Acta Geochim
https://doi.org/10.1007/s11631-024-00689-z

ORIGINAL ARTICLE

Harnessing machine learning tools for water quality assessment in the Kebili shallow aquifers, Southwestern Tunisia

Zohra Kraiem[1] · Kamel Zouari[1] · Rim Trabelsi[1]

Received: 27 December 2023 / Revised: 4 March 2024 / Accepted: 19 March 2024

Abstract An integrated method that implements multivariate statistical analysis and ML methods to evaluate groundwater quality of the shallow aquifers of the Djerid and Kebili district, Southern Tunisia, was adopted. An evaluation of their suitability for irrigation and/or drinking purposes is necessary. A comprehensive hydrochemical assessment of 52 samples with entropy weighted water quality index (EWQI) was also proposed. Eleven water parameters were calculated to ascertain the potential use of those resources in irrigation and drinking. Multivariate analysis showed two main components with Dim1 (variance = 62.3%) and Dim.2 (variance = 22%), due to the bicarbonate, dissolution, and evaporation and the intrusion of drainage water. The evaluation of water quality has been carried out using EWQI model. The calculated EWQI for the Djerid and Kebili waters (i.e., 52 samples) varied between 7.5 and 152.62, indicating a range of 145.12. A mean of 79.12 was lower than the median (88.47). From the calculation of EWQI, only 14 samples are not suitable for irrigation because of their poor to extremely poor quality (26.92%). The bivariate plot showed high correlation for EWQI ~ TH (r = 0.93), EWQI ~ SAR(r = 0.87), indicating that water quality depended on those parameters. Different ML algorithms were successfully applied for the water quality classification. Our results indicated high prediction accuracy (SVM > LDA > ANN > kNN) and perfect classification for kNN, LDA and Naive Bayes. For the purposes of developing the prediction models, the dataset was divided into two groups: training (80%) and testing (20%). To evaluate the models' performance, RMSE, MSE, MAE and R^2 metrics were used. kNN ($R^2 = 0.9359$, MAE = 6.49, MSE = 79.00) and LDA (accuracy = 97.56%; kappa = 96.21%) achieved high accuracy. Moreover, linear regression indicated high correlation for both training ($R^2 = 0.9727$) and testing data (0.9890). This well confirmed the validity of LDA algorithm in predicting water quality. Cross validation showed a high accuracy (92.31%), high sensitivity (89.47%) and high specificity (95%). These findings are fundamentally important for an integrated water resource management in a larger context of sustainable development of the Kebili district.

Keywords Water-resources management · Multivariate analysis · Machine learning · Kebili and Djerid shallow aquifers · EWQI · Water classification

1 Introduction

In arid and semi-arid regions, the population rely mainly on groundwaters. These resources are fundamentally important for several applications, mainly irrigation and drinking. It can be considered as the most available for humans (Isaac Sajan and Christopher 2023). Overexploitation of available and non-renewable resources, in conjunction with climatic changes, and rising agriculture and industry demands have led to challenging water resource management (Dassi 2011; Tarki et al. 2012; Khalfi et al. 2021a; Goodarzi et al. 2022). Thus, assessing water quality is an important issue for the sustainable management of the available resources. Some strategies and approaches to evaluate water potential use and valorisation are based on numerous indicators, of which the most used index is the

✉ Zohra Kraiem
zohra.kraiem@enis.tn

[1] Laboratory of Radio-Analyses and Environment, National Engineering School, University of Sfax, 1173, 3038 Sfax, Tunisia

Published online: 30 April 2024

water quality index (WQI). It was developed to summarise the pollution status and quality of a given water body for domestic and drinking purposes. Water quality index was usually calculated to eclectically evaluate the suitability of these resources for the desired use. The intricate tapestry of water quality necessitates accurate and efficient assessment methods. Traditionally, water quality indices (WQIs) have served as valuable tools, aggregating diverse parameters into a single, interpretable score. However, the limitations of conventional WQIs, including subjective weighting schemes and inadequate consideration of complex relationships between parameters, are increasingly recognized. This burgeoning field witnesses a transformative convergence of WQI development and machine learning (ML) techniques. Recent studies like Uddin et al. (2023b) and Sajib et al. (2023) showcase the development of novel WQI models using data science approaches and ML-powered groundwater assessment tools. Notably, Uddin et al. (2024), Sajib et al. (2023) delve into optimizing existing WQI models through ML, highlighting the potential for enhanced accuracy and predictive power. This section delves into this exciting intersection of WQIs and ML. We explore the latest advancements in developing data-driven WQI models: Examining how researchers utilize ML algorithms to build robust and adaptable WQIs that capture the intricate dynamics of water quality (Uddin et al. 2023a; Sajib et al. 2023). We can cite how optimizing traditional WQI models such as how investigating ML techniques are employed to refine existing WQIs, addressing limitations like subjectivity and static weighting (Uddin et al. 2023b, 2023c). The main event is how to predict water quality and how ML algorithms are harnessed to forecast water quality changes, enabling proactive management and pollution prevention (Georgescu et al. 2023).

Among various sources of water on Earth, groundwater can be considered the most available for humans (Isaac Sajan and Christopher 2023). Large quantities are mainly consumed by increased agricultural and industrial demands. Those applications may exert a deleterious effect on water. Thus, assessing water quality is an important issue for the sustainable management of the available resources. In arid and semi-arid regions, the population rely mainly on groundwater to supply drinking and agriculture needs.

According to Kumari and Rai (2020), more than 60% of groundwater is being pumped for drinking purposes; 20% for irrigation, and about 15% for industrial applications. Those overall percentages vary considerably with a huge stress on the hosting aquifers. Active measures to monitor groundwater in desert areas are usually shaped by local geo-environmental features and climate (i.e., hydrogeology, structures, evaporation, rock-water interaction, and weathering). In addition, the application of chemical fertilizers and harmful pesticides in agriculture has affected groundwater quality (Kumari and Rai 2020). Tunisia has been working for decades to develop strategies and action plans to rationalize the process of managing water resources. Gaaloul (2011) collected the available datasets on water resources in Tunisia. He stated that southern districts suffer from low annual rainfall (150 mm/year) and high evapotranspiration (≥ 2500 mm/year). Such irregular rainfall come to worsen an already dire situation of aridity and desertification-prone soils. Groundwater resources are mainly embedded in the southern underground exceeding 30% (221.36 million m^3) of the total phreatic aquifers reserves and 59.7% (722.2 106 m^3) in the deep aquifers of the Sahara Aquifer System (i.e., Continental Intercalaire and Complex terminal). The water of Plio-quaternary shallow aquifers is being used for irrigation and drinking purposes (Gaaloul 2011). He concluded that water issue, in the next 30 years, requires flexible and forward-thinking measures that may assist a thoughtful assessment, planning, conservation, and management. Subsequent trials were made to evaluate water resources status in the Kebili district, Southern Tunisia, where date palm agriculture is dominating the scenery with more than 1.4 million palm trees. Besser and Hamed (2019) ran a systematic analysis of water samples from the well-known Cretaceous aquifers. They recommended immediate prevention actions to predict and measure water quality coupled with systematic monitoring of the available resources in Kebili, especially oilfields.

Similarly, Haj-Amor et al. (2017) undertook a comprehensive study of the impact generated by excessive irrigation of date palms with low-quality water on soil salinity and shallow groundwater properties. They focussed on Douz oasis where low-quality irrigation water has led to high soil salinization risk. They urged local authorities and concerned farmers to develop effective strategies for water and soil management in the context of sustainable development. Another work was conducted by Dhaouadi et al. (2020) to explore the geochemical properties of water and its suitability for agriculture. The Sahara Aquifer System (SAS) was found to be affected by different anthropogenic factors. They concluded that the groundwater chemistry of the "continental intercalaire" cannot be used for agricultural purposes. In their recent work on ecologic and economic perspectives for sustainable agriculture in the arid climate of southern Tunisia, Besser et al. (2021) recommended some tips to prevent land degradation because of inefficient water management. Their attention has been directed to the SAS deep aquifer; they urged institutional and political actions for rehabilitation efforts. A study by Tarki et al. (2012) found that the water quality of the shallow aquifer in Chott Djerid had huge spatial variation from location to location. Water quality was generally better in the piedmont of the Dahar Mountains, where the aquifer is recharged by fresh groundwater. Water quality was worse in the areas closer to Chott Djerid, where the aquifer is more saline. The study also found that water quality of the shallow aquifer in Chott Djerid was declining over time, perhaps because of the increasing demand for water, degradation of the natural environment, and the discharge of untreated wastewater.

