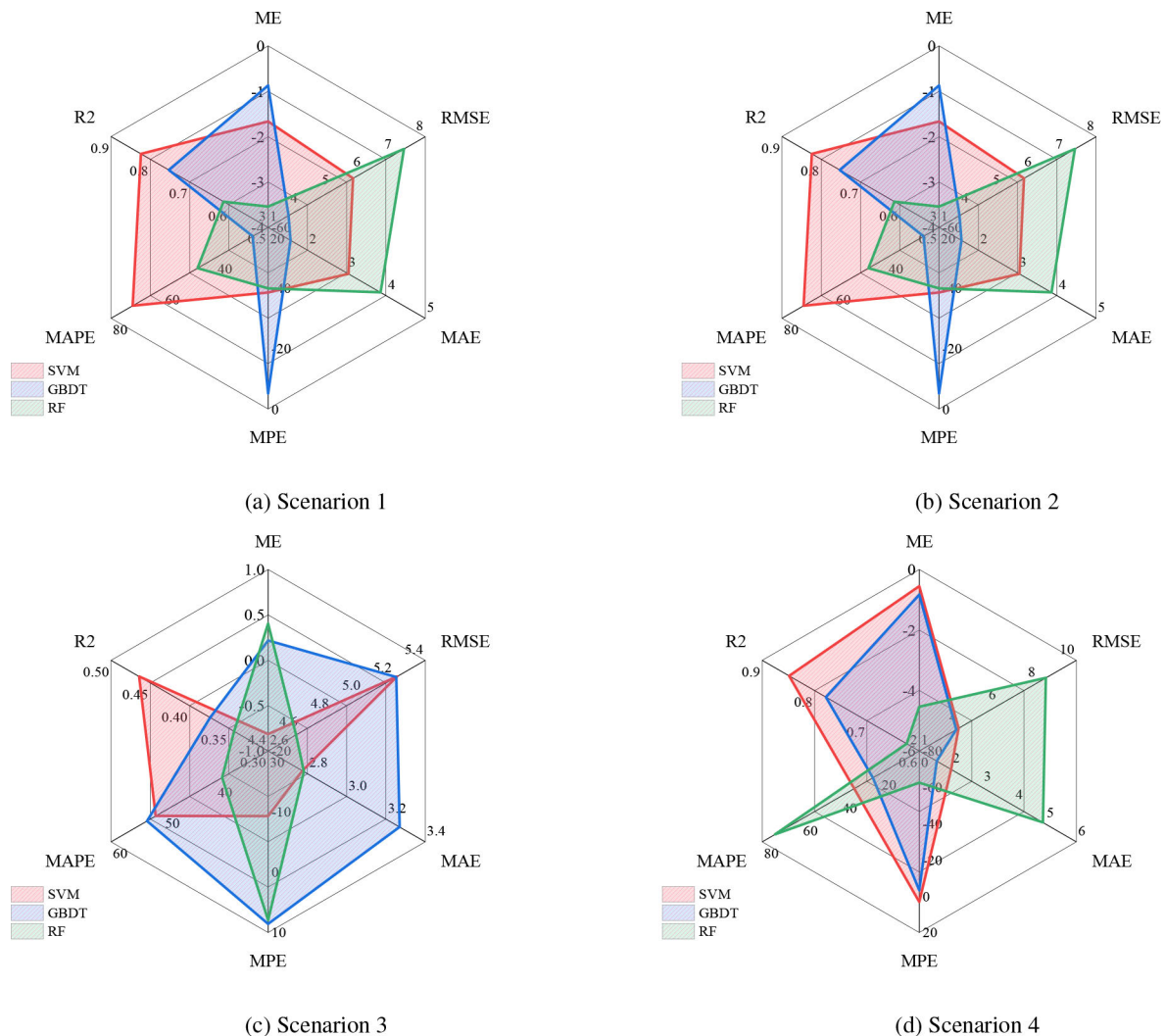


FIGURE 9. The spider plots for the applied predictive models over the four investigated modeling scenarios.

in Figure 9b, it can be inferred that the SVM model slightly is superior to the RF model in the prediction of TDS of soil in scenario 2. For the third modeling scenario, the spider plot Figure (9c), the SVM proves to be more accurate in determining the TSD of soil than other models when the target depends on the soil classification variables. According to the spider plot presented in Figure (9d), the SVM offers more promising results than the GBDT model in the estimation of TDS of soil. In general, based on the analysis performed for all the modeling scenarios, it can be found that Scenario 1 and 4 obtained the best results of applying the ML models with the superiority of GBDT (in the first scenario) and SVM (in the fourth scenario), respectively in capturing the non-linearity between soil TDS and the other variables.

For a better assessment of the modeling results, visualization was applied by using a scatter plot. The scatter plot was drawn for all the predictive models and for every investigated scenario as shown in Figures 10 and 11. The scatter plot has the important manner of evaluating the performance of the ML model to demonstrate the degree of deviation from the

ideal line. Figures 10 and 11 described the scatter plot for scenarios 1, 2 and 3, 4, respectively.

A careful examination of the scatter plots in the first scenario shows that the difference between the ranges of predicted and observed values in the GBDT and RF models is less than the SVM method and therefore has a better performance in capturing the oscillating and non-linear behavior of the data. In scenario 2, the SVM model with a correlation coefficient value equal to 0.34 gained the best performance in predicting TDS of soil in comparison with RF and GBDT models. Also, Figure 11 depicted that the SVM model has more compliance than RF and GBDT models with squared correlation 0.464 and 0.849 for scenarios 3 and 4 respectively. It is crystal clear that implementing the input combination of scenario 4 with the SVM model regarding the least acute angle of 95 percent confidence band is yielded the best correlation relationship.

In the next graphical validation, the performance visualization of physical trends soil's TDS in all scenarios was assessed which is demonstrated in Figures 12 and 13.

