Environmental Science and Pollution Research (2024) 31:48955–48971
https://doi.org/10.1007/s11356-024-34440-1

RESEARCH ARTICLE

Groundwater salinity modeling and mapping using machine learning approaches: a case study in Sidi Okba region, Algeria

Samir Boudibi[1] · Haroun Fadlaoui[1] · Fatima Hiouani[2] · Narimen Bouzidi[1] · Azeddine Aissaoui[1] · Zine-eddine Khomri[1]

Received: 29 April 2024 / Accepted: 16 July 2024 / Published online: 23 July 2024

Abstract

The groundwater salinization process complexity and the lack of data on its controlling factors are the main challenges for accurate predictions and mapping of aquifer salinity. For this purpose, effective machine learning (ML) methodologies are employed for effective modeling and mapping of groundwater salinity (GWS) in the Mio-Pliocene aquifer in the Sidi Okba region, Algeria, based on limited dataset of electrical conductivity (EC) measurements and readily available digital elevation model (DEM) derivatives. The dataset was randomly split into training (70%) and testing (30%) sets, and three wrapper selection methods, recursive feature elimination (RFE), forward feature selection (FFS), and backward feature selection (BFS) are applied to train the data. The resulting combinations are used as inputs for five ML models, namely random forest (RF), hybrid neuro-fuzzy inference system (HyFIS), K-nearest neighbors (KNN), cubist regression model (CRM), and support vector machine (SVM). The best-performing model is identified and applied to predict and map GWS across the entire study area. It is highlighted that the applied methods yield input variation combinations as critical factors that are often overlocked by many researchers, which substantially impacts the models' accuracy. Among different alternatives the RF model emerged as the most effective for predicting and mapping GWS in the study area, which led to the high performance in both the training (RMSE = 1.016, R = 0.854, and MAE = 0.759) and testing (RMSE = 1.069, R = 0.831, and MAE = 0.921) phases. The generated digital map highlighted the alarming situation regarding excessive GWS levels in the study area, particularly in zones of low elevations and far from the Foum Elgherza dam and Elbiraz wadi. Overall, this study represents a significant advancement over previous approaches, offering enhanced predictive performance for GWS with the minimum number of input variables.

Keywords Groundwater salinity · Digital elevation model · Machine learning algorithms · Random forest · Sidi Okba region

Introduction

In arid regions, groundwater is the principal source of irrigation, drinking, and industrial use (Kawo and Karuppannan 2018), including Sidi Okba region in Algeria. Groundwater provides 96% of the world's freshwater for around 2.4 billion people (Duran-Llacer et al. 2022; Pandey et al. 2023). Climate change, expansion of agricultural areas, and the scarcity of precipitations led to the increasing of the number of pumping wells in these regions (Hamamouche et al. 2018; Boudibi et al. 2021a; Li et al. 2020; Neshat and Pradhan 2017; Şen 2019). The overexploitation, the poor farming practices, and ineffective management of this valuable and scarce resource contributed to a deterioration of groundwater quality (Aouidane and Belhamra 2017; Afrasinei et al. 2017).

Groundwater salinization, expressed in terms of electrical conductivity (EC), is one of the major constraints for the agricultural production in the study area, because saline irrigation water is responsible for the alteration of physicochemical properties of the soil, which causes soil salinization and reduction of plants productivity (Boudibi 2021; Bradaï et al.

Responsible Editor: Marcus Schulz

✉ Samir Boudibi
Samir.boudhibi@gmail.com; s.boudibi@univ-biskra.dz

[1] Centre de Recherche Scientifique et Technique sur les Régions Arides, CRSTRA, Biskra, Algeria

[2] Department of Agricultural Sciences, University of Mohammed Khider, Biskra, Algeria

2016; Pulido-Bosch et al. 2018). Currently, salinization is rising at 10% annual rate (Barbieri 2023) and it is the primary issue with irrigation-related groundwater quality. Aquifer salinity can be directly or indirectly impacted by agricultural activities operations. As a result, changes in salinity brought about by irrigation water application can be categorized as direct impacts and those coming from irrigation abstraction as indirect impacts (Pulido-Bosch et al. 2018). Thus, managing the irrigation in any region requires precise prediction and mapping of the aquifer's groundwater salinity. In order to improve decision makers' and land use planners' capacity to accurately assess the spatial variability of GWS and to serve as a foundation for studies on groundwater quality risk assessment in various other regions of the world, it is mandatory to determine the accurate machine learning (ML) technique to assess the risk level of groundwater salinization using the adequate digital elevation model (DEM) derivatives.

ML algorithms and geostatistical models are the most suitable methods for digital mapping (DM) (Zhang et al. 2017; Qu et al. 2024). Ordinary kriging (OK) and its derivatives such as cokriging and regression kriging are the most applied geostatistical methods for GWS modeling studies. In this study, OK is used to estimate the salinity at unsampled points and to get an insight into the spatial distribution of GWS. DM uses DEM derivatives and satellite image indices as covariates based on ML, which is widely applied for modeling soil properties (Qu et al. 2024). To our best knowledge, the most recent studies are restricted to using GWS controlling factors (e.g., evaporation, transmissivity, water table, precipitation, and water cations and anions) to predict water salinity for areas of unknown sampling points. The application of such data to model GWS is not always possible because they are not available everywhere, costly, require extensive sampling, and time consuming. The major limitation associated with the application of DM based on ML for modeling GWS is the availability of the adequate covariates (inputs) such as DEM derivatives (e.g., elevation, slope, curvature, distance to rivers and streams, longitude, altitude, and aspect) (Sahour et al. 2020). Therefore, many scholars have applied ML techniques to deal with the non-linear and complex relationships between the target variable and the independent variables in order provide accurate predictions (Xiao et al. 2023; Tran et al. 2021; Meyer and Pebesma 2021).

In recent years, ML techniques have demonstrated their potential as effective tools in predicting GWS. For instance, Sahour et al. (2020) applied multiple linear regression (MLR), deep neural network (DNN), and extreme gradient boosting (EGB) to estimate the GWS in a coastal aquifer of the Caspian Sea. It was concluded that EGB method is the optimal alternative considering its better performance on the testing phase. Cui et al. (2021) used Gaussian processes (GPs) for GWSprediction in the NE part of South Australia. The findings suggested that GPs should be promoted actively in the prediction of groundwater researchers. Araya et al. (2023) employed random forest (RF) to make spatial predictions of GWS in the Horn of Africa. It was reported that RF is powerful tool for the geospatial predictive modeling. Al-Waeli et al. (2022) illustrated the ability of artificial neural networks (ANNs) to predict GWS at the Najaf–Kerbala plateau in Iraq using cations and anions as input data. Jamei et al. (2022) worked on GWS distribution of multi-aquifers in Bangladesh using adaptive neuro-fuzzy inference system (ANFIS) and Boruta-random forest (B-RF). The authors emphasized the great predictability of the applied methods. Lal and Datta (2020) compared the performance of four ML techniques, including ANNs, genetic programming (GP), Gaussian processes regression (GPR), and support vector regression (SVR), for GWS predictions in a coastal aquifer system and attested that GPR performed better than other models.

All the above previous studies have proved the high ability of ML algorithms to handle the complexity of GWS prediction related to groundwater research studies using numerous controlling factors including aquifer characteristics and the groundwater quality elements. Unfortunately, such numerous data are not available in many regions in addition to time consuming. Accordingly, the accuracy of these predictions varies widely depending on the adopted technique (Muniappan et al. 2023). Despite the high importance of the predictors' selection and its effect on the accuracy of the ML techniques, still maximum number of inputs increases computation time and may worsen learning accuracy (Cai et al. 2018). During this study, it is observed that the literature lacks comprehensive application of relevant feature selection methods and readily available influencing factors for modeling GWS using various ML technique. However, in this study the most effective ML method is identified as random forest (RF) with few input parameters, and its performance is shown numerically. Of the main steps in the execution of this study are (1) to map digitally the GWS using readily available DEM derivatives and a small sample dataset of EC; (2) to compare the performance of five commonly used ML techniques, namely random forest (RF), hybrid neuro-fuzzy inference system (HyFIS), K-nearest neighbors (KNN), cubist regression model (CRM), and support vector machine (SVM) for spatial modeling and digital mapping of GWS; and (3) to explore the effect of the different selection feature methods and the number of candidate inputs on the accuracy of these modeling techniques.

Material and methods

Description of the study area

The research was carried out in the Algerian Sahara at the Sidi Okba region (Wilaya of Biskra), which is located 19 km southeast of Biskra province, between 5°45′ N–6°2′ N longitudes and 34°39′ E–34°52′ E latitudes with a total surface area of 280 km^2 and an elevation ranging from 2 m, in the southern part of the study zone, to 126 m in the northern part. It is characterized by an arid climate with cold winters and hot, dry summers, and annual rainfall of less than 150 mm (Hamamouche et al. 2018). The mean annual temperature and evapotranspiration are 23 °C and 2500 mm, respectively (Boudibi et al. 2021b). It is crossed by two wadis, namely wadi Biskra and wadi El-Biraz, which ultimately flow into the natural depression of Chott Melghir (Fig. 1).

Geologically, the study area is a transitional zone, characterized by both structural and sedimentary features, positioned between the mountainous and folded Atlas domain in the north and the expansive, flat desert domain of the northern Sahara in the south (Abdennour et al. 2020; Chebbah 2016; Ghiglieri et al. 2020). Many geologist stated that the sedimentary formations in this region are a succession of Mesozoic to Cenozoic (Guiraud and Bosworth 1997). The Neogene stretches over a large surface area and unevenly cover a range of ages' formations, including Oligocene, Eocene, and Upper Cretaceous (Guiraud 1990; Ghiglieri et al. 2020). A large Quaternary formation discordantly overlies and covers these Neogene deposits (Chebbah 2016).

Hydrogeologically, the study area is recognized by the superposition of two main aquifer systems (Fig. 2): the continental intercalary aquifer (CIa), which is the deepest, and the terminal complex aquifer (CTa) (Besbes et al. 2003). These two aquifer systems are separated by a Cenomanian impermeable horizon and are a part of the North Western Sahara Aquifer System (NWSAS), often known as Système Aquifère du Sahara Septotrional (SASS), that extends over an area of 1 million km^2 shared by Libya, Tunisia, and Algeria, where the major part is in Algeria (about 700 000 km^2) (Al-Gamal 2011; Besser et al. 2018). The CTa comprises several minor aquifers extending from the Upper Cretaceous to The Mio-Pliocene (Edmunds et al. 2003; Ghiglieri et al. 2020). The Mio-Pliocene (called aquifer of sands) consists of alternating layers of clay, sand, and gravel (Reghais et al. 2024). It is the primary exploited aquifer in the eastern part of Biskra province, including Sidi-Okba region. The thickness of this aquifer reaches 1000 m and its depth varies from 90 to 300 m in the study area (Hamamouche et al. 2017; Reghais et al. 2024).