The study concluded that urgent action is needed to improve water quality of the shallow aquifer in Chott Djerid. In addition, Haj-Amor et al. (2017) studied the effects of excessive irrigation of date palm on soil salinization, and groundwater of the Kebili shallow aquifers. Later on, the impact of irrigation water quality and farmers' practices on agriculture sustainability in Kebili arid land was evaluated by Dhaouadi et al. (2020)

These studies provide further insights into the status of the Kebili shallow aquifers and the challenges they faces. They also provide recommendations for how to manage the aquifer sustainably.

Water quality is a measure of the physical, chemical, and biological characteristics of water. It is important for both human health and the environment. Water quality is especially important for irrigation and drinking water, as it can directly affect human health and crop yields.

There are different methods for water quality assessment; the most common include physical properties of water (e.g., temperature, turbidity, and color), chemical methods (concentration of chemicals in water, such as dissolved oxygen, pH, and nutrients), biological methods (measure the abundance of aquatic organisms) and remote sensing methods.

The best method for water quality assessment depends on the specific purpose of the assessment. For example, if the goal is to assess the suitability of water for drinking, then physical and chemical methods would be most appropriate. In some cases, a combination of methods may be used to get a more comprehensive assessment of water quality.

Tunisia is facing several water quality challenges, including pollution from agriculture, industry, and sewage. Artificial intelligence (AI) is a promising tool for helping Tunisia monitor and manage its water resources more effectively. Little is being done to use AI for water quality prediction in Tunisia. Nevertheless, attempts to use AI to predict water quality are being made. For instance, Msaddek et al. (2022) used AI for fractured aquifer vulnerability mapping. Smida et al. (2023) developed a model that can predict the quality of groundwater used for irrigation. The system uses integrated data from hydrogeological, hydrogeochemical, and GIS to develop a spatial database that was, in turn, processed by multicriteria analysis and artificial neural network to predict the risk of groundwater contamination. Tunisian government is also exploring the use of AI for water quality prediction. In 2022, the government launched a project to develop a national water quality monitoring and prediction system using AI. The system is expected to be operational in 2024. However, these studies considered the use of AI models trained on specific datasets; it might not generalize well to different regions or water bodies with unique characteristics, like Kebili aquifers. This can lead to inaccurate predictions. Those complex AI models were difficult to interpret, understand, and to apply in a generalized predictive strategy. Therefore, we proposed effective AI models with high-quality data for training and validation. Data collection on water quality followed strict guidelines for sampling and handling in the Kebili area (Kraiem et al. 2014). The choice of AI algorithms holds great promise for improving water quality monitoring and management in Tunisia, especially in the Kebili aquifers. We used both supervised and unsupervised algorithms for a comprehensive classification/prediction of water samples from the Kebili dataset. These are simple and handy tools for an accurate prediction of the Kebili water quality.

AI-based water quality prediction systems have the potential to help Tunisia to (1) identify and address water quality problems, (2) optimize water management and allocation and (3) protect public health and the environment.

However, some challenges need to be addressed before that AI-based water quality prediction systems can be widely deployed in Tunisia. One challenge is the lack of high-quality water quality data. Despite these challenges, the potential benefits of using AI for water quality prediction in Tunisia are significant. AI can help Tunisia to better manage its water resources and to protect public health and the environment.

AI-based water quality prediction systems, that can be used in Tunisia, include (1) water quality forecasting (An AI algorithm to forecast water quality conditions for future dates); (2) pollution source tracking (tracking the sources of water pollution), and (3) an early warning system for water pollution. An AI-based system could be used to monitor water quality and to alert managers to potential pollution events. This would allow managers to take preventive measures to protect public health and the environment, and develop targeted pollution control measures and water management plans for informed decisions about water use.

One study, conducted by the National Agency for Water and Wastewater, found that the water table in the Kebili shallow aquifer has declined by an average of 2 m per year since 1980. The study also found that the concentration of salinity in the aquifer has increased by an average of 10 mg per liter per year since 1980.

Another study, conducted by Gaaloul (2011), found that the Kebili shallow aquifer is vulnerable to pollution from agricultural runoff and industrial wastewater. The study also found that the aquifer is at risk of salinization.

Improving water quality of Chott Djerid is, therefore, a complex challenge, but it is essential to protect the environment and human health in the region. Despite the challenges, there are several ongoing efforts to improve water quality of Chott Djerid. These efforts are being led by the Tunisian government, as well as by international organizations such as the United Nations Environment Programme. The goal is to make Chott Djerid a more sustainable and liveable environment. The declining water table and the deteriorating water quality in the Kebili shallow aquifer pose some challenges, including reduced availability for drinking and irrigation. This is a problem for the growing population and the agricultural sector. The increasing salinity of water makes it unsuitable for drinking and irrigation. Therefore, adequate measures have been proposed to address the challenges facing the Kebili and Djerid shallow aquifers (More efficient water management practices, reducing pollution, and protecting the aquifer from drainage water intrusion (Kraiem et al. 2012).

To meet those challenges, AI has the potential to be a valuable tool for helping Tunisia monitor and manage its water resources more effectively. In this context, the present study has been undertaken to propose (1) an integrated approach of geochemical analysis of 52 samples, (2) evaluation of EWQI model and traditional water quality parameters (i.e., WQI, KR, SAR, MR, PI, TH, RSC, SSP, RSBC, NaR) and (3) application of convenient algorithms was proposed. Machine learning tools were applied to the hydrogeochemical properties and relevant indices for an accurate prediction of groundwater quality.

2 Study area and site description

Chott Djerid is a salt lake located in the south-central part of Tunisia. It is the largest salt lake in Tunisia and the third-largest salt lake in Africa. The lake is about 500 km long and 200 km wide. It is a terminal lake that has no outlet to the sea. The water in the lake is constantly evaporating, leaving behind salts and minerals, making the water very salty and alkaline, with high concentrations of dissolved solids, including calcium, magnesium, sodium, and chloride. Numerous oases that are supported by groundwater surround the lake. The water quality of the whole Chott Djerid depression is generally poor with high concentrations of dissolved solids.

In the Kebili area, shallow aquifers have been under increasing pressure in recent years due to population growth and increased agricultural activity. As a result, the water table has been declining and the quality of water has deteriorated. For instance, several pressure factors negatively affect water quality such as low rainfall (aridity), climate change (Uddin et al. 2021), intensive irrigation practices (Haj-Amor et al. 2017), overgrazing (Dhaouadi et al. 2021, 2022), inadequate sanitation and limited wastewater treatment infrastructure (Kachroud et al. 2019; Goodarzi et al. 2022; Uddin et al. 2022), among others.

Numerous studies have been conducted on the Kebili shallow aquifer to assess its status and to identify potential threats. These studies have shown that the aquifer is overexploited and water quality is declining.

The study area belongs to Southwestern Tunisia, surrounding the well-known Chott Djerid depression. It is bordered by the northern Chott range to the north, the Tebaga mountainous chain to the northeast, and the Tozeur uplift to the west; It opens on the oriental Erg southwards (Fig. 1).

3 Materials and methods

3.1 Samples collection and analysis

A set of 52 water samples were collected in the shallow aquifers of the Kebili district. The water sample was poured and stored in a 1.5 L polyethylene bottle and cooled to 4 ℃ until analysis for the desired physico-chemical parameters. Diluted nitric acid (10%) drops were added to obtain sub-samples for cationic (i.e., Na, K, Mg, Ca) analyses. The assessment of geochemical characteristics is crucial in developing strategies to protect water resources in this region. A clean, sanitized water sample bottle, a disinfectant, and a pair of gloves were used during the collection of water samples. Water was first filtered through a 0.45 µm pore size acetate filter. Then, *in-situ* measurements of electric conductivity (EC), temperature (T), and pH were evaluated using a HI 9828 Multiparameter meter portable device (Hanna Instruments, France). Chemical analysis was performed as described by Kraiem et al. (2014). Briefly, bicarbonates were ascertained by titration, as described elsewhere (Bradbury and Baeyens 2009). Major elements and nitrates concentrations were measured by a high-performance liquid chromatograph equipped with a Super-Sep column for anions.