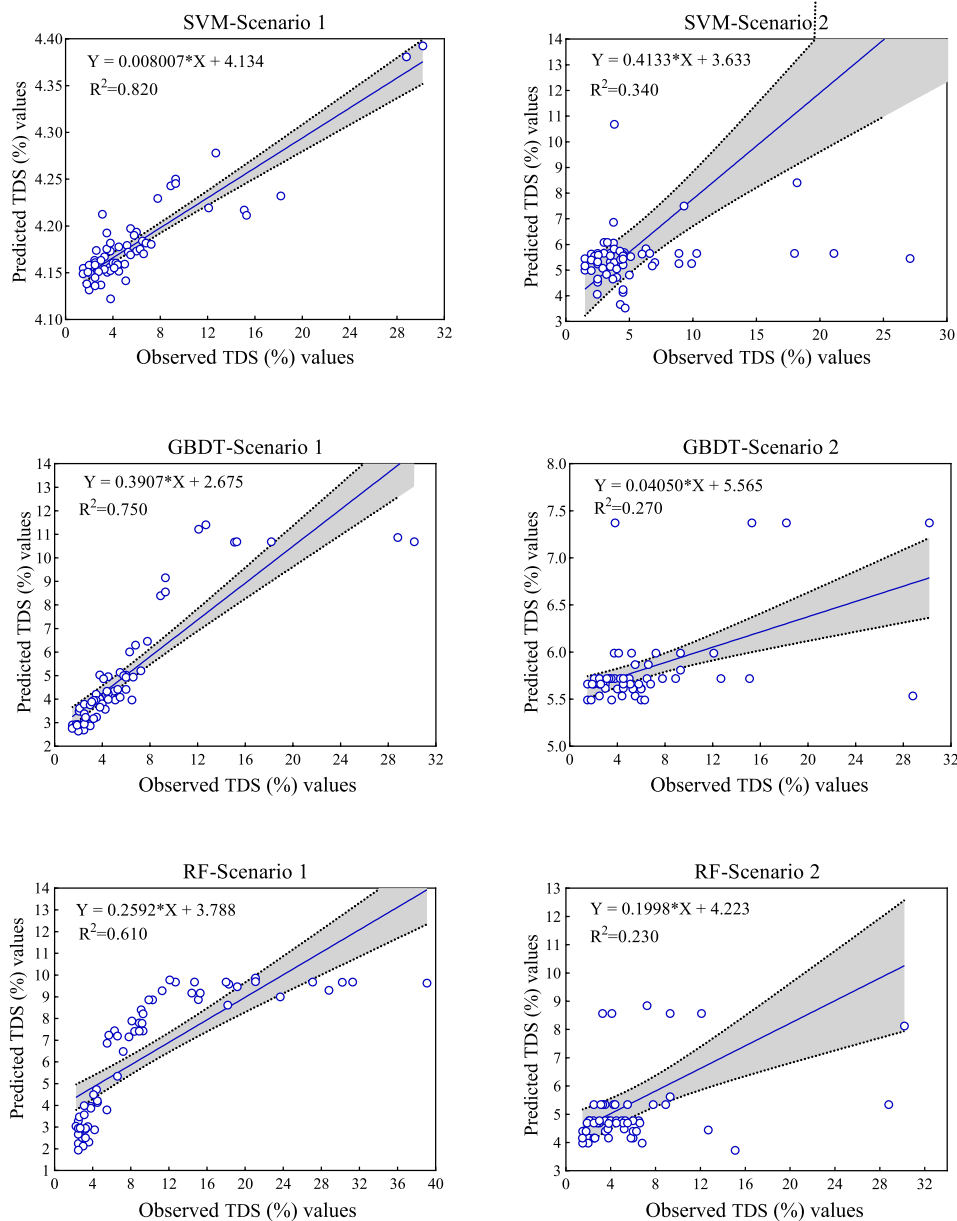


FIGURE 10. Scatter plots for the total dissolved salts predictions for Scenarios 1 and 2.

A comparison of the performance for each of the predictive methods and the four scenarios showed that the best performance is related to scenarios 1 and 4 which GBDT and SVM models were recognized as the superior models, respectively. Besides, Scenario 4 comprised of all effective input variables was yielded the most accurate results and successfully capturing the non-linear behavior of TDS data using the SVM model ($R^2 = 0.849$). It can be noticed that the SVM model in scenario 4 gave the highest performance with R-squared of 0.85 by comparing it to other models.

The relative errors owned by each predictive model were computed for all scenarios by using the violin plot which is drawn for all scenarios as shown in Figure 14. In Scenario 1,

SVR model has the highest relative error range ($-180.30\% \leq E_r \leq 85.45\%$) in comparison with GBDT model by relative error range ($-94.3\% \leq E_r \leq 64.60\%$) and RF model ($-35.30\% \leq E_r \leq 35.37\%$), respectively.

The violin plot of scenario 2 indicated that the highest range of relative error belongs to the GBDT model ($-282.40\% \leq E_r \leq 80.80\%$) and the lowest range of one was owned by the RF model ($-193.20\% \leq E_r \leq 83.50\%$) model. Besides, the distribution of the violin plot in scenario 3 demonstrates that GBDT model yielded the maximum value of relative error ($-328.40\% \leq E_r \leq 76.30\%$) followed by RF model ($-679.70\% \leq E_r \leq 64.30\%$) and SVR model ($-287.50\% \leq E_r \leq 78.15\%$), respectively.