Successive droughts and the expansion of the irrigated agriculture that characterize Sidi Okba regions lead to intensive exploitation of groundwater through deep wells tapping the MPa. In the last decade, groundwater of MPa became the main source of irrigation and drinking purposes in the study area (Hamamouche et al. 2015), despite the existence of Foum Elgherza dam that is used only to irrigate the palm groves of Sidi-Okba, Gharta, and Seriana Oases (Fig. 1). As the study area is experiencing a shortage of surface water from the dam, pumped groundwater is incorporated into the pre-existing irrigation infrastructure, resulting in the generation of an integrated surface and groundwater system. The agricultural sector is the largest groundwater consumer in the study area (more than 90% of the pumped groundwater) (Hamamouche et al. 2018).

Dataset acquisition and preparation

Groundwater salinity measurement

In order to acquire representative network of groundwater wells covering the entire study area as in Fig. 1 and capturing the Mio-Pliocene aquifer in Sidi Okba region, a total of 56 boreholes are used for agricultural and drinking purposes, and they were the subject of on-site measurements during Mai 2020 to obtain the electrical conductivity (EC in mS/cm), which is used to express the GWS. Most of the boreholes were operational during field sampling works. Otherwise, the well water was pumped for about 20 min before sampling to ensure that it represents the aquifer' current state. The EC measurements were carried out using the portable digital multiparameter (WTW multi 3430). The adopted methodology is summarized in Fig. 3.

Digital elevation model (DEM) derivatives

According to the recent studies, GWS is influenced by several factors including climatic, topographic, hydrologic, geologic, land use and land cover (LULC), and aquifer characteristics. In this study, the focus is on the most easily accessible influencing factors, which are the DEM derivatives, namely the slope, flow direction (FlowD), elevation, curvature, aspect, topographic wetness index (TWI), distance to streams, and wadis (DTS and DTBW). It is stated by Avand et al. (2020) that these factors can affect GWS salinity directly or indirectly. On the other hand, slope, elevation, curvature, and aspect play important roles in flushing and exporting saline materials from the soil into fluvial plains through transportation and accumulation of these materials in lowland areas (Mosavi et al. 2020). It is also well known that lower elevation areas often have higher GWS levels due to the accumulation of salts from evaporative concentration. Conversely, higher elevations might show lower GWS levels due to increased groundwater recharge and less evaporation (Leaney et al. 2003; Mosavi et al. 2021). Aspect, which represents the orientation of slopes and the direction of water flow, indirectly affects the amount of water infiltrating into the ground by influencing land cover, wind speed, precipitation direction, and evapotranspiration (Benjmel et al. 2022). Slope and curvature influence the flow and accumulation of water related to the rate of groundwater recharge (Avand et al. 2020; Benjmel et al. 2022), thus, affecting the distribution and concentration of salts in

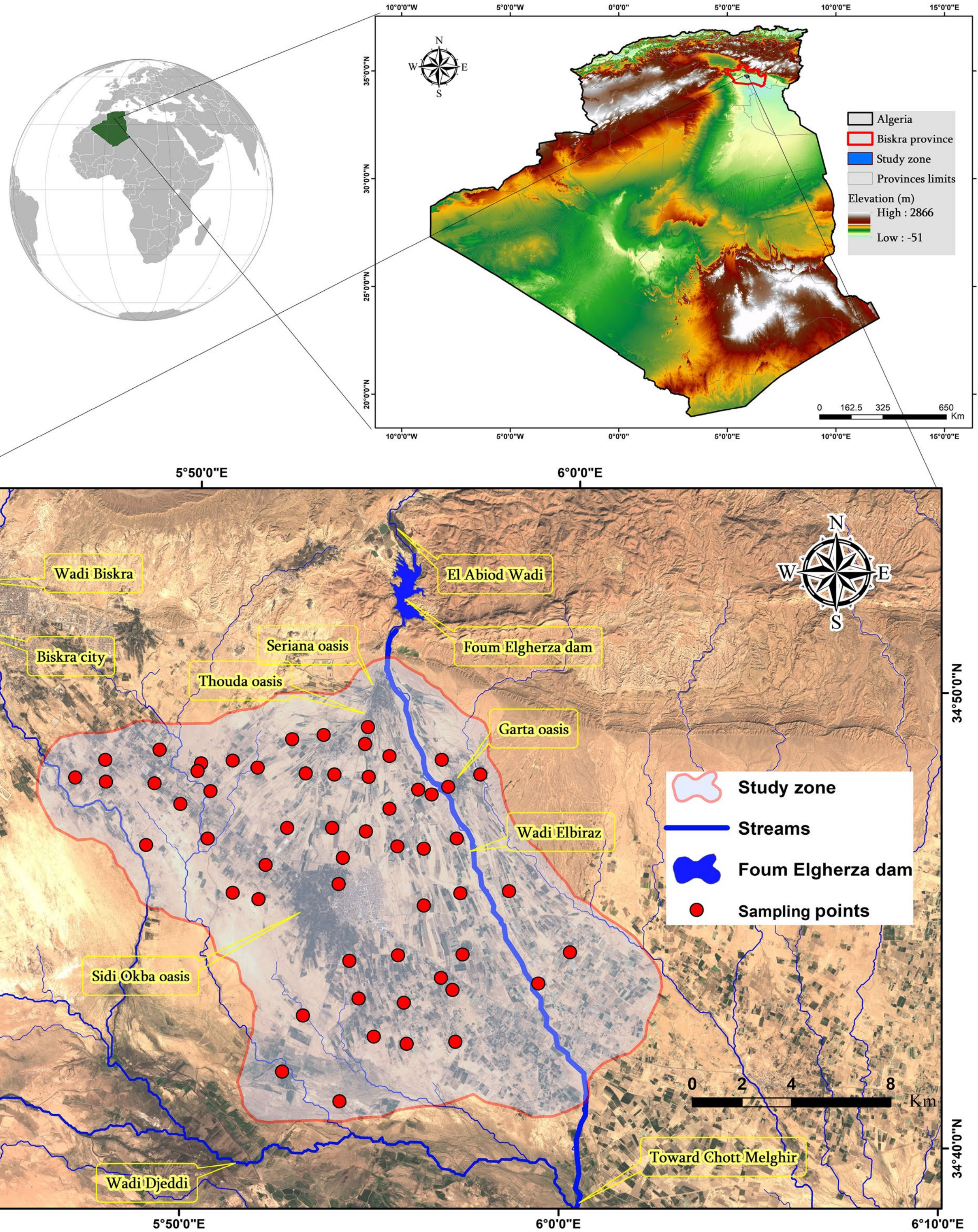

Fig. 1 Location map of the study area

groundwater storage. Topographic wetness index (TWI) is related to soil moisture patterns (Kalantar et al. 2019). Areas with high TWI values indicate higher soil moisture and potentially influence groundwater recharge through the infiltration of surface water, waterlogging, and leaching, which can dilute salinity levels (Mosavi et al. 2021; Benjmel et al. 2022). Streams and wadis play a crucial role in groundwater recharge within the study area. In

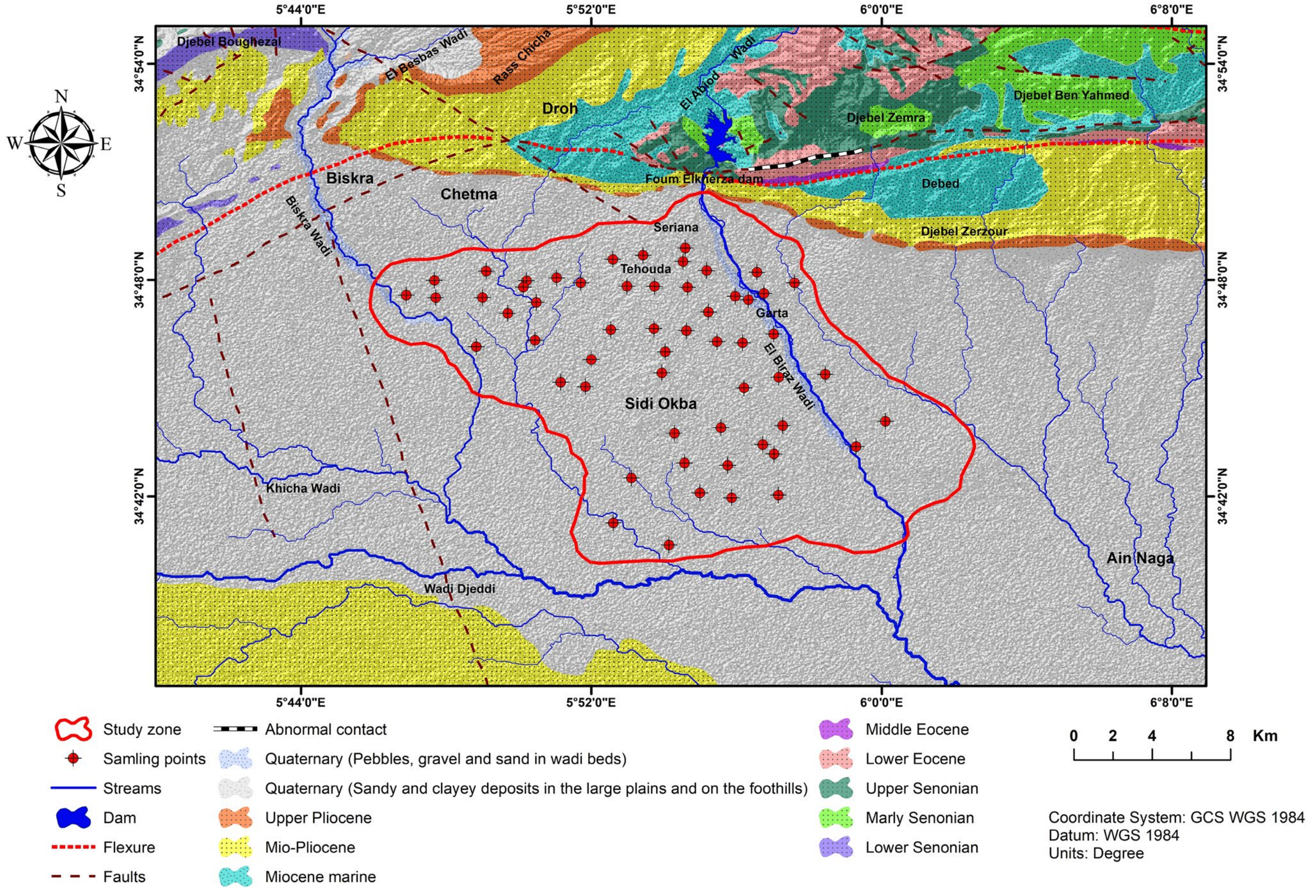

Fig. 2 Hydrogeological map of the study area

addition to serving as a primary source of groundwater recharge, streams and wadis also significantly influence the mobilization and distribution of salts within the aquifers (Balakrishnan et al. 2024). Geographic coordinates are in correlation with climatic precipitation and temperature conditions (Zhao et al. 2007), which affect groundwater recharge and evaporation rates. Longitude and latitude are considered due to their correlation with GWS. The ASTER DEM data of the study area are provided by the US Geological Survey (USGS) (https://earthexplorer.usgs.gov) at a spatial resolution of 30 m (raster cell size of 30 * 30 m). This DEM is utilized to prepare the maps (30 × 30 m of pixel resolution) of the 10 abovementioned derivatives that are extracted and calculated using ArcGIS 10.8 software (Fig. 4). The generated raster maps are imported into the R environment and run using the Raster package for GWS modeling in the entire study area.