To ensure a robust QA/QC measure, we followed a well-documented procedure; we proceeded to equipment calibration in each sampling campaign. Protocol, as described earlier, was meticulously designed from the comprehensive sampling technique to data treatment and management. Specific measures were undertaken for accuracy and precision, including analysis of blank samples as control, and measurements in triplicate to get errors and averages. The accuracy and precision of the measuring device were evaluated to a 95% confidence level (error $<$ 5%).

Regression metrics including root mean square error (RMSE), mean absolute error (MAE), and R-squared were used to evaluate the performance of models that predict EWQI values. Accuracy, sensitivity, specificity, and

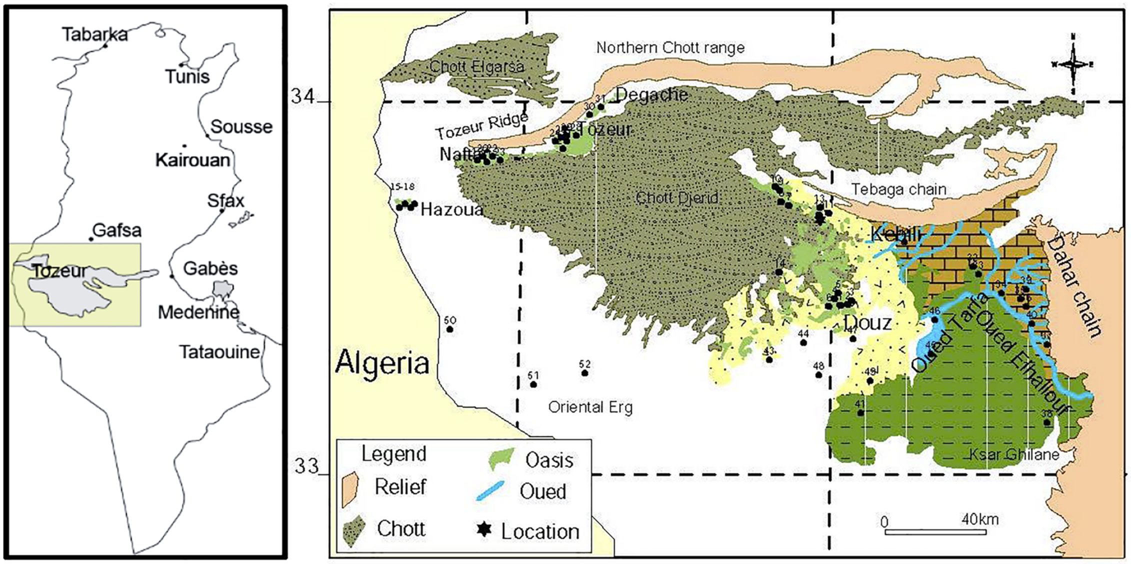


Fig. 1 Schematic geological map and location of the collected water samples

balanced accuracy (i.e., arithmetic mean of sensitivity and specificity) were also applied to test the quality and reliability of the models.

Positive Predictive Value (PPV) is an important metric used to evaluate the performance of diagnostic tests and classification models. It measures the probability that a positive test result indicates the presence of the condition. Negative Predictive Value (NPV) is another crucial metric used to evaluate diagnostic tests and classification models, alongside PPV (Çakir et al. 2023).

3.2 Water quality assessment

Water quality was assessed by the assignment of numerical values to various water quality parameters (i.e., water quality index). This is a useful tool for monitoring water quality over time and for comparing the water quality of different water bodies. Other water quality parameters included sodium absorption ratio (SAR), magnesium ratio (MR), Kelly's ratio (KR), sodium percentage (NR), permeability index (PI), residual sodium carbonate (RSC), potential salinity (PS), soluble sodium percentage (SSP) and residual sodium bicarbonate (RSBC). A detailed description of those indices can be found elsewhere (Hassen et al. 2016; Dhaouadi et al. 2020; Kumari and Rai 2020; Aouiti et al. 2021; Guasmi et al. 2022).

3.2.1 EWQI

EWQI is the method used in this study to summarize the quality of water by assigning numerical values to various water quality parameters (Kumar and Augustine 2022). The EWQI is then used to rank the water quality. The weighting factors for the EWQI are usually determined using a statistical technique called entropy, which is a measure of the uncertainty or randomness of a system. In the context of water quality, entropy can be used to measure the uncertainty in the distribution of water quality parameters.

EWQI is a valuable tool for water quality assessment because it provides a simple and concise way to communicate the overall quality of water. It can also be used to compare water quality between different water bodies and over time. It can be calculated as per Kumar and Augustine (2022).

An initial step should be allocated to the construction of the Eigenvalue matrix X based on the number of water samples (m) and parameters (n).

$$X = \begin{vmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{vmatrix} \tag{1}$$

Then, a standardization of data is needed to cancel the effects of magnitude and dimensions.

$$y_{i_j} = \frac{x_{i_j} - (x_{i_j})_{\min}}{(x_{i_j})_{\max} - (x_{i_j})_{\min}} \tag{2}$$

The standard matrix can be constructed as:

$$Y = \begin{vmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mn} \end{vmatrix} \tag{3}$$

The index value (*Pij*) and information entropy (e_j) can be calculated.

$$P_{i_j} = \frac{1 + y_{i_j}}{\sum_{i=1}^{m} (1 + y_{i_j})} \tag{4}$$

$$e_j = \frac{1}{\ln m} \sum_{i=1}^{m} P_{ij} \times \ln P_{ij} \tag{5}$$

Quality rating (q_j) and entropy weight (W_j) of j^{th} parameter:

$$q_j = \frac{Cj}{Sj} \times 100 \tag{6}$$

where Sj: drinking water standard (WHO 2017) and Cj: Value of the chemical parameter.

$$W_j = \frac{1 - e_j}{\sum_{i=1}^{m} (1 - e_j)} \tag{7}$$

Finally, the entropy-weighted water quality index (*EWQI*) can be given by the equation.

$$EWQI = \sum W_i \times q_i \tag{8}$$

The choice of the EWQI for Chott Djerid water quality assessment offers several advantages over other methods, particularly considering the unique characteristics of this hypersaline and endorheic lake (Lee et al. 2022; Sangare et al. 2023). That is, unlike traditional methods relying on expert-assigned weights, EWQI assigns weights based on the information inherent in the water quality data, eliminating subjective bias to reflect the actual contribution of each parameter to overall water quality (Ahmed et al. 2020; Bharani Baanu and Jinesh Babu 2022; Goodarzi et al. 2022). Moreover, EWQI allowed us to consider a wide range of water quality parameters specific to Chott Djerid, including salinity, specific ions (i.e., anions and cations), and relevant physical-chemical parameters. This adaptability is crucial for capturing the complex nature of this ecosystem (Najah Ahmed et al. 2019; Islam and Mostafa 2022). While relatively new, EWQI has shown promising results in various water quality assessments. Its data-driven approach and adaptability make it a potential leader in future water quality monitoring, particularly for unique environments like Chott Djerid. EWQI's unique advantages make it a compelling choice for comprehensive and data-driven evaluation of this unique and sensitive ecosystem (Aouiti et al. 2021b; Oğuz and Ertuğrul 2023).

3.2.2 *Multivariate statistical analysis*

A set of statistical methods including principal component analysis (PCA) and cluster analysis were applied as multivariate statistical tools to analyse the collected datasets with more than one independent variable and one dependent variable. PCA and cluster analysis are the most common multivariate statistical analyses (Sahu et al. 2021). PCA is often used to identify the most important variables. Hierarchical cluster analysis was used for grouping water samples together based on their similar water quality parameters (Aouiti et al. 2021; Schreiber et al. 2022; Liu et al. 2023; Meegahakumbura and Nanayakkara 2023).

3.3 Machine learning models

Artificial intelligence tools were used to accurately predict and monitor the studied water quality. Various artificial intelligence techniques have been used for the accurate classification of groundwater quality (Khullar and Singh 2021). In this work, special attention has been directed to support vector machine (SVM), linear discriminant analysis (LDA), artificial neural networks (ANN), Naïve Bayes (NB), and k-nearest neighbors (kNN).