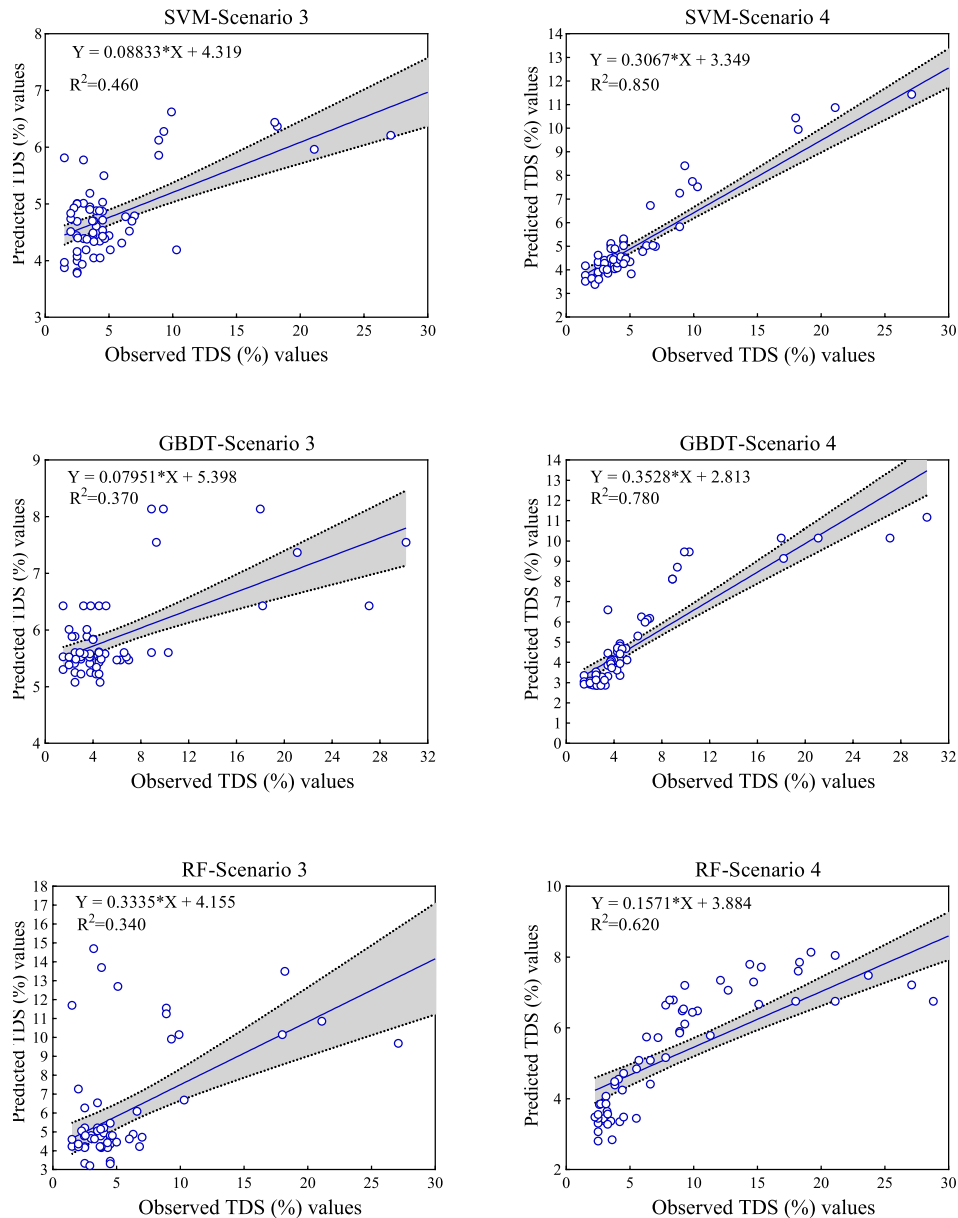


FIGURE 11. Scatter plots for the total dissolved salts predictions for Scenarios 3 and 4.

Eventually, the highest and lowest relative error ranges in scenario 4 belongs to SVM model ($-187.25\% \leq E_r \leq 68.60\%$) and RF model ($-54.8\% \leq E_r \leq 79.60\%$), respectively. However, the relative error concentration of the GBDT model was more compressed than RF and SVM models. Comparing the distribution of relative error of predictive models indicated the scenarios 1 and 4 have the best performance and most acceptable accuracy in estimating TDS of soil to the lowest error range, respectively. Furthermore, the RF model yielded the lowest amount of relative error in the superior scenarios, although the most consistent with the observational values is belong to the SVM model and GBDT is stood in the second rank.

To deeply appreciate the predictive model's efficiency, the cumulative frequency of absolute relative error (CFAE)

variation of all predictive models was assessed in each scenario for better validation and examination of the model's ability for each modeling scenario in Figure 15.

In scenario 1, over 80% of the estimated TDS of soil given by the GBDT model resulted in an absolute relative error of lower than 40% while RF and SVM models obtained 42%, and 73.0%, respectively. According to the CFAE variation against the relative absolute error for scenario 2, the RF model estimate 80% of the testing dataset with absolute relative error less than 93%, SVM model predict 80% of testing data points with absolute relative error less than 109%, and for GBDT model was less than 126%. However, the RF model gained the lowest correlation coefficient (0.272) in scenario 2, and SVM was identified as the superior model for accurate estimation of soil TDS. In scenario 3, the SVM model for more than 90% of