Modeling procedure and performance evaluation

All the steps of the modeling process are performed in RStudio/2022.12.00 software using Caret package.

Data standardization

Standardization, also known as centering and scaling, is a preprocessing technique commonly used in ML. This process is typically achieved by subtracting the mean value of each feature from all data points and then dividing by the standard deviation (Müller and Guido 2016). It involves transforming the features of a dataset so that they have a mean of 0 (centering) and a standard deviation of 1 (scaling) (Kraiem et al. 2024; Ouameur et al. 2020). Standardization ensures that features are on similar scales, which can improve the performance of many ML algorithms, particularly those sensitive to the scale of features (Shanker et al. 1996). In this study, standardization is applied automatically during feature selection and model training using preProcess argument passed to train function in R environment.

Feature selection

Feature selection (FS) serves as a valuable tool in ML, offering numerous benefits such as mitigating overfitting, enhancing model performance, and reducing computational

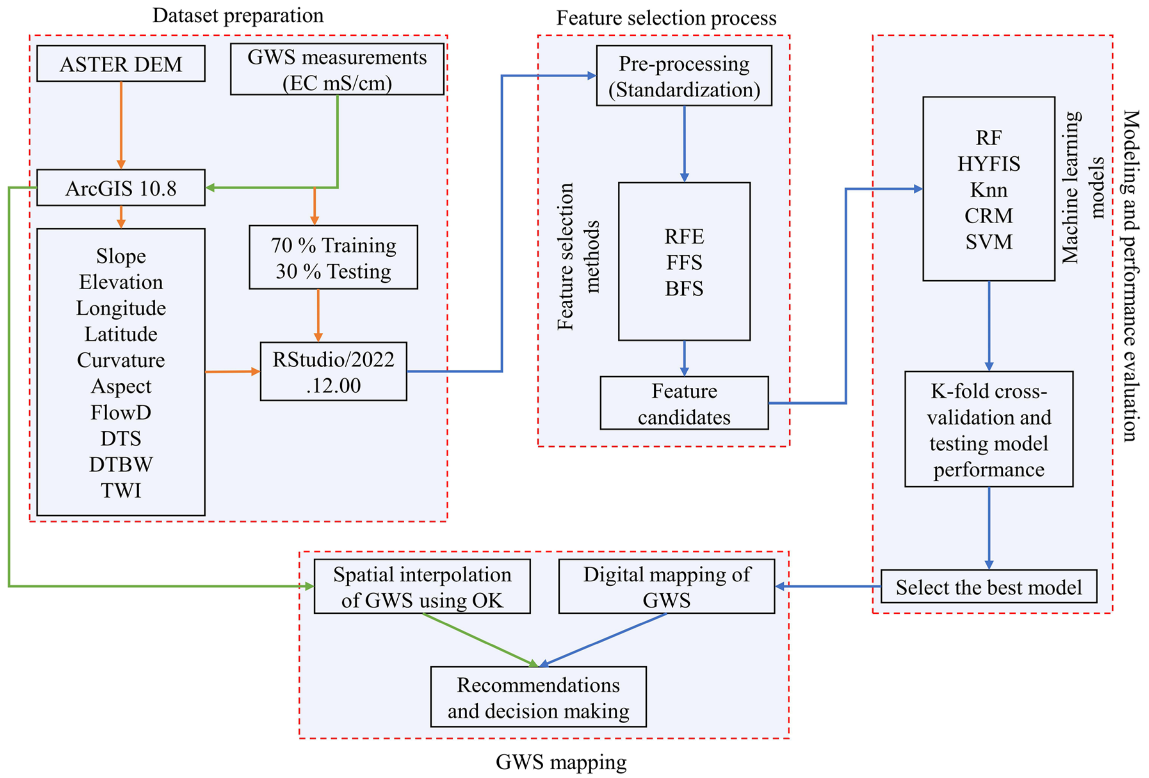
Dataset preparation
ASTER DEM
GWS measurements (EC mS/cm)
ArcGIS 10.8
Slope
Elevation
Longitude
Latitude
Curvature
Aspect
FlowD
DTS
DTBW
TWI
70 % Training
30 % Testing
RStudio/2022 .12.00
Feature selection process
Pre-processing (Standardization)
Feature selection methods
RFE
FFS
BFS
Feature candidates
RF
HYFIS
Knn
CRM
SVM
Machine learning models
Modeling and performance evaluation
K-fold cross-validation and testing model performance
Select the best model
Spatial interpolation of GWS using OK
Digital mapping of GWS
Recommendations and decision making
GWS mapping

Fig. 3 Flowchart methodology

complexity by strategically removing irrelevant or redundant features (Cai et al. 2018; Tran et al. 2021). FS methods can be divided into three principal groups, unsupervised, supervised, and semi-supervised alternatives (Cai et al. 2018). There are three primary categories of supervised feature selection methods: embedded methods, filters, and wrappers (Lualdi and Fasano 2019; Cai et al. 2018). These methods utilize machine learning algorithms and search strategies to iteratively train and test feature subsets, integrating feature selection into model training (Lualdi and Fasano 2019; Jamei et al. 2022). Empirical evidence favors wrappers in terms of performance. In this study, wrappers of FS methods, i.e., backward feature selection (BFS), forward feature selection (FFS), and recursive feature elimination (RFE), are used to pick the best candidate input combinations for the different ML modeling techniques.

Machine learning models

For modeling GWS, five ML models were employed, namely, RF, HyFIS, KNN, SVM, and CRM. A succinct overview of each ML model is provided below.

RF is a tree-based machine learning algorithm (Cutler et al. 2012) harnessing the collective strength of multiple decision trees. Ho (1995) developed the first such algorithm, and Breiman (2001) and Cutler et al. (2012) expanded upon her work, refining and popularizing the algorithm for broader applications in predictive modeling where the fundamental concept is to construct numerous decision trees using the dataset and then amalgamate them to create a predictive model known as the random forest (Parzinger et al. 2022; Kim et al. 2024). RF is attractive and widely applied by researchers due to its high accuracy and efficiency, rapid convergence, and exhibit lower susceptibility to overfitting (Wang et al. 2024; Li et al. 2021). Another notable advantage of this method is its high flexibility, as it does not rely on assumptions about data distribution or necessitate detailed physical models (Parzinger et al. 2022). Additionally, RF possesses the capability to effectively handle missing data and outliers, and thus can be used for tackling both classification and regression problems (Kim et al. 2024; Zhang et al. 2024).

HyFIS is a hybrid neuro-fuzzy system proposed by Kim and Kasabov (1999) for constructing and enhancing fuzzy models through the combination of fuzzy logic principles with learning capabilities of ANNs (Saleh et al. 2023). It is widely applied in ML, where the learning is optimized by hybrid learning scheme that consists of two phase: rule finding using knowledge acquisition module in the initial phase, followed by the parameter learning phase using an error backpropagation learning scheme for a neural fuzzy system (Kim and Kasabov 1999; Hassan and Arman 2023).

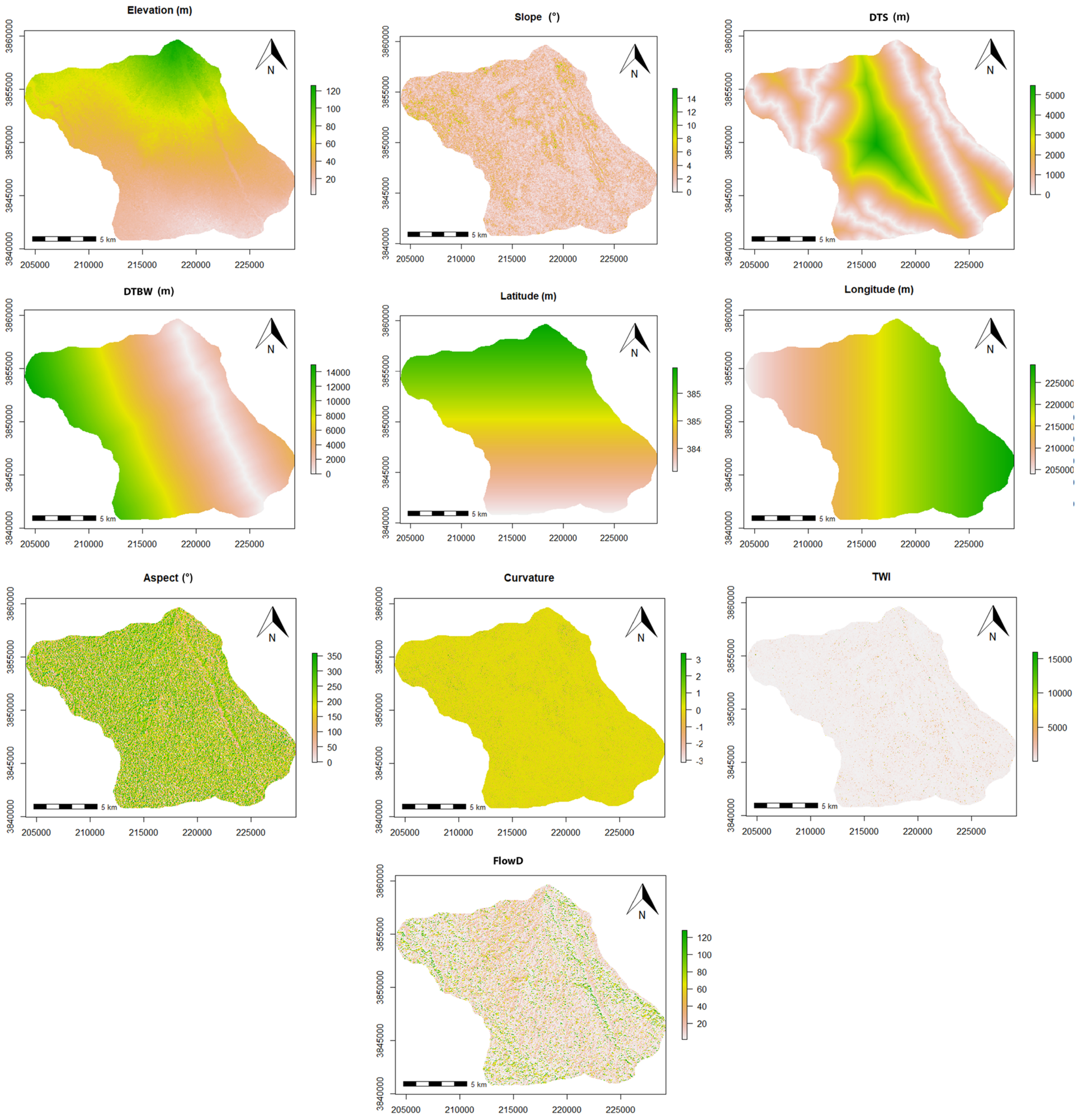

Fig. 4 Digital elevation model (DEM) derivatives

Wang and Mendel (1992) proposed a fuzzy technique for the extraction of fuzzy rules in the HyFIS model, which is a simple method that segments the input and output data into an optimal number of fuzzy sets, then assigns a fuzzy membership function (MF) to each segment (Verma et al. 2022). The procedure is a supervised learning approach that employs gradient descent-based learning algorithms with a multilayer perceptron (Hassan and Arman 2023). In HyFIS, Gaussian function is applied as the MF and, subsequently, during the prediction stage the standard Mamdani methodology can be employed (Ali et al. 2018).