3.3.1 *Support Vector Machine (SVM)*

SVM is a supervised learning algorithm that can be used for both classification and regression tasks (Behzad et al. 2009). However, it is most used for classification. SVMs work by finding a hyperplane in the feature space that separates the data points into two classes with the largest possible margin. The margin is the distance between the hyperplane and the closest data points from each class. SVMs are very powerful algorithms that can be used to solve a wide variety of problems (Khalil et al. 2005). They can be used to solve both linear and non-linear problems with an easy interpretation (Che Nordin et al. 2021). SVM is used in this study as a powerful machine learning tool to assess the quality of the Kebili PQ waters.

SVM has been used for groundwater quality evaluation by Khalil et al. (2005) who outlined interesting algorithms for predicting groundwater quality (i.e., artificial neural network, support vector machines, locally weighted projection regression, and relevance vector machines). The obtained results indicated rigorous application of those machine learning tools to build models with strong predictive capabilities.

SVM was found to have higher prediction efficiency than other relevant AI-based models (ANN and ANN-GA models). It is an effective ML classifier that may significantly contribute to a better groundwater quality assessment (Che Nordin et al. 2021).

3.3.2 *Linear discriminant analysis*

Linear discriminant analysis (LDA) is a supervised machine learning algorithm used for classification and dimensionality reduction. It finds a linear combination of features that best discriminates between two or more classes (Zhu et al. 2022a, b). LDA works by first computing the mean vector and covariance matrix for each class. Then, it calculates the between-class scatter matrix and the within-class scatter matrix. The between-class scatter matrix measures how different the classes are from each other, while the within-class scatter matrix measures how similar the data points within each class are to each other. LDA has been used for decades due to its successful supervised feature extraction. A brief recap of the LDA was provided by Zhu et al. (2022a, b). They stated that LDA samples from the same class should be independently and identically distributed, as the case in the present dataset. LDA is a powerful classification algorithm that finds its successful application in face recognition (Benkhaira and Layeb 2020; Wan et al. 2023), medical imagery diagnosis (e.g., X-rays and MRIs), and text classification (Lin et al. 2010). It is relatively easy to implement and can be used in both small and large datasets. Thus, it has been applied to the present water samples.

3.3.3 *Artificial neural network (ANN)*

ANN works by adjusting the weights of the connections between the nodes in the network. The weights are adjusted so that the network learns to produce the desired output for a given input (Hameed et al. 2017). The backpropagation algorithm is an iterative process that starts by randomly assigning weights to the connections. Then, the network is presented with a training example. The output of the network is compared to the desired output, and the weights are adjusted accordingly. This process is repeated for several training examples until the network learns to produce the desired output for all of the training examples. The network converges when the error is minimized. The learning rate is a parameter that controls how much the weights are updated in each iteration of the backpropagation algorithm. A high learning rate can make the network converge faster, but it can also make the network more prone to overfitting (Zhu et al. 2022a, 2022b; Tao et al. 2022; Bashar et al. 2023). The more training examples, the better the network will perform. Here a large enough dataset was collected to run ANN efficiently.

ANN is, thus, a powerful tool for learning complex relationships between input and output data. Data points were split into training (80%) and test (20%) datasets for model training and evaluation. The back propagation algorithm was used for ANN training as it performs smoothly to provide faster convergence with a lower iteration number (Keskin et al. 2015; Zaqoot et al. 2018; Al-Mukhtar 2021; Islam Khan et al. 2022; Zhu et al. 2022a, 2022b).

3.3.4 *k nearest neighbour (kNN)*

K-nearest neighbors (kNN) is another simple, supervised machine learning algorithm that can be used for both classification and regression tasks. It works by finding the K most similar data points in the training set to a new data point, and then using the labels or values of those K neighbors to predict the desired property of the new data point. To classify a new data point using kNN, the algorithm first calculates the distance between the new data point and each data point in the training set (Wan et al. 2023).

The algorithm predicts the label of the new data point by taking a majority vote of the K nearest neighbors. If most of the K nearest neighbors have a certain label, then that label is predicted for the new data point.

To perform regression using kNN, the algorithm works similarly, but instead of predicting a label, it predicts a value. The algorithm predicts the value of the new data point by taking the average of the values of the K nearest neighbors.

kNN is a simple, effective, and popular algorithm for machine learning tasks; it is often used as a baseline algorithm to compare machine-learning algorithms.

According to Çakir et al. (2023), kNN is a controlled machine learning algorithm that can be applied for classification and regression. They have successfully applied numerous algorithms to detect the main physical–chemical parameters of water in aquaculture.

They concluded that kNN achieved high accuracy (i.e., 91.3%), but somewhat lower than the SVM model. k-nearest neighbors measure the closest distance between data points using distance functions (e.g., Euclidean, Manhattan, and Minkowski distance).

3.3.5 *Naïve Bayes*

Naive Bayes is a simple but surprisingly powerful probabilistic machine learning algorithm used for predictive modeling and classification tasks (Taheri and Mammadov 2013). Based on Bayes' theorem, Naive Bayes classifiers assume that data points are independent, which is often not the case in real-world data. However, despite this assumption, Naive Bayes classifiers can often achieve high accuracy on classification tasks (Zhang et al. 2016).

To train a Naive Bayes classifier, the algorithm first calculates the probability of each class in the training set. The algorithm then calculates the probability of each feature given each class. Once the algorithm has been trained, it can be used to classify new data points by calculating the probability of each class given the features of the new data point (Zhang et al. 2016). The class with the highest probability is then predicted for the new data point. Naive Bayes is a powerful and versatile machine learning algorithm that was used, in the current study, to classify Kebili water samples according to several indices.

4 Results and discussions

4.1 Statistical analyses

Statistical analyses were used to evaluate the main relevant ratios that can estimate the real properties of a given water sample. Pairs plots of the main ratios were plotted based on Pearson correlation (Fig. 2).

The calculated water quality parameters included total hardness (TH), permeability index (IP), residual sodium carbonate (RSC), Kelly ratio (KR), magnesium ratio (MG), sodium absorption ratio (SAR), sodium percentage (NR), total dissolved solids (TDS), potential salinity (PS), soluble sodium percentage (SSP) and residual sodium bicarbonate (RSBC).

4.1.1 Descriptive statistics

Descriptive statistics indices are given in Table 1. It appeared that total hardness varied between 18.8 and 461; the IP index varied from 20.2 to 59.3 with a mean of 43.6. TDS values exceeded 360 mg/L to reach more than 11,000 mg/L, in some water points. This is an indication of high salinity. Thus, one should expect low-quality water. However, this statement needs to be confirmed by the calculated ratios and indices. Among 11 water parameters, special attention was devoted to SAR, NR, and MR due to their direct contribution to soil degradation (Hassen et al. 2016; Dhaouadi et al. 2020; Besser et al. 2021).

4.1.2 Pairwise correlations

Pairwise correlations of the calculated ratios, relevant to the shallow groundwater samples from Kebili, indicated that highly correlated ratios included TDS ~ PS (r = 0.98), KR ~ NR (r = 0.97), KR ~ SAR (r = 0.91), SAR ~ NR (r = 0.9), TH ~ RSBC (r = 0.90), TDS ~ RSBC (r = 0.89), PS ~ RSBC (r = 0.85), TH ~ TDS (0.84), IP ~ NR (r = 0.78), IP ~ KR (r = 0.78), NR ~ SSP(r = 0.78), SAR ~ PS(r = 0.74) and PS ~ SSP (r = 0.73). Negatively correlated ratios are RSC ~ TH (r = – 0.99), RSC ~ RSBC(r = – 0.88), RSC ~ TDS(r = – 0.84), RSC ~ PS(r = – 0.79) and RSC ~ SAR(r = – 0.76). High relationships between the studied ratios and parameters may suggest a strong interdependence (Fig. 2). This is a further confirmation of the statement given by Aouiti et al. (2021). Islam and Mostafa (2022) proposed six hazard classes to develop a custom water quality index for irrigation. The proposed WQI included the maximum number and type of water parameters to evaluate water quality for irrigation purposes. This will corroborate with the findings of Parween et al. (2022). In the same context, Satish Kumar et al. (2016) performed a comprehensive study about groundwater quality for drinking and irrigation use. They determined the chemical composition of the studied water samples for possible evaluation of the suitability for irrigation. They stated that the use of shallow water for irrigation depended on several indices like TDS, EC, SAR, KR, RSC, SSP, IP, and WQI. They have, then, addressed each water parameter or ratio separately from the other without considering the possible interaction between those ratios. Based on the pairwise correlation of those parameters, one can draw further clarifying information about the studied samples' geochemical properties and subsequently the overall water quality index. Similarly, Dhaouadi et al. (2020) evaluated the ability of groundwater samples to supply drinking and irrigation needs. They targeted the complex terminal aquifer, which is a deep confined aquifer, lying below the current shallow PQ aquifer. They applied an in-depth assessment of water quality based on several ratios, as in the present study. A quick look at the correlation matrix may highlight the high correlation between TDS and PS, TDS, and TH due to the dissolution of Ca and Mg. SAR versus KR, NR, PS, and RSBC confirmed the interrelation between those indices and Ca, Mg, and the existing dissolved anions. In contrast, RSC exhibited too high negative correlation with TH, RSBC, TDS, PS, and SAR perhaps due to the low carbonates (HCO_3 and CO_3) when compared to Ca and Mg. Similar results were found by Dhaouadi et al. (2020) when they correlated RSC to SAR and NR, but to a much lower extent.