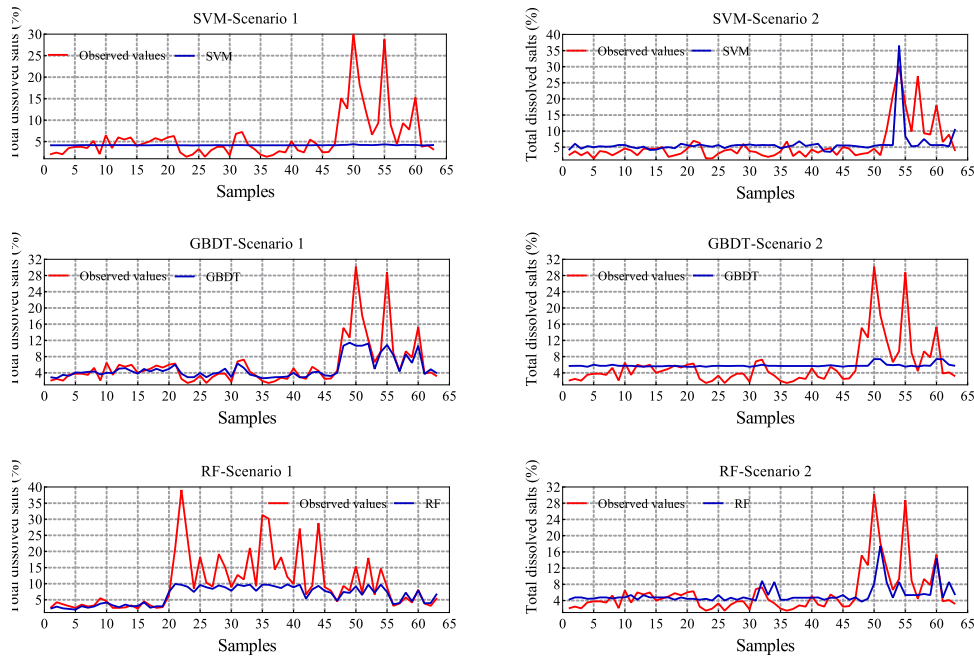


FIGURE 12. Total dissolved salts predictions for Scenarios 1 and 2.

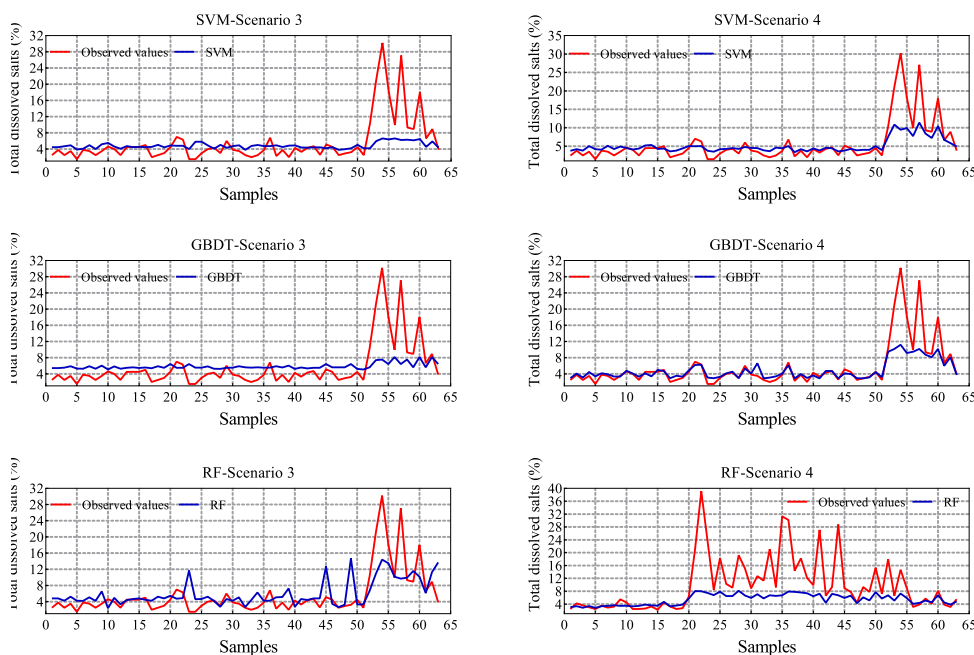


FIGURE 13. Total dissolved salts predictions for Scenarios 3 and 4.

all testing data points estimated TDS of soil with an absolute relative error less than 118%, and around 95% dataset with a relative absolute error less than 141%, whereas the RF model can around 80% of dataset predict the TDS with a relative absolute error less than 118%. Therefore, it can be inferred that the SVM model achieved the most promising results among all models in the third scenario. The error analysis of scenario 4, as the most efficient combination, showed that the

RF model for more than 90% of the testing dataset predict TDS of soil with a relative absolute error of 68% and GBDT and SVM model for the same percentage of data points have less than 54% and 80.8%, respectively.

An accurate comparison of the four modeling scenarios indicated scenario 4 by using all input variables, respect to capability in capturing non-linearity of the dataset, gaining the highest squared correlation values, and lower error ranges

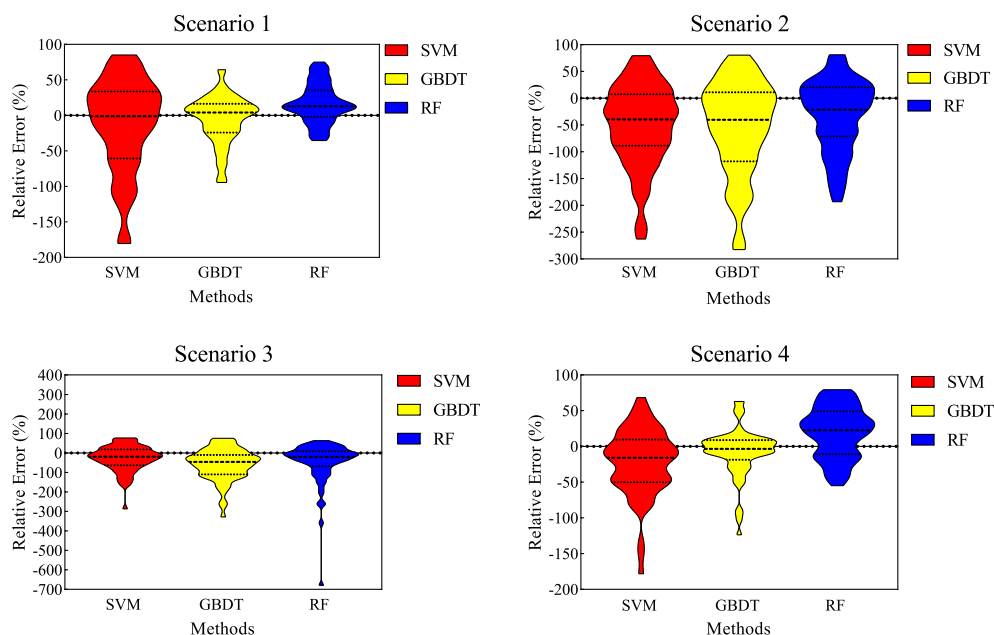


FIGURE 14. Percentage Relative errors for four discussed scenarios.