KNN is nonparametric and lazy learning algorithm (Silverman and Jones 1989) proposed by Fix and Hodges (1951) and expanded by Cover and Hart (1967). It is one of the widely employed supervised ML algorithms for forecasting, classification, and regression problems (Chacón et al. 2023). The fundamental concept behind the KNN algorithm is that when examining the feature space, if a significant proportion of the k-nearest neighbors surrounding a particular sample are categorized within a specific group, then that sample is also categorized within that group (Liu et al. 2022; Chacón et al. 2023; Gomez-Gil et al. 2024). Alternatively, the KNN algorithm categorizes an unknown data point by selecting the category of the most similar data point from the training dataset, which is often determined by calculating the Euclidean distance between them (Motevalli et al. 2019; Zamri et al. 2022).

SVM is a powerful supervised ML technique developed by Cortes and Vapnik (1995), originally conceived for tackling classification tasks (He et al. 2022), and later extended to solve regression problems due to its empirical success and application in various research areas (Onyekwena et al. 2022). The SVM gained widespread attention in the first two decades of the twenty first century due to its robust statistical and mathematical foundations, supporting the principles of generalization, optimization, and notable characteristics (Wang et al. 2024; Onyekwena et al. 2022), including independence from data distribution, straightforward algorithm structure, manageable computational complexity, and remarkable generalization capabilities (Joshi 2020). The SVM prioritizes fitting the best line within a specified margin over minimizing the variation between observed and predicted values (Ouameur et al. 2020). This margin delineates the separation between the boundary line and the hyperplane with the nearest data points on both sides designated as support vectors (Majumdar et al. 2023).

CRM is a rule-based model developed based on Quinlan's M5 model tree (Quinlan 1992) and further it utilizes ensemble learning principles to enhance accuracy by combining multiple model trees (Ao et al. 2024). In the ensemble model, each tree mimics a regression tree but it substitutes constant values in terminal nodes with linear regression models (Quinlan 1992; Li et al. 2020). Terminal nodes represent distinct areas in the input space with explanatory variables included in linear regression models, if they significantly influence the response variable within those areas (Li et al. 2020). The main advantage of the cubist model lies in its ability to handle complex non-linear relationships between the inputs (explanatory variables) and the output (target variable) (Ao et al. 2024).

Validation and performance criteria

Validation assesses the efficiency of the applied models. The models are built using the hold out method, where 70% of the training data sets are utilized for training, while the remaining 30% are reserved for testing. K-fold cross-validation is an appropriate evaluation method for a limited dataset (Suleymanov et al. 2023; Chacón et al. 2023). In this study, fivefold cross-validation ($K=5$) is applied and repeated five times on the training set throughout all the modeling techniques via Caret package.

Three performance metrics are utilized to describe the accuracy of the ML models, namely root mean squared error (RMSE), mean absolute error (MAE), and the correlation coefficient (R).

RMSE is defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(GWS_i^O - GWS_i^P\right)^2}{n}} \tag{1}$$

MAE is defined as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|GWS_i^O - GWS_i^P\right| \tag{2}$$

R is defined as

$$R = \frac{\sum_{i=1}^{n}(GWS_i^O - \overline{GWS_i^P})(GWS_i^O - \overline{GWS_i^P})}{\sqrt{\sum_{i=1}^{n}(GWS_i^O - \overline{GWS_i^P})}\sqrt{\sum_{i=1}^{n}(GWS_i^O - \overline{GWS_i^P})}} \tag{3}$$

where ${GWS_i}^O$ denotes the measured groundwater salinity at a location i, ${GWS_i}^P$ signifies the predicted groundwater salinity at a location i, and n represents the total number of sampling points.

Spatial interpolation of groundwater salinity using kriging

Spatial interpolation methods are frequently applied in various fields to estimate values of a variable at locations lacking direct measurements. Kriging is a robust geostatistical interpolation method founded upon the theory of regionalized variable (Delhomme 1978; Şen 1989; Miao and Wang 2024). Ordinary kriging (OK) seeks to offer unbiased and optimal estimates of variables by examining the spatial relationships between data points within the analyzed area through semi-variance (Qu et al. 2024; Zhu et al. 2021). The geostatistical analyst extension of ArcGIS 10.8 is used for conducting OK and spatial prediction of GWS.

Results and discussion

Descriptive statistics of GWS

The descriptive statistics of GWS in terms of EC(mS/cm) are summarized in Table 1. The EC values range from 1.45 to 9.62 mS/cm, with a mean value of 4.573 mS/cm. The results indicate a significant variability, with a coefficient of variation of 41.5%. The Shapiro–Wilk (SK) test is used to assess the normal distribution of EC. The test yielded a significant value of 0.122. Additionally, skewness with a value of 0.226 (near to zero) and kurtosis with a value of 2.239 (near to three) provide further confirmation for the normal distribution assumption of EC (Fig. 5).

Table 1 Descriptive statistics of EC (mS/cm)

Statistical parameters	EC (mS/cm)
Number of samples	56
Mean	4.573
Median	4.645
Minimum	1.450
Maximum	9.620
Variance	3.602
Standard deviation	1.897
Coefficient of variation (%)	41.5
Skewness	0.226
Kurtosis	2.539
Shapiro–Wilk test (*p*-value)	0.122

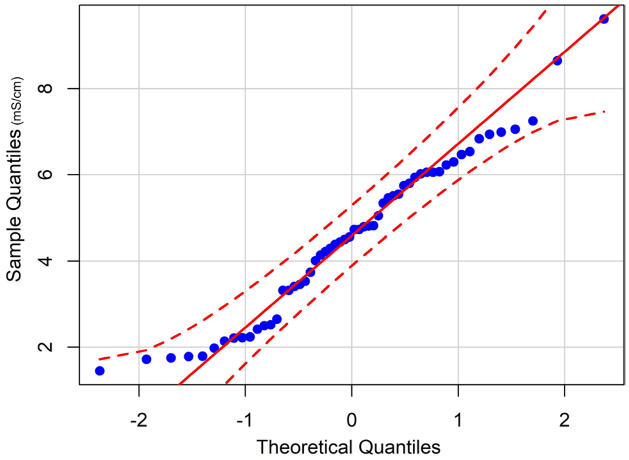

Fig. 5 QQ plot of EC distribution

Spatial interpolation of groundwater salinity

In this research, OK is employed to generate the spatial distribution map of GWS using only the EC field data. The Shapiro–Wilk test indicated that the data of EC follow a normal distribution; therefore, no transformation is required. The primary processing step in Kriging approaches involves fitting a theoretical semivariogram model to the empirical semivariogram. Various models, such as spherical, Gaussian, and exponential are tested, and the most suitable model is selected based on the results of cross-validation, namely by identifying the model with the lowest MAE and RMSE. The parameters of the best-fitted semivariogram are given in Table 2. The cross-validation results of the OK method show good accuracy with a low RMSE of 1.045 and MAE of 0.775, as well as a high *R* value of 0.832.

The generated map of GWS using OK as in Fig. 6 demonstrates the spatial distribution of the various GWS classes in Sidi Okba region. According to the US Salinity Laboratory Staff (Richards 1954), EC (mS/cm) measurements categorize water salinity for irrigation into five classifications: $0 \leq$ low (C1) ≤ 0.25, $0.25 <$ medium (C2) ≤ 0.75, $0.75 <$ high (C3) ≤ 2.25, $2.25 <$ very high (C4) ≤ 5, and excessive (C5) > 5. Each category delineates the suitability of water for different crops and soil types, ranging from low risk of salinization to unsuitability for irrigation. C1 is suitable for most crops, while category 5 is deemed unsuitable. Categories 2 to 4 require varying degrees of caution and crop selection based on salt tolerance and drainage conditions. Ayers and Westcot (1988) showed through extensive experiments that a groundwater EC of 3 mS/cm is acceptable for irrigating most crops. In this research, since the EC values range from 1.45 to 9.62 mS/cm, the groundwater of the study area is classified into three categories: C3, C4, and C5.

From the spatial distribution map in Fig. 6 and values in Table 3, the dominance of groundwater is apparent as very high risk of salinity (C4) with 47.85% of the total surface area. It is localized particularly in the middle and the southern parts of the study area. Groundwater at excessive risk of salinity (C5) occupies a large part of the study area (38.57% of the total surface area). This class is located specifically in the northwestern part with an extension into the middle of the study area (Sidi Okba oasis) along the main road linking the city of Biskra and the town of Sidi Okba. The least saline groundwater (C3) is located in the northeastern part of the study area near the Foum El Gherza dam and the beginning of Wadi Elbiraz, covering 13.58% of the total surface area. According to Ayers and Westcot's recommendations, only 27% of the groundwater in the study area is deemed acceptable for irrigation, and it is located in the eastern part of the study area along Wadi Elbiraz.

Digital mapping of groundwater salinity using machine learning techniques

Covariate selection and model performance

The results in Tables 4, 5, and 6 show the selected covariates using REF, FFS and BFS, respectively, as inputs to the five ML techniques, their tuning parameters, and performance statistics in terms of RMSE, *R*, and MAE criteria.