4.2 Saturation index

Saturation index (SI) is a measure of a solution's ability to dissolve or deposit a particular mineral. It is often used as an indicator of water quality. Water chemical composition should be evaluated for its saturation versus minerals. Dissolution and precipitation mechanisms may take place depending on the desired mineral concentration. If a mineral is below the threshold limit of a given concentration, it will be dissolved. An oversaturated water with respect to an element may stimulate its precipitation.

Calcite and aragonite are both calcium carbonate minerals, but aragonite is less stable than calcite. This means that aragonite will tend to convert to calcite over time, especially

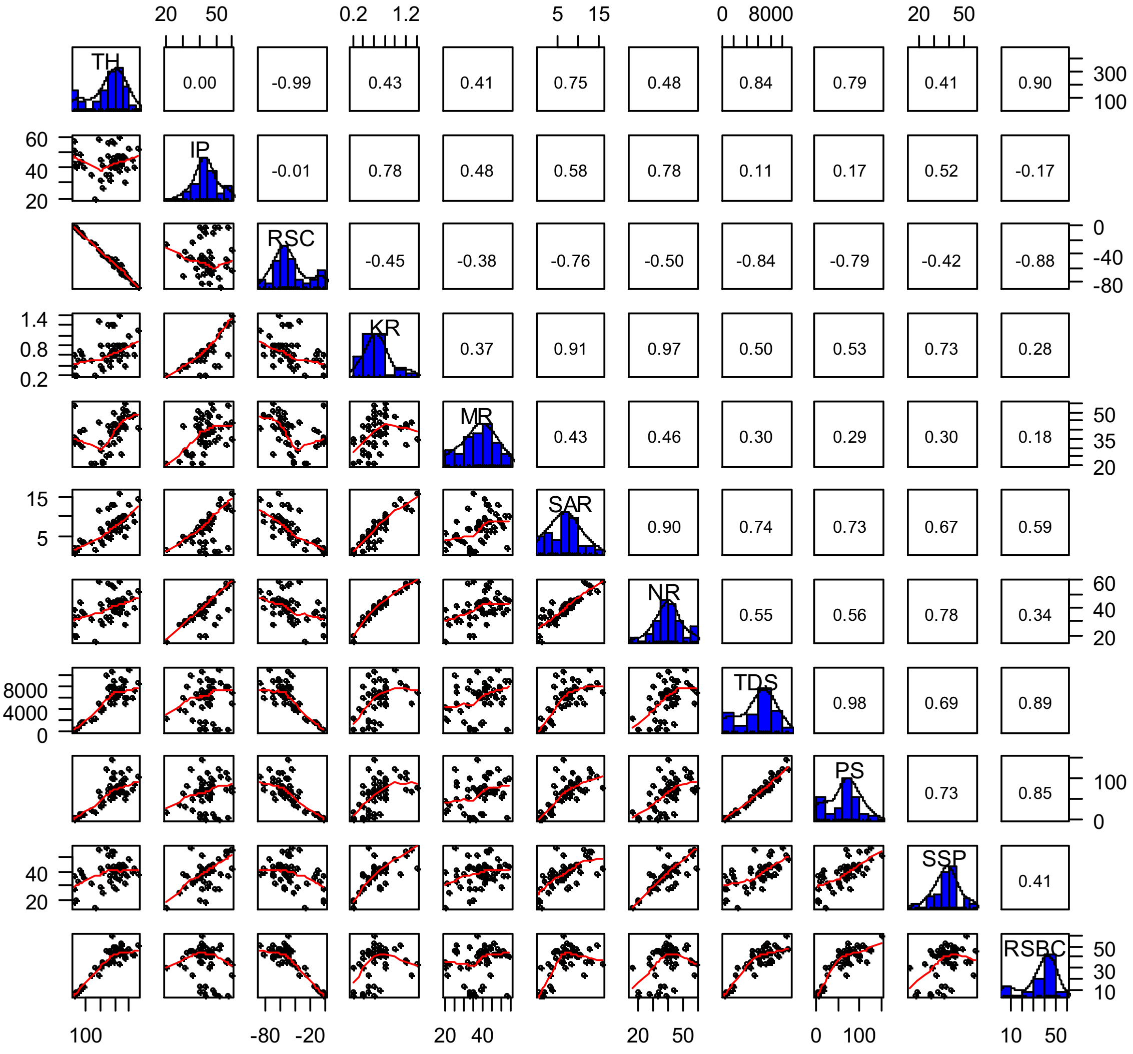


Fig. 2 Pairwise correlations of water quality ratios

Table 1 Descriptive statistics for the studied water parameters

Vars	n	Mean	SD	Median	Trimmed	MAD	Min	Max	Range	Skew	Kurtosis	SE
TH	52	256.65	118.50	291.15	264.73	87.32	18.80	461	442.20	−0.68	−0.51	16.43
IP	52	43.60	7.97	43.55	43.74	5.34	20.21	59.32	39.11	−0.24	0.43	1.11
RSC	52	−46.68	22.31	−51.60	−48.12	16.31	−86.42	−2.00	84.42	0.65	−0.41	3.09
KR	52	0.67	0.26	0.61	0.65	0.22	0.22	1.41	1.21	0.69	0.42	0.04
MR	52	38.23	8.68	38.90	38.51	8.60	20.61	54.52	33.91	−0.28	−0.64	1.20
SAR	52	6.95	3.66	6.95	6.83	3.63	0.73	15.93	15.20	0.23	−0.48	0.51
NR	52	40.00	9.35	40.15	40.09	7.71	15.90	58.81	42.91	−0.22	0.13	1.30
TDS	52	5833.01	2895.43	6405.02	5971.76	2461.12	360.03	11,010	10,650.03	−0.50	−0.72	401.52
PS	52	64.06	36.12	67.55	64.01	29.95	1.72	149.42	147.70	−0.11	−0.57	5.01
SSP	52	38.58	9.21	39.02	38.77	7.19	14.31	57.62	43.31	−0.29	0.20	1.28
RSBC	52	35.39	14.66	40.70	36.91	9.12	3.82	59.63	55.81	−0.96	−0.24	2.03

in warm water. A positive relationship between aragonite and calcite; dolomite and calcite and, dolomite and aragonite may further confirm their common origin (Fig. 3). Significant correlations ($p \leq 0.001$) were observed for dolomite-aragonite-calcium and gypsum-anhydrite-halite, which mean their interdependence. Gypsum and anhydrite are both calcium sulfate minerals, but anhydrite is more stable than gypsum. This means that gypsum will tend to convert to anhydrite over time, especially in dry environments.

A positive SI indicates that the solution is oversaturated with respect to the mineral and precipitation is likely to occur. A negative SI indicates that the solution is undersaturated with respect to the mineral and dissolution is likely to occur.

Calculated values of saturation indices for calcite, aragonite, anhydrite, gypsum, and halite are shown in Fig. 4. SI values >0 are indicative of precipitation. In other words, one can expect an oversaturated water with those minerals, leading to a precipitation process.