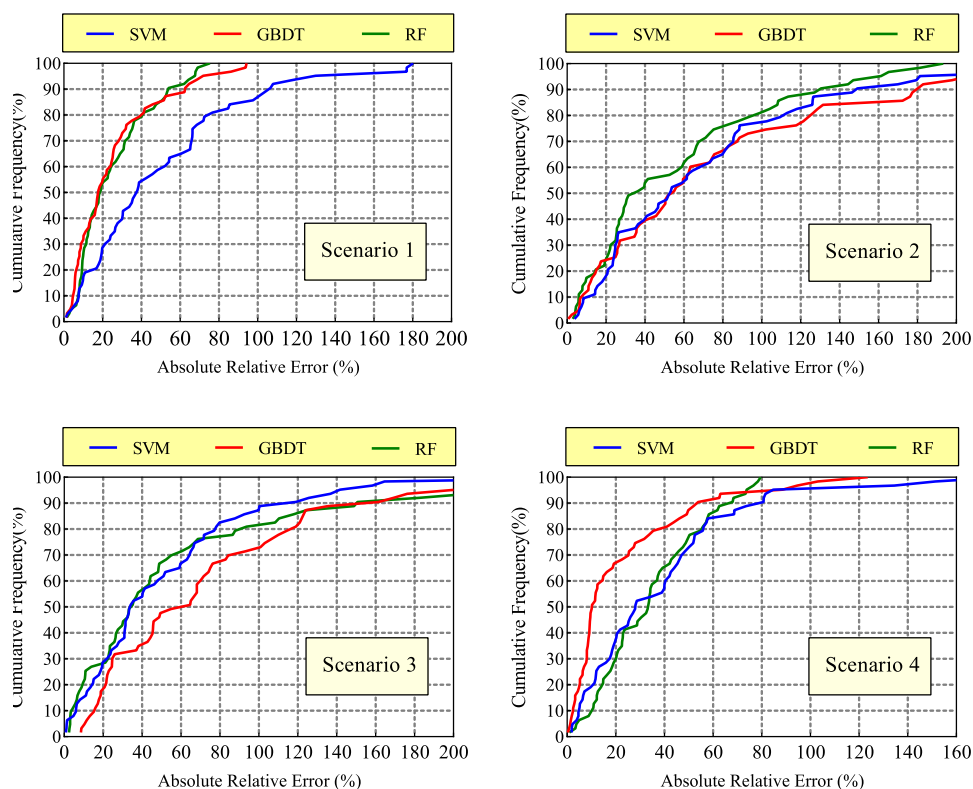


FIGURE 15. Cumulative frequency versus absolute relative errors for four discussed scenarios.

leads to more accurate results in estimating TSD of soil and more efficiency in all predictive models. Besides, the summary of all the above analyzes demonstrated that the SVM model with the highest squared correlation value (0.848) was

identified as the most accurate model in the prediction of TDS value followed by GBDT and RF models, respectively.

Based on the reported modeling results, the authors recommend that the current study can be measured using

the committee (stacking) and ensemble-based machine learning models to combine the advantages of three under study ML-based models for improving the precision of results modeling.

V. CONCLUSION

The motivation of the current study was to develop a robust and reliable machine learning predictive model (SVM, RF, and GBDT) for TDS simulation using several associated soil physicochemical variables. Dataset was collected from five locations within the central and southern parts of the Iraq region. Four different modeling scenarios initiated based on the type of the selected dataset. The applied models are analyzed and discussed based on several statistical measures and graphical presentations. The research findings are summarized as follows:

- Based on the correlation analysis, Gypsum concentration, Sulfur trioxide SO_3 , Chloride (Cl), and organic matter (OR) are the essential soil properties for the TDS concentration influence.
- Among all the investigated modeling scenarios, the first and the fourth scenarios attained satisfactory prediction results.
- The best modeling results of the applied ML models with the superiority of GBDT model for the first scenario and SVM model for the fourth scenario. In the quantitative presentation, the SVM model reported maximum ($R^2 = 0.849$) and minimum (RMSE=3.882) for the fourth scenario over the testing phase.
- The best prediction process attained for the fourth scenario where a combination of all the utilized input variables was used for the models learning process.
- Overall, the current research provided new methodological technology for the TDS of soil prediction within an acceptable degree of accuracy.