Table 2 Parameters of the best fitted semivariogram

Kriging method	Anisotropy	Best model	Direction (degrees)	Nugget effect $(mS/cm)^2$	Partial sill $(mS/cm)^2$	Major range (m)	Minor range (m)	Lag size (m)
OK	True	Gaussian	144.84	0.75	3.55	12,456.87	6531.9	1038.07

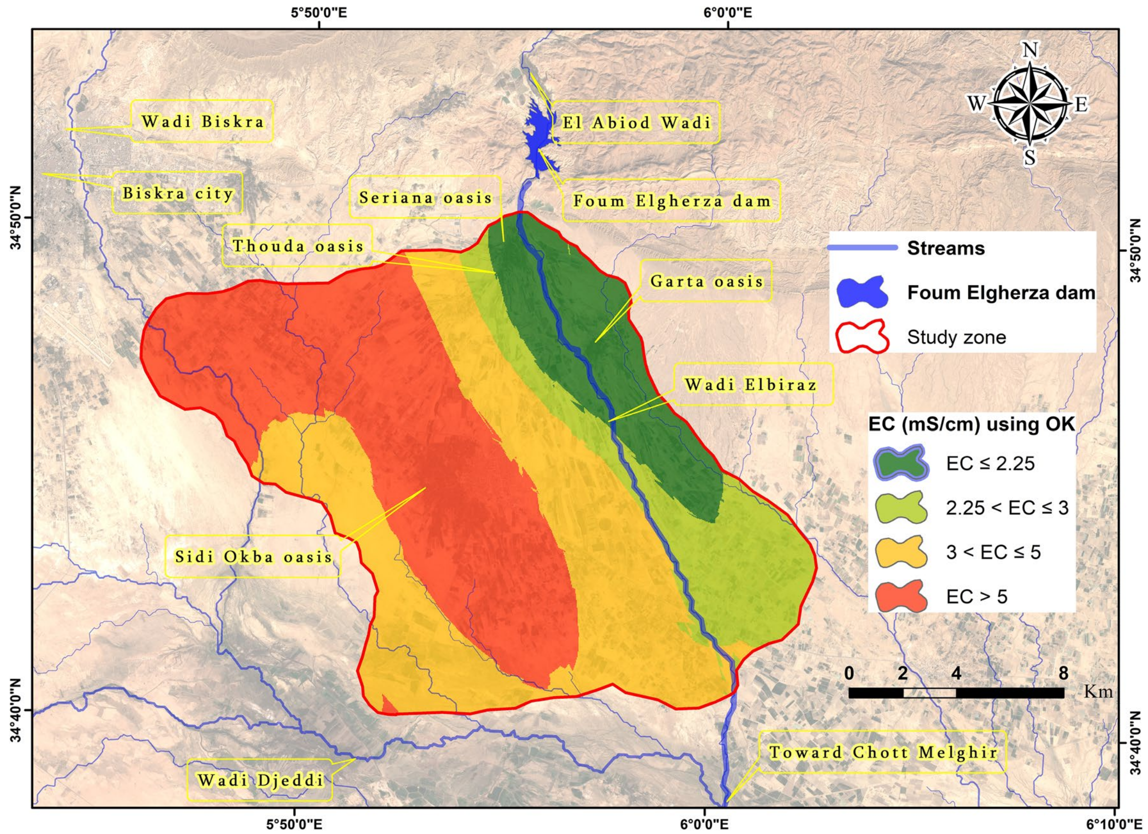

Fig. 6 EC spatial distribution map using OK

For REF selection method (Table 4), among the original ten candidates, only five covariates were selected as inputs to the RF, HyFIS, KNN, CRM, and SVM modeling techniques using REF. From the most important to the least important, they are the distance to Elbiraz Wadi (DTBW), longitude (X), elevation, distance to streams (DTS), and aspect. During the training phase, the SVM model achieved slightly better predictions of GWS with RMSE = 1.010, $R = 0.865$, and MAE = 0.750 compared to the RF model, CRM model, KNN model, and HyFIS model. During the testing phase, the RF model demonstrated superior predictions of GWS with RMSE = 1.069, $R = 0.831$, and MAE = 0.921 compared to the SVM model. However, both the SVM and RF models outperformed the CRM model, the HyFIS model, and the KNN model.

For the FFS selection method as shown in Table 5, nine out of 10 covariates (Y, TWI, slope, curvature, elevation, aspect, FlowD, DTBW, and DTS) were selected as inputs for different ML techniques. During the modeling process, RF model outperformed all other models in both the training and testing phases. The performance metrics were significantly better in the training phase with an RMSE of 1.113,

Table 3 Surface areas of different classes predicted using OK and RF

		OK		RF	
Classes	EC (mS/cm)	Area (ha)	Area (%)	Area (ha)	Area (%)
C1	0 < EC ≥ 0.25	0	0	0	0
C2	0.25 < EC ≥ 0.75	0	0	0	0
C3	0.75 < EC ≥ 2.25	3806.22	13.58	1974.64	7.04
C4	2.25 < EC ≥ 5	13,414.08	47.85	12,974.26	46.28
C5	EC > 5	10,813.55	38.57	13,084.96	46.68
Ayers and Scott threshold	EC ≤ 3	7668.40	27.35	3974.68	14.18
	EC > 3	20,365.45	72.65	24,059.17	85.82

Table 4 Model performance metrics for GWS prediction using REF selection method

Selection method	Variables	Model	Tuning parameter	Training			Testing		
				RMSE	R	MAE	RMSE	*R*	MAE
REF	DTBW, X, Elevation DTS, Aspect	RF	mtry = 4	**1.016**	**0.854**	**0.759**	**1.069**	**0.831**	**0.921**
		HyFIS	Num. labels = 5, max. iter. = 10	1.193	0.821	0.999	1.557	0.727	1.237
		KNN	*K* = 5	1.078	0.858	0.851	1.649	0.612	1.304
		CRM	Committees = 1, Neighbors = 9	1.042	0.861	0.874	1.322	0.771	0.937
		SVM	Kernel: RBF, Sigma = 0.378, *C* = 1	1.010	0.865	0.750	1.129	0.859	0.878

Table 5 Model performance metrics for GWS prediction using FFS selection method

Selection method	Variables	Model	Tuning parameter	Training			Testing		
				RMSE	*R*	MAE	RMSE	*R*	MAE
FFS	Y, TWI, Slope, Curvature, Elevation, Aspect, FlowD, DTBW, DTS	RF	Mtry = 8	1.113	0.817	0.903	1.150	0.862	0.878
		HyFIS	Num. labels = 9, max. iter. = 10	1.752	0.501	1.377	1.480	0.701	1.245
		KNN	*K* = 9	1.446	0.708	1.121	1.537	0.722	1.293
		CRM	Committees = 1, Neighbors = 9	1.160	0.840	0.885	1.331	0.777	0.961
		SVM	Kernel: RBF, Sigma = 0.068, *C* = 1	1.290	0.775	0.964	1.464	0.742	1.188

Table 6 Model performance metrics for GWS prediction using BFS selection method

Selection method	Variables	Model	Tuning parameter	Training			Testing		
				RMSE	*R*	MAE	RMSE	*R*	MAE
BFS	Y, X, TWI, Slope, Curvature, Elevation, Asp, FlowD, DTBW, DTS	RF	Mtry = 6	1.171	0.812	0.958	1.163	0.858	0.918
		HyFIS	Num. labels = 5, max. iter. = 10	1.677	0.629	1.181	1.566	0.670	1.250
		KNN	*K* = 9	1.387	0.693	1.116	1.560	0.678	1.264
		CRM	Committees = 10, Neighbors = 5	1.193	0.810	0.936	1.172	0.826	0.936
		SVM	Kernel: RBF, Sigma = 0.089, *C* = 1	1.232	0.796	0.986	1.393	0.771	1.128

an *R*-value of 0.817, and a MAE of 0.903. In the testing phase, the RF model continued to perform well, with an RMSE of 1.150, an *R*-value of 0.862, and a MAE of 0.878. In addition, the CRM model, which is the second-best performer, demonstrated superior performance compared to SVM and other models.

For the BFS method results in Table 6, all the covariates including *Y*, *X*, TWI, slope, curvature, elevation, aspect, FlowD, DTBW, and DTS were used as inputs for the various ML techniques. The RF model is still the best performer, outperforming all other models with higher performance metrics. In the training phase, the RF model achieved an RMSE of 1.171, an *R*-value of 0.812, and a MAE of 0.958. In the testing phase, the RF model attained an RMSE of 1.163, an *R*-value of 0.858, and a MAE of 0.918. These results indicate the strong performance of the RF model in both phases. Additionally, the CRM model exhibited superiority over the remaining models by achieving the lowest RMSE and MAE, as well as the highest *R*-value in both phases.

How well the model fits the training dataset perform during the training phase measure? However, this measure alone does not evaluate the model's prediction and generalization abilities (Tran et al. 2021). In contrast, the model's predictive performance, which evaluates its accuracy during the testing phase, better demonstrates its ability to predict outcomes reliably (Rahmati et al. 2019; Tran et al. 2021). Therefore, the optimal model to predict GWS in the study area was selected by evaluating both its goodness-of-fit performance during the training phase and its generalization capabilities as demonstrated by its prediction performance during the testing phase.

The chosen model for predicting and mapping GWS in the study area is the RF alternative that uses DTBW, *X*, elevation, DTS, and aspect as inputs selected through the REF selection method. This model exhibited the lowest RMSE

(1.016 and 1.069) and MAE (0.759 and 0.831), as well as the highest *R*-value (0.854 and 0.831) for the training and testing phases, respectively.

Groundwater salinity map

While all input variables are available in continuous raster maps across the entire study area, the representation of GWS in terms of EC (mS/cm) is limited to specific samples distributed within it. However, the selected best model is applied to generate a digital map of GWS (Fig. 7) across the entire Sidi Okba region. This generated GWS map is entered in ArcMap for classification and layout.

The GWS classes of the digital map generated using RF are consistent with those of the spatial distribution map from OK, displaying the same general structure and distribution. The difference lies in the extension of each class with variations in their extent. This congruence enhances the reliability of the modeling results. The surface areas of each class are presented in Table 3. Analysis of Fig. 7 and this table reveals that RF model tends to underestimate the salinity class with high risk (C4), accounting for 7.04% of the entire surface area. Conversely, RF tends to overestimate the salinity class with excessive levels (C5), covering 46.68% of the total surface area. Additionally, 46.28% of the land is covered by groundwater within the C3 category, indicating a very high salinity level. According to Ayers and Westcot's threshold, acceptable groundwater for irrigation covers less than 15% of the entire research area.

Discussions and predictor variables' importance

The type and number of input combinations play a crucial role in the accuracy of ML techniques as demonstrated by the results of three selection methods (REF, FFS, and BFS) applied in this study to identify appropriate predictors (DEM derivatives) of GWS in the Sidi Okba region. The results revealed that the selection methods can yield different input combinations as aspect often overlooked by researchers. This finding is consistent with the results of Theng and Bhoyar (2024), who concluded that the presence of redundant and irrelevant features can lead to less effective ML algorithms as in several research papers in the literature. The REF selection method identified five DEM derivatives as top predictors as shown in Table 4, while the FFS selection method identified nine variables as in Table 5, and BFS considered all variables as top predictors (Table 6). The application of these various input combinations to five ML modeling