Xiao et al. (2015) addressed water quality via the assessment of dissolution and precipitation in natural water from an extremely arid area in China. They calculated high SI values for dolomite, calcite, and aragonite, as the case in the present study. Figure 4 shows supersaturated water samples with dolomite, calcite, and aragonite. Our results indicated that SI ranges were as follows: anhydrite (− 2.09 to − 0.11; range = 1.98); aragonite (− 0.52 to 1.05, range = 1.57); calcite (− 0.38 to 1.19; range = 1.57) and dolomite (− 0.77 to 2.5; range = 3.27). These values are used for precipitation process. In contrast, halite showed a negative saturation index (SI.Hal. from − 7.8 to − 3.8; range = 4), indicative of favorable dissolution of NaCl. The only equilibrium state was observed for gypsum (SI.Gyp. value from − 1.87 to 0.10; range = 1.97) and with its boxplot at the zero line (SI = 0 means equilibrium). A similar interpretation was adopted by Hassen et al. (2016) who applied multivariate statistical analysis to evaluate water quality from the Jelma shallow watershed.

4.3 Multivariate analysis of water quality

A PC (noted Dim here) is defined based on eigenvalue > 1. Eigenvalue and variance (in %) may give a clear idea about each component's contribution. Dim1 contributed an eigenvalue of 6.85 with a variance of 62.3% highly explained by two main groups of indexes that reflect two real-time processes. The first cluster is associated with RSBC (8.46%, r = 0.76, r2 = 0.58), TDS (11.74%, r = 0.9, $r^2 = 0.8$), PS (11.62%, r = 0.89, $r^2 = 0.8$) and TH (10.49%, r = 0.85, $r^2 = 0.72$). This can be primarily explained by the bicarbonate dissolution and its effects on salinity increase. The second cluster is associated with SAR (12.84%, r = 0.94, $r^2 = 0.88$), NR (10.08%, r = 0.83, $r^2 = 0.69$), KR (9.16%, r = 0.79, $r^2 = 0.63$), SSP (8.71%, r = 0.77, $r^2 = 0.6$) because of dissolution, evaporation and the intrusion of drainage water (Kraiem et al. 2012). Surprisingly, the RSC ratio was negatively correlated to Dim1 (10.54%,

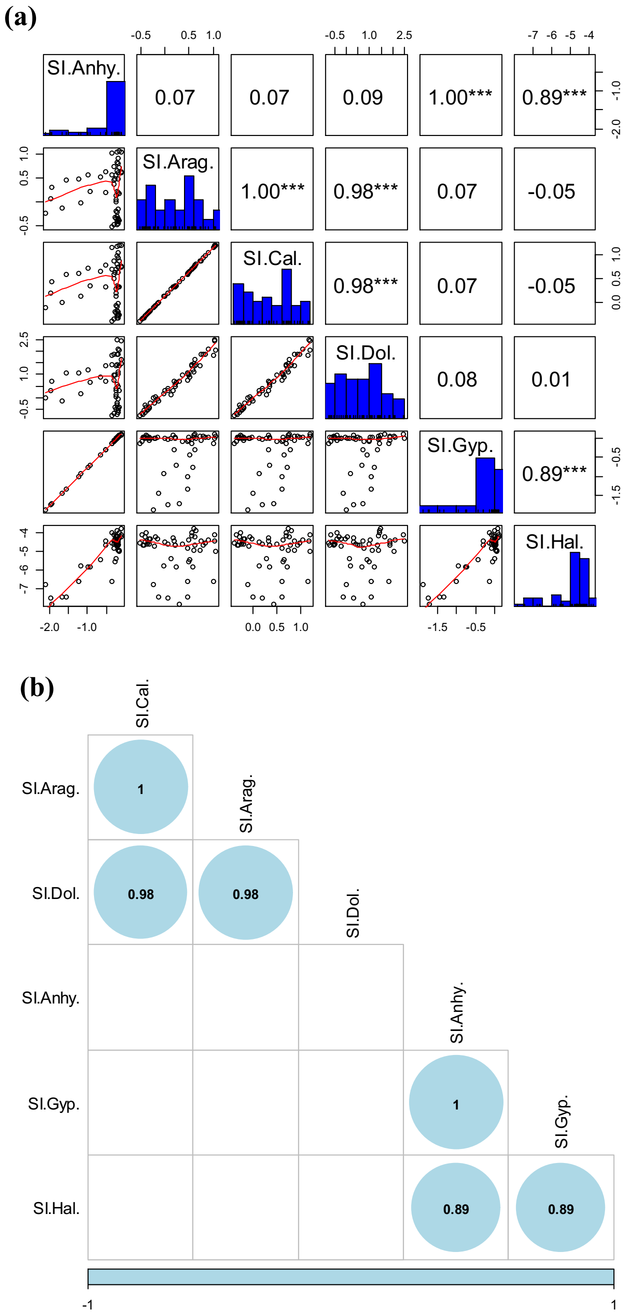


Fig. 3 Pairs representation of saturation indices for all minerals (**a**) and significantly correlated indices (**b**; $p = 0.001$)

r = – 0.85, r^2 = 0.72), reflecting its weak contribution to mineralization probably because of the unstable carbonates states that were tightly related to pH value. Dim2 was controlled by IP (30.48%, r = 0.86, r^2 = 0.74) perhaps related to the structure of the underlying soil. It is well known that sandy soil has high enough permeability to allow the percolation of infiltration water (Khalfi et al. 2021b; Kraiem et al. 2024). Thus, only Dim1 and Dim2 were shown in the PCA presentation (Fig. 5). Isaac and Siddiqui (2022) adopted a similar interpretation when they applied multivariate statistical techniques to evaluate water quality around Yamuna River, India.

4.4 Water quality classification

SAR values varied between 0.7 and 15.9 (range = 15.2) with a mean of 6.95 (SD = 3.66). The density plot of the studied water samples showed a positively skewed (right-tailed distribution) and platykurtic features (Table 1). SAR is an important parameter for assessing the suitability of water for irrigation, as high SAR water can lead to soil salinization and reduced crop yields. Fortunately, this is not the case with the present water samples. Almost all samples showed two main groups with 44 excellent samples and only 8 as good for irrigation (Fig. 6). The violin plot of the whole dataset (i.e., 52 water samples) showed a higher distribution of around 7.5 for the "excellent" group and around 13 for the "good" samples.

Sodium/alkali hazard is a deleterious threat to soil permeability and structure (Satish Kumar et al. 2016). MR index varied between 0.2 and 1.4. Out of 51 samples, only 4 samples are not suitable for irrigation as per the MR ratio. It is well known that magnesium excess in irrigation water negatively affects the soil alkalinity. This was observed in water samples K6, K21, K23, and K41 which represented less than 10% of the collected samples. Alkalinity hazard can be assessed by sodium percentage (NR) ranging between 15.9% and 58.8% with a mean of 40 and median of 40.15%. Classification of water samples based on the NR index showed 50% "good" and 50% "permissible" water quality. This is an indication of the short residence time of water, low dissolution of minerals from the housing lithology, and the rationale use of chemical fertilizers. Kumari and Rai (2020) attributed higher concentrations of Na in groundwater to longer residence time, high dissolution, and abundant fertilizers input.

According to TDS values (Table 1), most of the samples are unsuitable for irrigation (41 samples), 2 samples are doubtful (K39 and K43), three samples are permissible (K38, K44, and K48) and 6 samples belonged to the "good" class (i.e., K34–37, K40, and K41). This indicated that, based on TDS values, more than 76% are not recommended for irrigation. PS also indicated that 40 samples are unsuitable for irrigation. It exceeded the threshold limit of 35% by 78.4%, ranging between 1.7 and 149.4. This was further confirmed by the SSP index which showed 34 samples as "unsuitable" and 18 samples were found suitable for irrigation purposes. Similarly, the RSBC index showed a minimum of 3.8 (> 2.5), indicating "unsuitable" irrigation water. This information corroborated well with the study of Dhaouadi et al. (2020) (Table 2).