REFERENCES

- [1] R. A. V. Rossel, D. J. J. Walvoort, A. B. McBratney, L. J. Janik, and J. O. Skjemstad, "Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties," *Geoderma*, vol. 131, nos. 1–2, pp. 59–75, Mar. 2006.
- [2] I. O. A. Odeh and A. B. McBratney, "Using AVHRR images for spatial prediction of clay content in the lower Namoi Valley of eastern Australia," *Geoderma*, vol. 97, nos. 3–4, pp. 237–254, Sep. 2000.
- [3] S. Vincent, B. Lemerrier, L. Berthier, and C. Walter, "Spatial disaggregation of complex soil map units at the regional scale based on soil-landscape relationships," *Geoderma*, vol. 311, pp. 130–142, Feb. 2018.
- [4] W. D. Hively, G. W. McCarty, J. B. Reeves, M. W. Lang, R. A. Oesterling, and S. R. Delwiche, "Use of airborne hyperspectral imagery to map soil properties in tilled agricultural fields," *Appl. Environ. Soil Sci.*, vol. 2011, pp. 1–13, Jan. 2011.
- [5] L. Winowiecki, T.-G. Vågen, and J. Huising, "Effects of land cover on ecosystem services in tanzania: A spatial assessment of soil organic carbon," *Geoderma*, vol. 263, pp. 274–283, Feb. 2016.
- [6] R. A. V. Rossel, S. R. Cattle, A. Ortega, and Y. Fouad, "In situ measurements of soil colour, mineral composition and clay content by vis-NIR spectroscopy," *Geoderma*, vol. 150, nos. 3–4, pp. 253–266, May 2009.
- [7] R. Baritz, L. Wiese, I. Verbeke, and R. Vargas, "Voluntary guidelines for sustainable soil management: Global action for healthy soils," in *International Yearbook of Soil Law and Policy*. Cham, Switzerland: Springer, 2018, pp. 17–36.
- [8] B. Kuang, H. Mahmood, M. Quraishi, W. Hoogmoed, A. Mouazen, and E. V. Henten, "Chapter four-sensing soil properties in the laboratory, in situ, and on-line: A review," *Adv. Agronomy*, vol. 114, pp. 155–223, Dec. 2012.
- [9] E. Wallor, K.-C. Kersebaum, K. Lorenz, and R. Gebbers, "Soil state variables in space and time: First steps towards linking proximal soil sensing and process modelling," *Precis. Agricult.*, vol. 20, no. 2, pp. 313–334, Apr. 2019.
- [10] R. Li, X. Hou, Z. Jia, and Q. Han, "Soil environment and maize productivity in semi-humid regions prone to drought of Weibei highland are improved by ridge-and-furrow tillage with mulching," *Soil Tillage Res.*, vol. 196, Feb. 2020, Art. no. 104476.
- [11] Y. Li, Z. Li, S. Cui, and Q. Zhang, "Trade-off between soil pH, bulk density and other soil physical properties under global no-tillage agriculture," *Geoderma*, vol. 361, Mar. 2020, Art. no. 114099.
- [12] F. I. Massoud, "Salt affected soils at a global scale and concepts for control," FAO Land Water Develop. Division, Food Agricult. Org. United Nations, Rome, Italy, Tech. Paper, 1981, p. 21.
- [13] D. B. Lobell, "Remote sensing of soil degradation: Introduction," *J. Environ. Qual.*, vol. 39, no. 1, pp. 1–4, Jan. 2010.
- [14] D. B. Lobell, S. M. Lesch, D. L. Corwin, M. G. Ulmer, K. A. Anderson, D. J. Potts, J. A. Doolittle, M. R. Matos, and M. J. Baltes, "Regional-scale assessment of soil salinity in the red river valley using multi-year MODIS EVI and NDVI," *J. Environ. Qual.*, vol. 39, no. 1, pp. 35–41, Jan. 2010.
- [15] J. Rhoades, "Salinity: Electrical conductivity and total dissolved solids," *Methods Soil Anal., Chem. Methods*, vol. 5, pp. 417–435, Jan. 1996.
- [16] M. Pansu and J. Gautheryou, *Handbook of Soil Analysis: Mineralogical, Organic and Inorganic Methods*. Berlin, Germany: Springer-Verlag, 2006, doi: 10.1007/978-3-540-31211-6.
- [17] P. K. Weber-Scan and L. K. Duffy, "Effects of total dissolved solids on aquatic organisms: A review of literature and recommendation for salmonid species," *Amer. J. Environ. Sci.*, vol. 3, no. 1, pp. 1–6, Jan. 2007.
- [18] S. Arora and J. Dagar, "Salinity tolerance indicators," in *Research Developments in Saline Agriculture*. Singapore: Springer, 2019, pp. 155–201.
- [19] A. Morellos, X.-E. Pantazi, D. Moshou, T. Alexandridis, R. Whetton, G. Tziotziou, J. Wiebenson, R. Bill, and A. M. Mouazen, "Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy," *Biosyst. Eng.*, vol. 152, pp. 104–116, Dec. 2016.
- [20] P. A. Burrough, P. F. M. van Gaans, and R. Hootsmans, "Continuous classification in soil survey: Spatial correlation, confusion and boundaries," *Geoderma*, vol. 77, nos. 2–4, pp. 115–135, Jun. 1997.
- [21] P. Scull, J. Franklin, O. A. Chadwick, and D. McArthur, "Predictive soil mapping: A review," *Prog. Phys. Geogr., Earth Environ.*, vol. 27, no. 2, pp. 171–197, Jun. 2003.
- [22] M. Kovačević, B. Bajat, and B. Gajić, "Soil type classification and estimation of soil properties using support vector machines," *Geoderma*, vol. 154, nos. 3–4, pp. 340–347, Jan. 2010.
- [23] C.-W. Chang, D. A. Laird, M. J. Mausbach, and C. R. Hurburgh, "Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties," *Soil Sci. Soc. Amer. J.*, vol. 65, no. 2, pp. 480–490, Mar. 2001.
- [24] K. Islam, B. Singh, and A. McBratney, "Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy," *Soil Res.*, vol. 41, no. 6, pp. 1101–1114, 2003.
- [25] A. M. Mouazen, B. Kuang, J. De Baerdemaeker, and H. Ramon, "Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy," *Geoderma*, vol. 158, nos. 1–2, pp. 23–31, Aug. 2010.
- [26] G. W. McCarty, J. B. Reeves, V. B. Reeves, R. F. Follett, and J. M. Kimble, "Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement," *Soil Sci. Soc. Amer. J.*, vol. 66, no. 2, pp. 640–646, Mar. 2002.
- [27] M. A. Pukalchik, A. M. Katrutsa, D. Shadrin, V. A. Terekhova, and I. V. Oseledets, "Machine learning methods for estimation of the indicators of phosphogypsum influence in soil," *J. Soils Sediments*, vol. 19, no. 5, pp. 2265–2276, May 2019.
- [28] C. O. Martinez-Ojeda, T. M. Amado, and J. C. D. Cruz, "In field proximal soil sensing for real time crop recommendation using fuzzy logic model," in *Proc. Int. Symp. Multimedia Commun. Technol. (ISMATC)*, Aug. 2019, pp. 1–5.