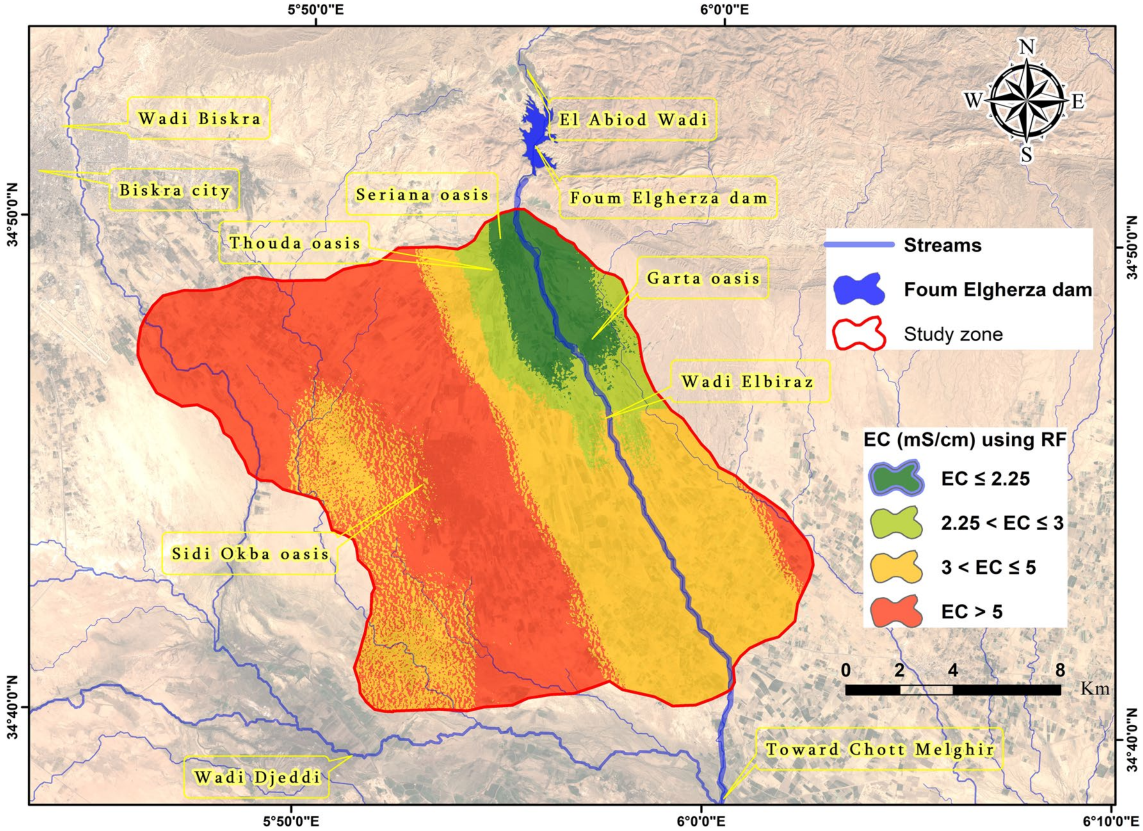

Fig. 7 Digital map of GWS in terms of EC generated using RF

techniques (RF, HyFIS, KNN, CRM, and SVM) demonstrated that utilizing only the five most important predictors consistently yielded the highest accuracy across all models during both the training and testing phases. This result is also supported by the comprehensive survey conducted by Li et al. (2017), which provides empirical evidence that fewer features often lead to better model performance, especially in terms of accuracy and generalization. This is the case also in this paper.

The results of this research indicated that RF is the most effective modeling technique for predicting GWS using distance to Elbiraz wadi (DTBW), *X*, elevation, distance to streams (DTS), and aspect as input variables. The best RF model was employed to generalize prediction results across the entire study area and generate the digital map of GWS using raster formats of the five best predictors. The current result is aligned with recent research displaying the effectiveness and superior predictive capability of RF compared to other ML methods in predicting groundwater quality parameters such as groundwater nitrate pollution (Ouedraogo et al. 2019), groundwater contamination by ammonia concentration (Madani et al. 2022), and groundwater arsenic contamination (Guo et al. 2023; Iqbal et al. 2024).

The feature importance analysis indicated that the distance to Elbiraz wadi, which is fed by releases from the Foum Elgherza dam and other tributaries, is the most impactful variable for predicting GWS in the Mio-Pliocene aquifer in the Sidi Okba region, with a variable importance score (VIS) of 16.37 (see Fig. 8). The digital map clearly shows that as we move further away from the Foum Elgherza dam and Elbiraz wadi, groundwater salinity (GWS) increases. This increase is likely due to the reduced recharge of the aquifer as one moves away from the wadi. In this context, Balakrishnan et al. (2024) confirmed the essential role of streams and wadis not only in groundwater recharge but also in the dissolution, mobilization, and distribution of salts within the aquifers. Longitude (*X*) is the second most important variable affecting the accuracy of the prediction results (VIS = 10.93). This is due to its strong negative correlation with the direction of GWS salinity changes, which generally decrease from the western to the eastern parts of the study area. Moreover, the northeastern part of the study area, which is characterized by lower GWS, is associated with higher elevation areas. In contrast, the northwestern and southeastern parts, which display excessive GWS salinity levels, are associated with lower elevation areas. Elevation (VIS = 8.91) can influence groundwater flow paths as water typically flows from higher to lower elevations transporting minerals and salts along the way. This process can increase salinity in lower elevation areas. Similar observations were made by Mosavi et al. (2020), who indicated that lower elevation areas in Sarvastan plain (Iran) are highly susceptible to GWS evolution. The middle of the study area, which exhibits excessive GWS, is linked to areas with the greatest DTS and is the fourth most important variable (VIS = 6.45). This can be attributed to the low recharge rates in the middle of the study area, which is characterized by intensive agriculture and a high number of wells. Due to these circumstances, the overexploitation of groundwater contributes to the increase of GWS. Pulido-Bosch et al. (2018) and Foster et al. (2018) indicated that poor irrigation practices can intensify GWS by soil salts leaching into groundwater, and thus leading to further groundwater salinity increases. The aspect, which refers to the orientation of slope, is identified as the last significant predictor of GWS in the study area with the lowest VIS of 0.75. It affects GWS dynamics through its impact on precipitation patterns and runoff direction influencing the movement of water and salts within the landscape (Benjmel et al. 2022).

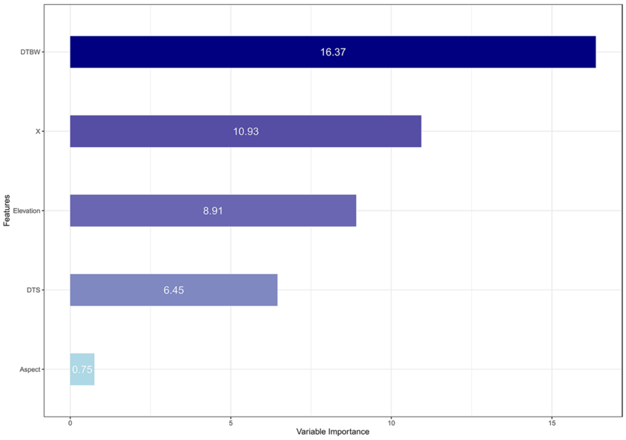

Fig. 8 Predictor variables' importance

The GWS classes on the digital map generated using RF align with those on the spatial distribution map from OK, sharing the same general structure and distribution. This congruence enhances the reliability of the modeling results. However, there are variations in the extent of each class because the RF model tends to moderately underestimate the high-risk salinity class (C4) and slightly overestimate the excessive salinity class (C5). The final digital map of GWS illustrates the alarming state of groundwater quality in the study area. The high levels of GWS may lead to significant environmental problems and economic costs posing a substantial risk to global health (Jamei et al. 2022), emphasizing the urgent need for intervention and remedial measures. Despite this alarming situation, farmers in the study area continue to use groundwater for irrigation without implementing sufficient measures to mitigate the ongoing deterioration situation of groundwater quality. Authorities and policymakers must adapt to increased groundwater salinity, by taking some urgent interventions, including careful freshwater resource management and planning, (e.g., Foum

Elgherza dam); communicate with public, especially farmers, the need to prevent overexploitation, and contamination of groundwater resources due to anthropogenic activities; raise farmers' awareness of the saline water hazards for irrigation and support the implementation of effective methods such as leaching, drainage systems, reverse osmosis desalination of groundwater, and fertigation using modern technologies to prevent the infiltration of saltwater back into the aquifer and the subsequent increase of GWS; and cease pumping from groundwater wells that are no longer suitable for irrigation due to excessive groundwater salinity. After all what have been explained, comparison of the main observations in this study with similar groundwater salinity (GWS) modeling studies in various regions worldwide, such as Bangladesh (Jamei et al. 2022), Vietnam (Tran et al. 2021), Algeria (Tachi et al. 2023), and Iran (Gharechaee et al. 2024), demonstrates that the combination of appropriate machine learning techniques and the effective selection of readily available digital elevation model (DEM) derivatives can result in accurate GWS predictions, even in the absence of specific aquifer parameters, which are costly and not available everywhere.

Conclusion

Although modeling and mapping of groundwater salinity (GWS) procedures are essential for groundwater resources management in any region but especially in arid regions, five machine learning (ML) techniques are used in this study based on electrical conductivity measures and digital elevation model (DEM) derivatives for GWS mapping in the Sidi Okba region. A limited number of strategically positioned wells and 10 DEM derivatives are used with input parameter combinations through RFE, FFS, and BFS methods. The ML models (RF, HyFIS, Knn, CRM, and SVM) are evaluated using RMSE, *R*, and MAE error measurement metrics. The following points are among the main key interpretations obtained from this study.

- It is explained that the type and number of input combinations significantly influence the accuracy of machine learning (ML) techniques. Three of these methods (RFE, FFS, and BFS) are used in the study to identify different sets of predictors (DEM derivatives) for GWS in the Sidi Okba region based on fewer strategically selected features leading to better model performance.
- The random forest (RF) model with five key predictors (distance to Elbiraz wadi, *X*, elevation, distance to streams, and aspect) is found as the most effective alternative for GWS prediction. This model outperformed other ML techniques and therefore, it is used to generate a digital map of GWS in the study area.
- The most impactful variables for predicting GWS are identified as the distance to Elbiraz wadi, longitude (*X*), and elevation. These variables are significantly influenced by the spatial distribution of GWS, especially in higher salinity levels further away from the wadi and in lower elevation areas.
- The study emphasizes the need for authorities and policymakers to implement interventions such as careful groundwater resource management as fresh water resource coupled with public awareness campaigns, and the adoption of effective irrigation practices to mitigate the ongoing deterioration of groundwater quality due to salinity.

For the future research directions, this study has some limitations such as the dataset used is relatively small and the resolution of the DEM derivatives is limited to 30×30 m. Expanding the dataset to include more groundwater wells covering the entire Mio-Pliocene aquifer in the region can enhance the accuracy and reliability of GWS predictions in order to obtain higher resolution remote sensing derivatives. As another alternative future study, a long-term monitoring and temporal analysis of groundwater salinity are necessary to assess changes and salinity trends over time. Despite these limitations, the results of this study mark a notable accomplishment and carry substantial significance for groundwater managers in the study area, as well as researchers globally engaged in similar investigations.

Author contribution Dr. Boudibi Samir: study conception and design, manuscript writing, data collection, modeling, mapping, and corresponding author. Fadlaoui Haroun: data collection and fieldwork. Dr. Hiouani Fatima: assistance in study conception. Bouzidi Narimen: data collection support. Dr. Aissaoui Azeddine: manuscript revision assistance. Dr. Khomri Zinne-eddine: fieldwork support.