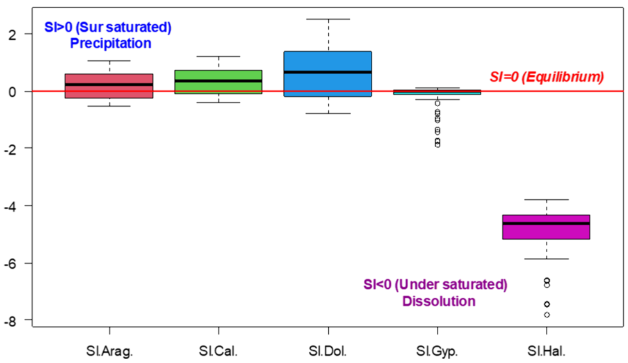


Fig. 4 Variation of SI values for aragonite, calcite, dolomite, and halite minerals

4.4.1 EWQI

WQIs are valuable tools for simplifying complex water quality data into a single, easy-to-understand value. However, several issues can affect their accuracy and reliability. Advantages of EWQI over other indices reside in the fact that EWQI addresses some of the limitations of traditional WQI methods by incorporating information theory concepts. It uses entropy to assign weights to individual parameters based on their variability in the dataset. This reduces subjectivity in weight selection compared to assigning arbitrary weights. It has the potential for mitigating eclipsing by assigning higher weights to parameters with greater variability. The choice of the most suitable WQI for a specific application depends on various factors like data availability, water quality concerns, and intended use of the results. In this context, EWQI offers potential advantages in addressing model uncertainty and potentially mitigating eclipsing. The evaluation of water quality has been conducted using EWQI as detailed by Kumar and Augustine (2022). They explained the methodology of EWQI calculation based on relative WHO standards (WHO 2017). The calculated EWQI for Kebili water samples (i.e., 52 water samples) varied between 7.5 and 152.62, indicating a range of 145.12 (Fig. 7). A mean of 79.12 was lower than the median (88.47); it showed

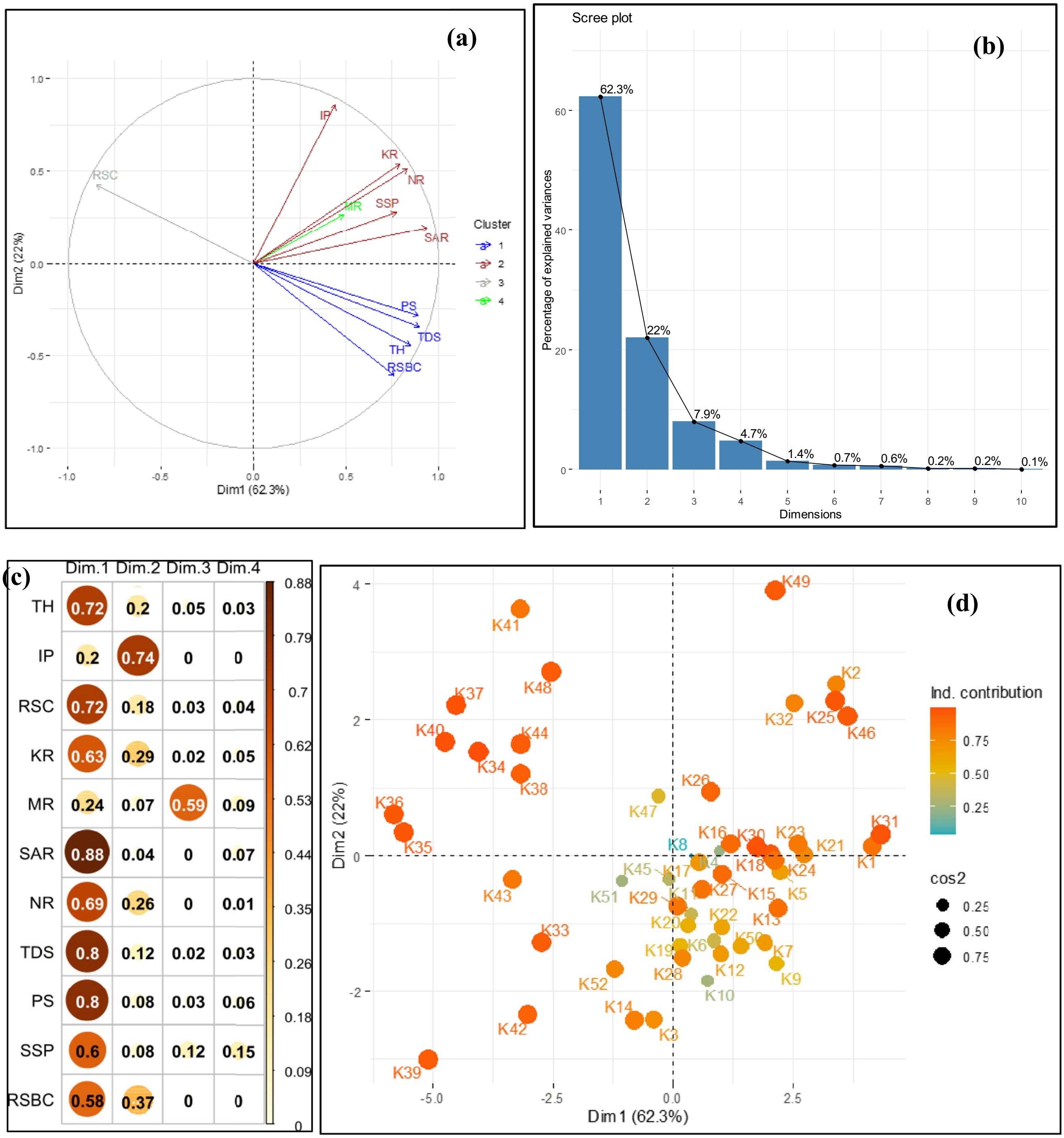


Fig. 5 PCA of water quality parameters (**a**), scree plot (**b**), correlation to Dim1, Dim2 and Dim3 components, coefficient of determination (cos2, **c**) and individual contribution of the studied water samples (**d**)

a negative skewness and platykurtic distribution. This is a confirmation of the previously discussed quality indices. From the calculation of EWQI, eight samples were of "excellent" class (K34–37, K40, K41, K44, and K48), 3 "good" samples (K38, K39, and K43), and 27 medium-quality samples. Thirteen poor quality (K1, K2, K4, K7, K13, K18, K21, K23, K24, K25, K31, K32 and K46) and 1 extremely poor sample (K5). Therefore, out of fifty-two water samples, only 14 samples are not suitable for irrigation because of their poor to extremely poor quality (26.92%). The lowest values indicated higher quality. The bivariate plot for the thirteen studied parameters can be

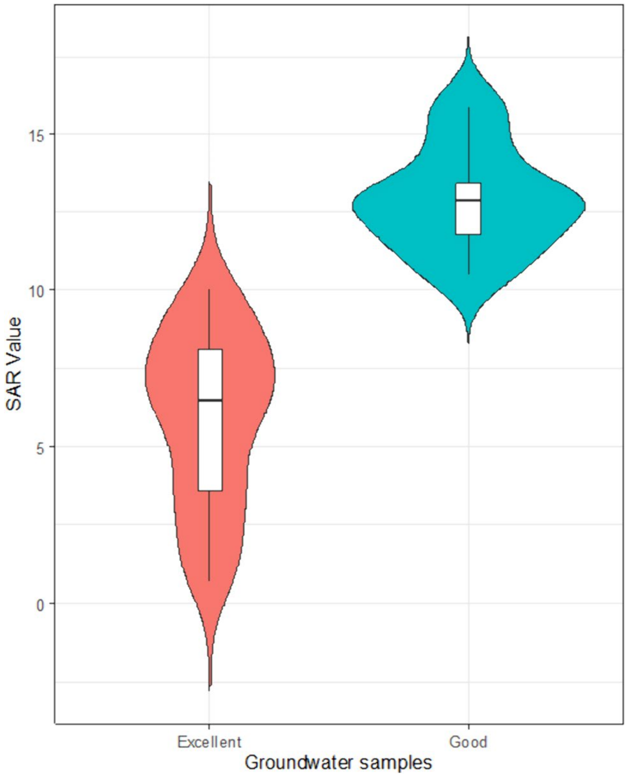


Fig. 6 Groundwater samples classification according to SAR

Table 2 Classification of the studied water samples based on several ratios

Index	Class	Number of samples	Percentage (%)
SAR	Excellent	44	84.62
	Good	8	15.38
TH	Suitable	52	100
IP	Permissible	51	98
	Good	1	2
RSC	Safe	52	100
MR	Suitable	47	90.38
	Unsuitable	5	9.62
NR	Good	26	50
	Permissible	26	50
TDS	Good	6	11.54
	Permissible	3	5.77
	Doubtful	2	3.85
	Unsuitable	41	78.85
PS	Permissible	11	21.15
	Unsuitable	41	78.85
SSP	Suitable	18	36.62
	Unsuitable	34	65.38
RSBC	Marginal	1	2
	Unsuitable	51	98

found in Fig. 8. A high positive correlation was observed for EWQI ~ TH (r = 0.93), and EWQI ~ SAR (r = 0.87). A somewhat lower correlation of EWQI with TDS (r = 0.84) and RSBC (r = 0.8) can be seen. Other ratios exhibited significant correlation with EWQI, but to a much lower extent. In contrast, an extremely high negative relationship between EWQI and RSC ratio (r = – 0.94). The tight correlation of TH and SAR with EWQI indicated that the water quality depended on those parameters; they are very well distributed in the southern edge of the studied area. Most of the medium to excellent-quality waters lay within this sub-area of Kebili, except for a few wells (e.g., K5). The poor samples are mostly collected from the northern edge of Chott Djerid, near Tozeur (Figs. 7, 8).