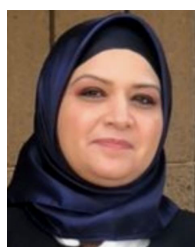
- [29] A. A. Besalatpour, S. Ayoubi, M. A. Hajabbasi, A. Y. Jazi, and A. Gharipour, "Feature selection using parallel genetic algorithm for the prediction of geometric mean diameter of soil aggregates by machine learning methods," *Arid Land Res. Manage.*, vol. 28, no. 4, pp. 383–394, Oct. 2014.
- [30] S. Ahmad, A. Kalra, and H. Stephen, "Estimating soil moisture using remote sensing data: A machine learning approach," *Adv. Water Resour.*, vol. 33, no. 1, pp. 69–80, Jan. 2010.
- [31] R. A. V. Rossel and T. Behrens, "Using data mining to model and interpret soil diffuse reflectance spectra," *Geoderma*, vol. 158, nos. 1–2, pp. 46–54, Aug. 2010.
- [32] X. Guan, S. Wang, Z. Gao, and Y. Lv, "Dynamic prediction of soil salinization in an irrigation district based on the support vector machine," *Math. Comput. Model.*, vol. 58, nos. 3–4, pp. 719–724, Aug. 2013.
- [33] B. Kuang, Y. Tekin, and A. M. Mouazen, "Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon, pH and clay content," *Soil Tillage Res.*, vol. 146, pp. 243–252, Mar. 2015.
- [34] C. W. Brungard, J. L. Boettinger, M. C. Duniway, S. A. Wills, and T. C. Edwards, "Machine learning for predicting soil classes in three semi-arid landscapes," *Geoderma*, vols. 239–240, pp. 68–83, Feb. 2015.
- [35] D. Vermeulen and A. Van Niekerk, "Machine learning performance for predicting soil salinity using different combinations of geomorphometric covariates," *Geoderma*, vol. 299, pp. 1–12, Aug. 2017.
- [36] W. Wu, C. Zucca, A. S. Muhaimeed, W. M. Al-Shafie, A. M. F. Al-Quraishi, V. Nangia, M. Zhu, and G. Liu, "Soil salinity prediction and mapping by machine learning regression in central mesopotamia, Iraq," *Land Degradation Develop.*, vol. 29, no. 11, pp. 4005–4014, Nov. 2018.
- [37] T. Wu, J. Luo, W. Dong, Y. Sun, L. Xia, and X. Zhang, "Geo-object-based soil organic matter mapping using machine learning algorithms with multi-source geo-spatial data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1091–1106, Apr. 2019.
- [38] P. Hoa, N. Giang, N. Binh, L. Hai, T.-D. Pham, M. Hasanlou, and D. T. Bui, "Soil salinity mapping using SAR sentinel-1 data and advanced machine learning algorithms: A case study at ben tre province of the Mekong River delta (Vietnam)," *Remote Sens.*, vol. 11, no. 2, p. 128, Jan. 2019.
- [39] H. Keskin, S. Grunwald, and W. G. Harris, "Digital mapping of soil carbon fractions with machine learning," *Geoderma*, vol. 339, pp. 40–58, Apr. 2019.
- [40] F. Wang, S. Yang, W. Yang, X. Yang, and D. Jianli, "Comparison of machine learning algorithms for soil salinity predictions in three dryland oases located in Xinjiang Uyghur autonomous region (XJUAR) of China," *Eur. J. Remote Sens.*, vol. 52, no. 1, pp. 256–276, Jan. 2019.
- [41] S. Nawar and A. M. Mouazen, "On-line vis-NIR spectroscopy prediction of soil organic carbon using machine learning," *Soil Tillage Res.*, vol. 190, pp. 120–127, Jul. 2019.
- [42] J. I. Rivera and C. A. Bonilla, "Predicting soil aggregate stability using readily available soil properties and machine learning techniques," *Catena*, vol. 187, Apr. 2020, Art. no. 104408.
- [43] S. Katuwal, M. Knadel, T. Norgaard, P. Moldrup, M. H. Greve, and L. W. de Jonge, "Predicting the dry bulk density of soils across Denmark: Comparison of single-parameter, multi-parameter, and vis-NIR based models," *Geoderma*, vol. 361, Mar. 2020, Art. no. 114080.
- [44] S. A. Muhaimeed, A. J. Saloom, K. A. Saliem, K. A. Alani, and W. M. Muklef, "Classification and distribution of iraqi soils," *Int. J. Agricult. Innov. Res.*, vol. 2, no. 6, pp. 997–1002, 2014.
- [45] V. K. Sissakian, N. Al-Ansari, N. Adamo, M. K. Al-Azzawi, A. Mukhalad, and J. Laue, "Geomorphology of the mesopotamian plain: A critical review," *J. Earth Sci. Geotech. Eng.*, vol. 10, no. 4, pp. 1–25, 2020.
- [46] S. Y. Yacoub, "Stratigraphy of the Mesopotamia plain," *Iraqi Bull. Geol. Mining*, no. 4, pp. 47–82, 2011.
- [47] V. K. Sissakian, N. Al-Ansari, N. Adamo, A. Mukhalad, and J. Laue, "Stratigraphy of the mesopotamian plain: A critical review," *J. Earth Sci. Geotech. Eng.*, vol. 10, no. 4, pp. 27–56, 2020.
- [48] S. Y. Yacoub, S. H. Roffa, and J. M. Tawfiq, "The geology Al-Amara—Al-Nasiriya—Al-Basrah area," *Iraq Geol. Surv. Library, Iraq, Tech. Rep.* 1386.
- [49] A. Alwash, H. Istepanian, R. Tollast, and Y. Zeyad, "Towards sustainable water resources management in Iraq," *Iraq Energy Inst., London, U.K., Tech. Rep.*, 2018, pp. 1–45.
- [50] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Workshop Comput. Learn. Theory (COLT)*, 1992, pp. 144–152.
- [51] V. Vapnik, S. E. Golowich, and A. J. Smola, "Support vector method for function approximation, regression estimation and signal processing," in *Proc. Adv. Neural Inf. Process. Syst.*, 1997, pp. 281–287.
- [52] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004.
- [53] Z. Yaseen, M. Ehteram, A. Sharafati, S. Shahid, N. Al-Ansari, and A. El-Shafie, "The integration of nature-inspired algorithms with least square support vector regression models: Application to modeling river dissolved oxygen concentration," *Water*, vol. 10, no. 9, p. 1124, Aug. 2018.
- [54] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2013.
- [55] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [56] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
- [57] A. M. Prasad, L. R. Iverson, and A. Liaw, "Newer classification and regression tree techniques: Bagging and random forests for ecological prediction," *Ecosystems*, vol. 9, no. 2, pp. 181–199, Mar. 2006.
- [58] R. E. Schapire, "The boosting approach to machine learning: An overview," in *Nonlinear Estimation and Classification*. New York, NY, USA: Springer, 2003, pp. 149–171.
- [59] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2009.
- [60] G. Ridgeway, "Generalized boosted models: A guide to the GBM package," *Update*, vol. 1, no. 1, p. 2007, 2007.