Availability of data and materials The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

Abdennour MA, Douaoui A, Barrena J et al (2020) Geochemical characterization of the salinity of irrigated soils in arid regions (Biskra, SE Algeria). Acta Geochim 1–17. https://doi.org/10.1007/s11631-020-00426-2

Afrasinei GM, Melis MT, Buttau C et al (2017) Classification methods for detecting and evaluating changes in desertification-related features in arid and semiarid environments. Euro-Mediterranean J Environ Integr 2:1–19. https://doi.org/10.1007/s41207-017-0021-1

Al-Gamal SA (2011) An assessment of recharge possibility to North-Western Sahara Aquifer System (NWSAS) using environmental isotopes. J Hydrol 398:184–190. https://doi.org/10.1016/j.jhydrol.2010.12.004

Ali D, Hayat MB, Alagha L, Molatlhegi OK (2018) An evaluation of machine learning and artificial intelligence models for predicting the flotation behavior of fine high-ash coal. Adv Powder Technol 29:3493–3506. https://doi.org/10.1016/j.apt.2018.09.032

Al-Waeli LK, Sahib JH, Abbas HA (2022) ANN-based model to predict groundwater salinity: a case study of West Najaf-Kerbala region. Open Eng 12:120–128. https://doi.org/10.1515/eng-2022-0025

Ao X, Qian J, Lu Y, Yang X (2024) Mapping fine-scale anthropogenic heat flux in Shanghai by integrating multi-source geospatial big data using Cubist. Sustain Cities Soc 101:105125. https://doi.org/10.1016/j.scs.2023.105125

Aouidane L, Belhamra M (2017) Hydrogeochemical processes in the Plio-Quaternary Remila aquifer (Khenchela, Algeria). J African Earth Sci 130:38–47. https://doi.org/10.1016/j.jafrearsci.2017.03.010

Araya D, Podgorski J, Berg M (2023) Groundwater salinity in the Horn of Africa: spatial prediction modeling and estimated people at risk. Environ Int 176:107925. https://doi.org/10.1016/j.envint.2023.107925

Avand M, Janizadeh S, Tien Bui D et al (2020) A tree-based intelligence ensemble approach for spatial prediction of potential groundwater. Int J Digit Earth 13:1408–1429. https://doi.org/10.1080/17538947.2020.1718785

Ayers RS, Westcot D (1988) La qualité de l'eau en agriculture. Bull FAO Irrig Drain, Rome, p 170

Balakrishnan JV, Bailey RT, Jeong J et al (2024) Quantifying climate change impacts on future water resources and salinity transport in a high semi-arid watershed. J Contam Hydrol 261:104289. https://doi.org/10.1016/j.jconhyd.2023.104289

Barbieri M (2023) Editorial: Groundwater salinity: origin, impact, and potential remedial measures and management solutions. Front Water 5. https://doi.org/10.3389/frwa.2023.1202576

Benjmel K, Amraoui F, Aydda A et al (2022) A multidisciplinary approach for groundwater potential. Water 14:1553

Besbes M, Abdous B, Abidi B et al (2003) The north western Sahara aquifer system. Joint management of a transborder basin. Houille Blanche 6368:128–133. https://doi.org/10.1051/lhb/2003102

Besser H, Mokadem N, Redhaounia B et al (2018) Groundwater mixing and geochemical assessment of low-enthalpy resources in the geothermal field of southwestern Tunisia. Euro-Mediterranean J Environ Integr 3:1–15. https://doi.org/10.1007/s41207-018-0055-z

Boudibi S, Sakaa B, Benguega Z et al (2021a) Spatial prediction and modeling of soil salinity using simple cokriging, artificial neural networks, and support vector machines in El Outaya plain, Biskra, southeastern Algeria. Acta Geochim 40:390–408. https://doi.org/10.1007/s11631-020-00444-0

Boudibi S, Sakaa B, Benguega Z (2021b) Spatial variability and risk assessment of groundwater pollution in El-Outaya region, Algeria. J African Earth Sci 176:104135. https://doi.org/10.1016/j.jafrearsci.2021.104135

Boudibi S (2021) Modeling the impact of irrigation water quality on soil salinieation in an arid region, case of Biskra, p 176. https://doi.org/10.13140/RG.2.2.12406.93768

Bradaï A, Douaou A, Bettahar N, Yahiaoui I (2016) Improving the prediction accuracy of groundwater salinity mapping using indicator kriging method. J Irrigat Drain Eng 142:11. https://doi.org/10.1061/(ASCE)IR.1943-4774.0001019

Breiman L (2001) RFRSF: employee turnover prediction based on random forests and survival analysis. In: Machine Learning, pp 5–32. https://doi.org/10.1023/A:1010933404324

Cai J, Luo J, Wang S, Yang S (2018) Feature selection in machine learning: a new perspective. Neurocomputing 300:70–79. https://doi.org/10.1016/j.neucom.2017.11.077

Chacón PAM, Segovia Ramírez I, García Márquez FP (2023) K-nearest neighbour and K-fold cross-validation used in wind turbines for false alarm detection. Sustain Futur 6:0–5. https://doi.org/10.1016/j.sftr.2023.100132

Chebbah M (2016) A Miocene-restricted platform of the Zibane zone (Saharan Atlas, Algeria), depositional sequences and paleogeographic reconstruction. Arab J Geosci 9:1–14. https://doi.org/10.1007/s12517-015-2132-9

Cortes C, Vapnik VN (1995) Support vector networks. Mach Learn 20:273–297

Cover TM, Hart PE (1967) Nearest neighbour pattern classification. IEEE Trans Info Theory IT 13:21–27

Cui T, Pagendam D, Gilfedder M (2021) Gaussian process machine learning and Kriging for groundwater salinity interpolation. Environ Model Softw 144:105170. https://doi.org/10.1016/j.envsoft.2021.105170

Cutler A, Cutler DR, Stevens JR (2012) Random forests. In: Zhang C, Ma Y (eds) Ensemble machine learning: methods and applications. Springer Science+Business Media, New York, pp 157–175. https://doi.org/10.1007/978-1-4419-9326-7_5

Delhomme JP (1978) Kriging in the hydrosciences. Adv Water Resour 1(5):252–266

Duran-Llacer I, Arumí JL, Arriagada L et al (2022) A new method to map groundwater-dependent ecosystem zones in semi-arid environments: a case study in Chile. Sci Total Environ 816. https://doi.org/10.1016/j.scitotenv.2021.151528

Edmunds WM, Guendouz AH, Mamou A et al (2003) Groundwater evolution in the Continental Intercalaire aquifer of southern Algeria and Tunisia: trace element and isotopic indicators. Appl Geochem 18:805–822. https://doi.org/10.1016/S0883-2927(02)00189-0

Fix E, Hodges JL (1951) Discriminatory analysis. Nonparametric discrimination: consistency properties. Project number 21–49–004, USAF School of Aviation Medicine, Randolph Field, Texas, pp 1–24

Foster S, Pulido-Bosch A, Vallejos Á et al (2018) Impact of irrigated agriculture on groundwater-recharge salinity: a major sustainability concern in semi-arid regions. Hydrogeol J 26:2781–2791. https://doi.org/10.1007/s10040-018-1830-2

Gharechaee H, Nazari Samani A, Khalighi Sigaroodi S et al (2024) Introducing a novel approach for assessment of groundwater salinity hazard, vulnerability, and risk in a semiarid region. Ecol Inform 81:102647. https://doi.org/10.1016/j.ecoinf.2024.102647

Ghiglieri G, Buttau C, Arras C et al (2020) Using a multi-disciplinary approach to characterize groundwater systems in arid and semi-arid environments: the case of Biskra and Batna regions (NE Algeria). Sci Total Environ 757:143797. https://doi.org/10.1016/j.scitotenv.2020.143797

Gomez-Gil FJ, Martínez-Martínez V, Ruiz-Gonzalez R et al (2024) Vibration-based monitoring of agro-industrial machinery using a k-nearest neighbors (kNN) classifier with a harmony search (HS) frequency selector algorithm. Comput Electron Agric 217. https://doi.org/10.1016/j.compag.2023.108556

Guiraud R, Bosworth W (1997) Senonian basin inversion and rejuvenation of rifting in Africa and Arabia: synthesis and implications to plate-scale tectonics. Tectonophysics 282:39–82. https://doi.org/10.1016/S0040-1951(97)00212-6

Guiraud R, (1990) Evolution post-triasique de l'avant pays de la chaîne alpine en Algérie d'après l'étude du bassin du Hodna et des régions voisines. Office National de la Géologie, Alger, p 259

Guo W, Gao Z, Guo H, Cao W (2023) Hydrogeochemical and sediment parameters improve predication accuracy of arsenic-prone groundwater in random forest machine-learning models. Sci Total

Environ 897:165511. https://doi.org/10.1016/j.scitotenv.2023.165511

Hamamouche MF, Kuper M, Lejars C (2015) Émancipation des jeunes des oasis du Sahara algérien par le déverrouillage de l'accès à la terre et à l'eau. Cah Agric 24:412–419. https://doi.org/10.1684/agr.2015.0777

Hamamouche MF, Kuper M, Riaux J, Leduc C (2017) Conjunctive use of surface and ground water resources in a community-managed irrigation system — the case of the Sidi Okba palm grove in the Algerian Sahara. Agric Water Manag 193:116–130. https://doi.org/10.1016/j.agwat.2017.08.005

Hamamouche MF, Kuper M, Amichi H et al (2018) New reading of Saharan agricultural transformation: Continuities of ancient oases and their extensions (Algeria). World Dev 107:210–223. https://doi.org/10.1016/j.worlddev.2018.02.026

Hassan MY, Arman H (2023) HYFIS vs FMR, LWR and Least squares regression methods in estimating uniaxial compressive strength of evaporitic rocks. Sci Rep 13:1–15. https://doi.org/10.1038/s41598-023-41349-1

He B, Jia B, Zhao Y et al (2022) Estimate soil moisture of maize by combining support vector machine and chaotic whale optimization algorithm. Agric Water Manag 267:107618. https://doi.org/10.1016/j.agwat.2022.107618

Ho TK (1995) Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition, pp 278–282

Iqbal J, Su C, Ahmad M et al (2024) Hydrogeochemistry and prediction of arsenic contamination in groundwater of Vehari, Pakistan: comparison of artificial neural network, random forest and logistic regression models. Environ Geochem Health 46:14. https://doi.org/10.1007/s10653-023-01782-7

Jamei M, Karbasi M, Malik A et al (2022) Computational assessment of groundwater salinity distribution within coastal multi-aquifers of Bangladesh. Sci Rep 12:1–28. https://doi.org/10.1038/s41598-022-15104-x

Joshi A (2020) Support vector machines. In: Joshi AV (ed) Machine learning and artificial intelligence. Springer Nature, Switzerland, pp 65–71. https://doi.org/10.1007/978-3-031-12282-8_8

Kalantar B, Al-Najja HAH, Pradhan B et al (2019) Optimized conditioning factors using machine learning techniques for groundwater potential mapping. Water 11(9):1909. https://doi.org/10.3390/w11091909

Kawo SN, Karuppannan S (2018) Groundwater quality assessment using water quality index and GIS technique in Modjo River Basin, central Ethiopia. J African Earth Sci 147:300–311. https://doi.org/10.1016/j.jafrearsci.2018.06.034

Kim J, Kasabov N (1999) HyFIS: Adaptive neuro-fuzzy inference systems and their application to nonlinear dynamical systems. Neural Netw 12:1301–1319. https://doi.org/10.1016/S0893-6080(99)00067-2

Kim JH, Lee DH, Mendoza JA, Lee MY (2024) Applying machine learning random forest (RF) method in predicting the cement products with a co-processing of input materials: optimizing the hyperparameters. Environ Res 248:118300. https://doi.org/10.1016/j.envres.2024.118300

Kraiem Z, Zouari K, Chkir N (2024) Accurate prediction of salinity in Chott Djerid shallow aquifers, southern Tunisia: Machine learning model development. Water Sci 38:33–47. https://doi.org/10.1080/23570008.2023.2294535

Lal A, Datta B (2020) Performance evaluation of homogeneous and heterogeneous ensemble models for groundwater salinity predictions: a regional-scale comparison study. Water Air Soil Pollut 2031–320. https://doi.org/10.1007/s11270-020-04693-w

Leaney FW, Herczeg AL, Walker GR (2003) Salinization of a fresh palaeo-groundwater resource by enhanced recharge. Ground Water 41:84–92. https://doi.org/10.1111/j.1745-6584.2003.tb02571.x

Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H (2017) Feature selection: a data perspective. ACM Comput Surv 50(6):1–45. https://doi.org/10.1145/3136625

Li Y, Hernandez JH, Aviles M et al (2020) Empirical Bayesian kriging method to evaluate inter-annual water-table evolution in the Cuenca Alta del Río Laja aquifer, Guanajuato, México. J Hydrol 582:124517. https://doi.org/10.1016/j.jhydrol.2019.124517

Li X, Liu J, Liu D et al (2021) Measurement and analysis of regional agricultural water and soil resource composite system harmony with an improved random forest model based on a dragonfly algorithm. J Clean Prod 305. https://doi.org/10.1016/j.jclepro.2021.127217

Liu G, Zhao H, Fan F et al (2022) An enhanced intrusion detection model based on improved kNN in WSNs. Sensors 22:1–18. https://doi.org/10.3390/s22041407

Lualdi M, Fasano M (2019) Statistical analysis of proteomics data: a review on feature selection. J Proteome 198:18–26

Madani A, Hagage M, Elbeih SF (2022) Random forest and logistic regression algorithms for prediction of groundwater contamination using ammonia concentration. Arab J Geosci 15:1619. https://doi.org/10.1007/s12517-022-10872-2

Majumdar P, Mitra S, Bhattacharya D (2023) Soil moisture simulation of rice using optimized support vector machine for sustainable agricultural applications. Sustain Comput Informatics Syst 40:100924. https://doi.org/10.1016/j.suscom.2023.100924

Meyer H, Pebesma E (2021) Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol Evol 12:1620–1633. https://doi.org/10.1111/2041-210X.13650

Miao C, Wang Y (2024) Interpolation of non-stationary geo-data using kriging with sparse representation of covariance function. Comput Geotech 169:106183. https://doi.org/10.1016/j.compgeo.2024.106183

Mosavi A, Hosseini FS, Choubin B et al (2020) Groundwater salinity susceptibility mapping using classifier ensemble and Bayesian machine learning models. IEEE Access 8:145564–145576. https://doi.org/10.1109/ACCESS.2020.3014908

Mosavi A, Sajedi Hosseini F, Choubin B et al (2021) Susceptibility mapping of groundwater salinity using machine learning models. Environ Sci Pollut Res 28:10804–10817. https://doi.org/10.1007/s11356-020-11319-5

Motevalli A, Naghibi SA, Hashemi H et al (2019) Inverse method using boosted regression tree and k-nearest neighbor to quantify effects of point and non-point source nitrate pollution in groundwater. J Clean Prod 228:1248–1263. https://doi.org/10.1016/j.jclepro.2019.04.293

Müller A, Guido S (2016) Introduction to machine learning with Python. O'Reilly Media, Sebastopol

Muniappan A, Jarin T, Sabitha R et al (2023) Bi-LSTM and partial mutual information selection-based forecasting groundwater salinization levels. Water Reuse 13:525–544. https://doi.org/10.2166/wrd.2023.050

Neshat A, Pradhan B (2017) Evaluation of groundwater vulnerability to pollution using DRASTIC framework and GIS. Arab J Geosci 10. https://doi.org/10.1007/s12517-017-3292-6

Onyekwena CC, Xue Q, Li Q et al (2022) Support vector machine regression to predict gas diffusion coefficient of biochar-amended soil. Appl Soft Comput 127. https://doi.org/10.1016/j.asoc.2022.109345

Ouameur MA, Caza-Szoka M, Massicotte D (2020) Machine learning enabled tools and methods for indoor localization using low power wireless network. Internet Things (netherlands) 12:100300. https://doi.org/10.1016/j.iot.2020.100300

Ouedraogo I, Defourny P, Vanclooster M (2019) Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale. Hydrogeol J 27:1081–1098. https://doi.org/10.1007/s10040-018-1900-5

Pandey HK, Kumar Singh V, Kumar Singh S, Kumar Sharma S (2023) Mapping and validation of groundwater dependent ecosystems (GDEs) in a drought-affected part of Bundelkhand region, India. Groundw Sustain Dev 23:100979. https://doi.org/10.1016/j.gsd.2023.100979

Parzinger M, Hanfstaengl L, Sigg F et al (2022) Comparison of different training data sets from simulation and experimental measurement with artificial users for occupancy detection — using machine learning methods Random Forest and LASSO. Build Environ 223:109313. https://doi.org/10.1016/j.buildenv.2022.109313

Pulido-Bosch A, Rigol-Sanchez JP, Vallejos A et al (2018) Impacts of agricultural irrigation on groundwater salinity. Environ Earth Sci 77:197. https://doi.org/10.1007/s12665-018-7386-6

Qu L, Lu H, Tian Z et al (2024) Spatial prediction of soil sand content at various sampling density based on geostatistical and machine learning algorithms in plain areas. CATENA 234:107572. https://doi.org/10.1016/j.catena.2023.107572

Quinlan JR (1992) Learning with continuous classes. In: 5th Australian joint conference on artificial intelligence, vol. 92, Singapore, pp 343–348

Rahmati O, Choubin B, Fathabadi A et al (2019) Predicting uncertainty of machine learning models for modelling nitrate pollution of groundwater using quantile regression and UNEEC methods. Sci Total Environ 688:855–866. https://doi.org/10.1016/j.scitotenv.2019.06.320

Reghais A, Drouiche A, Ugochukwu E et al (2024) Compositional data analysis (CoDA) and geochemical signatures of the terminal complex aquifer in an arid zone (northeastern Algeria). J African Earth Sci 210:105162. https://doi.org/10.1016/j.jafrearsci.2023.105162

Richards LA (1954) Diagnosis and improvement of saline and alkali soils. In: Agriculture handbook No. 60, US Department of Agriculture, Washington, DC. https://doi.org/10.1097/00010694-195408000-00012

Sahour H, Gholami V, Vazifedan M (2020) A comparative analysis of statistical and machine learning techniques for mapping the spatial distribution of groundwater salinity in a coastal aquifer. J Hydrol 591:125321. https://doi.org/10.1016/j.jhydrol.2020.125321

Saleh MH, Alkhawaldeh RS, Jaber JJ (2023) A predictive modeling for health expenditure using neural networks strategies. J Open Innov Technol Mark Complex 9:100132. https://doi.org/10.1016/j.joitmc.2023.100132

Şen Z (1989) Cumulative semivariogram models of regionalized variables. Int J Math Geol 21(3):891–903

Şen Z (2019) Groundwater recharge level estimation from rainfall record probability match methodology. Earth Syst Environ 3:603–612. https://doi.org/10.1007/s41748-019-00130-z

Shanker MS, Hu MY, Hung MS (1996) Effect of data standardization on neural network training. Omega 24:385–397. https://doi.org/10.1016/0305-0483(96)00010-2

Silverman BW, Jones MC (1989) E. Fix and J.L. Hodges (1951): An important contribution to nonparametric discriminant analysis and density commentary on fix and Hodges (1951). Int Stat Rev 57:233–247

Suleymanov A, Tuktarova I, Belan L et al (2023) Spatial prediction of soil properties using random forest, k-nearest neighbors and cubist approaches in the foothills of the Ural Mountains, Russia. Model Earth Syst Environ 9:3461–3471. https://doi.org/10.1007/s40808-023-01723-4

Tachi A, Metaiche M, Messoul A et al (2023) Forecasting groundwater quality parameters using machine learning models: a case study of Khemismiliana Plain, Algeria. Dokl Earth Sc 512:907–914. https://doi.org/10.1134/S1028334X23600792

Theng D, Bhoyar KK (2024) Feature selection techniques for machine learning: a survey of more than two decades of research. Knowl Inf Syst 66:1575–1637. https://doi.org/10.1007/s10115-023-02010-5

Tran DA, Tsujimura M, Ha NT et al (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 Indic 127:107790. https://doi.org/10.1016/j.ecolind.2021.107790

Verma B, Prasad R, Srivastava PK et al (2022) Investigation of optimal vegetation indices for retrieval of leaf chlorophyll and leaf area index using enhanced learning algorithms. Comput Electron Agric 192:106581. https://doi.org/10.1016/j.compag.2021.106581

Wang LX, Mendel JM (1992) Generating fuzzy rules by learning from examples. IEEE Trans Syst Man Cybern 22:1414–1427. https://doi.org/10.1109/21.199466

Wang C, Wang K, Liu D et al (2024) Development and application of a comprehensive assessment method of regional flood disaster risk based on a refined random forest model using beluga whale optimization. J Hydrol 633. https://doi.org/10.1016/j.jhydrol.2024.130963

Xiao C, Ji Q, Chen J et al (2023) Prediction of soil salinity parameters using machine learning models in an arid region of northwest China. Comput Electron Agric 204. https://doi.org/10.1016/j.compag.2022.107512

Zamri N, Pairan MA, Azman WNAW et al (2022) River quality classification using different distances in k-nearest neighbors algorithm. Procedia Comput Sci 204:180–186. https://doi.org/10.1016/j.procs.2022.08.022

Zhang GL, Liu F, Song XD (2017) Recent progress and future prospect of digital soil mapping: a review. J Integr Agric 16:2871–2885. https://doi.org/10.1016/S2095-3119(17)61762-3

Zhang X, Shen H, Huang T et al (2024) Improved random forest algorithms for increasing the accuracy of forest aboveground biomass estimation using Sentinel-2 imagery. Ecol Indic 159:111752. https://doi.org/10.1016/j.ecolind.2024.111752

Zhao D, Zheng D, Wu S et al (2007) Climate changes in northeastern China during last four decades. Chin Geogr Sci 17:317–324. https://doi.org/10.1007/s11769-007-0317-1

Zhu X, Liang Y, Tian Z et al (2021) Simulating soil erodibility in southeastern China using a sequential Gaussian algorithm. Pedosphere 31:715–724. https://doi.org/10.1016/S1002-0160(20)60021-2

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.