NEERAJ DHANRAJ BOKDE received the M.E. degree in embedded systems from the EEE Department, BITS Pilani, Pilani Campus, India, and the Ph.D. degree in data science from the Visvesvaraya National Institute of Technology, Nagpur, India. He is currently working as a Postdoctoral Researcher with the Department of Mechanical and Production Engineering - Renewable Energy and Thermodynamics, Aarhus University, Aarhus, Denmark. His major research contributions are in the domain of data science topics, focused majorly on time series analysis, software package development, and prediction applications in renewable energy. He is serving in editorial positions in *Data in Brief*, *Frontiers in Energy Research*, *Energies*, and *Information journals*.



ZAINAB HASAN ALI is currently pursuing the Ph.D. degree in civil engineering with the University of Diyala. Her main research interest includes construction civil engineering. She has some expertise in soft computing and expert system.



MAYSAM TH. AL-HADIDI has been an Assistant Professor of geotechnical engineering with the Department of Water Resources Engineering, University of Baghdad, since 2013. She has published several research articles internationally and teaches undergraduate and graduate students with the Department of Water Resources Engineering and Civil Engineering, since 2002. Her major is in geo-science engineering.



AITAZAZ AHSAN FAROOQUE is currently working as an Associate Professor with the Faculty of Sustainable Design Engineering, University of Prince Edward Island. His research focuses on fundamental understanding and development of state of the art precision agriculture (PA) technologies for Eastern Canada's agriculture industry. His development of innovative and novel PA systems utilizes knowledge of engineering design, development and management, instrumentation,

design and evaluation of sensors and controllers, development of hardware and software for automation of machines to sense targets in real-time for spot application of agrochemicals on an as-needed basis to improve farm profitability while maintaining environmental sustainability. He is actively working on machine vision, application of multispectral and thermal imagery using drone technology, delineation of management zones for site-specific fertilization, electromagnetic induction methods, remote sensing, and digital photography technique for mapping, bio-systems modeling, artificial neural network, deep learning, analog and digital sensor integration into agricultural equipment for real-time soil, plant, and yield mapping. He has been evaluating the variable rate technologies for potential environmental risks.



MEHDI JAMEI received the M.Sc. and Ph.D. degrees in civil engineering from Shahid Chamran University of Ahvaz, Iran, in 2005 and 2015, respectively. He worked 12 years as a signor engineering and a project manager in the famous consulting companies in Iran. His research contributions are in the domain of data mining topics, focused majorly on applied soft computing, numerical method in porous media and prediction applications in hydrology, scouring, nanofluids, and energy.



ALI ABDULRIDHA AL MALIKI (Member, IEEE) received the B.S. degree in geology and the M.S. degree in remote sensing from the University of Baghdad, Iraq, in 1993 and 2002, respectively, and the Ph.D. degree from the University of South Australia, Australia, in 2015, which was focused on the reflectance spectroscopy and spatial distribution modeling and to be applied for soil contamination and implications for human health. He has worked with the Environment Research

Centre, as a Senior Geologists Chief and a Senior Scientific Researcher in Environment and Water Directorate, Ministry of Higher Education and Scientific Research, Science and Technology, Baghdad, Iraq. His current research interests include environmental management, soil and water contamination, spatial and spectral analyses, water quality assessment, climate change, and applications of geographic information systems in environmental engineering.



BESTE HAMIYE BEYAZTAS received the Ph.D. degree in statistics from Dokuz Eylul University. She is currently an Assistant Professor with the Department of Statistics, Istanbul Medeniyet University, Turkey. Her research interests include time series analysis, resampling methods, forecasting, and panel data analysis.



HOSSAM FARIS received the B.A. degree in computer science from Yarmouk University, Jordan, in 2004, the M.Sc. degree in computer science from Al-Balqa' Applied University, Jordan, in 2008, and the Ph.D. degree in e-business from the University of Salento, Italy, in 2011. In 2016, he worked as a Postdoctoral Researcher with the GeNeura Team, Information and Communication Technologies Research Center (CITIC), University of Granada, Spain. He is currently

a Professor with the School of Computing and Informatics, Al Hussein Technical University, and the Information Technology Department, King Abdullah II School for Information Technology, The University of Jordan. His research interests include applied machine learning, evolutionary computation, knowledge systems, and data mining. He was awarded a full-time competition-based scholarship from the Italian Ministry of Education and Research for his Ph.D. degree. He co-founded The Evolutionary and Machine Learning (Evo-ML.com) Research group.



ZAHER MUNDHER YASEEN received the master's and Ph.D. degrees from the National University of Malaysia (UKM), Malaysia. He is currently a Senior Lecturer and the Director of the Scientific Research Center, Al-Ayen University. He is major in civil engineering applications. He has an excellent expertise in machine learning and advanced data analytics. He has published over 190 research articles in international journals with a Google Scholar H-index of 36, and a total of 4180 citations.

...