4.5 Contribution of ML algorithms to the prediction of water quality

Machine learning algorithms have made significant contributions to water quality assessment in recent years. They offer several advantages over traditional methods due to their accuracy, efficiency, and flexibility. This allows them to make more accurate predictions than traditional methods. ML algorithms can be used to assess a wide range of water quality parameters, including physical, chemical, and biological indicators (Abuzir and Abuzir 2022; Hirvasniemi et al. 2023; Lu et al. 2023; Rashid and Kumari 2023; Zhen et al. 2023). They can also be used to assess water quality in different types of water bodies (Khullar and Singh 2021). ML algorithms have been used to develop a variety of tools for water quality assessment (Zaqoot et al. 2018). As mentioned earlier, we have applied both prediction and classification models to deepen knowledge about the actual groundwater status in the arid region of southern Tunisia.

4.5.1 SVM algorithm

SVM model has been performed on 39 samples and 12 predictors with radial basis function Kernel. No pretreatment of the dataset has been undertaken that was cross validated tenfold while taking a constant tuning parameter "sigma" (0.704). The application of the confusion matrix to the predicted data indicated an overall high accuracy (92.31%; balanced accuracy = 98.48%) with the highest sensitivity (100%) and specificity (96.97%) to the "Excellent" class. Similarly, the SVM classifier showed 100% sensitivity and 95% specificity to the "Medium" class water samples (balanced accuracy was 97.50%). In the third position, "Poor" quality samples exhibited high balanced accuracy (93.28%) with 90% of poor-quality water samples predicted correctly. True negatives (i.e., “non-poor” quality samples) were correctly predicted to a rate of 96.55%.

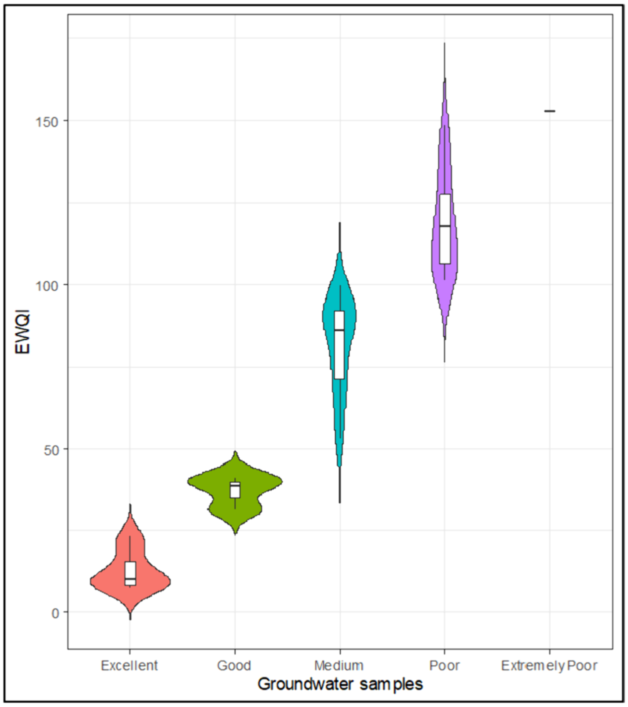


Fig. 7 Groundwater samples classification according to EWQI

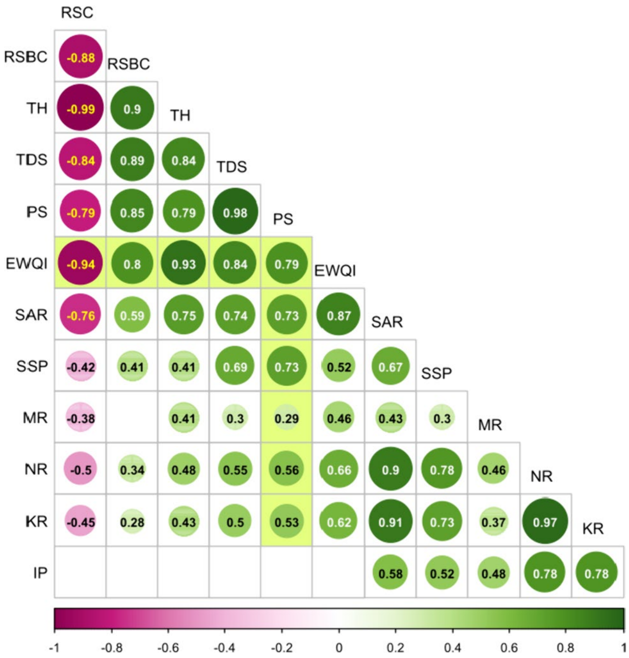


Fig. 8 Correlation plot of the studied groundwater parameters (EWQI correlation is highlighted)

Table 4 summarizes the main testing criteria for SVM model performances. Balanced accuracy (i.e., the arithmetic means of sensitivity and specificity) exceeded 83.33% for all water classes, except the "Extremely poor" class which showed only 50%. The predicted water quality, given in Table 1, showed 6 true positive predicted samples for the Excellent class with one sample predicted as "Good" (called false negative predicted). This generates a lower sensitivity for the "Good" class, which showed 2 samples as true positive and one false predicted from the "Excellent" class. This means that only two-thirds were truly predicted as good-quality samples, corresponding to the sensitivity (i.e., 66.67%). The overall prevalence was highest in "Medium" (48.72%), "Excellent" (15.38%) and "Good" (7.69%) classes. Nevertheless, "Poor" quality samples represented 25.64% with sensitivity of 90% and specificity of 96.55%. The kappa statistic test of 0.88 suggested an almost perfect agreement with the observed results (McHugh 2012; Table 3).

4.5.2 ANN prediction

An artificial network algorithm is a powerful tool for learning complex relationships between input and output data. In this work, 10 input neurons (i.e., water quality ratios), 1 hidden layer with 4 neurons, and one output (EWQI) were used for ANN model construction (Fig. 9). Data points were split for training (70%) and test (30%) datasets for model training and evaluation. The back propagation algorithm was used for ANN training as it performs smoothly to provide faster convergence with a lower iteration number (Keskin et al. 2015; Zaqoot et al. 2018; Al-Mukhtar 2021; Islam Khan et al. 2022; Zhu et al. 2022a, 2022b). For instance, Keskin et al. (2015) applied ANN for the prediction of water quality through 13 geochemical parameters. They successfully predicted the property of water samples by using one hidden layer with five nodes which is slightly different from the present ANN structure. The simple ANN structure has given high accuracy and sensitivity, as in the case of the present study. The present model has converged after 89 steps with a threshold of 0.009 s and very low error (0.037), indicating a high efficiency. Moreover, linear regression indicated a high correlation for both training ($R^2 = 0.9727$) and testing data ($R^2 = 0.9890$; Fig. 10). This is well correlated with cross validation tests that showed a high accuracy (92.31%), high sensitivity (89.47%), and high specificity (95%). A Kappa value of 0.8458 confirmed the validity of the prediction model.

4.5.3 Linear discriminant analysis (LDA)

LDA is a supervised learning algorithm that can be used for both classification and dimensionality reduction. It works by finding a linear combination of features that best discriminates between two or more classes. LDA is a simple algorithm, but it can be very effective in practice. Fitting the LDA algorithm to the studied sample water indicated an almost perfect agreement with experimental data. Overall

Table 3 Contingency table of the main testing criteria (in %)

Criterion	Class				
	Excellent	Good	Medium	Poor	Extremely poor
Prediction	6 (1)	2	19 (1)	9 (1)	0
Sensitivity	100.00	66.67	100.00	90.00	0.00
Specificity	96.97	100.00	95.00	96.55	100.00
Positive Predictive Value	85.71	100.00	95.00	90.00	-
Negative Predictive Value	100.00	97.30	100.00	96.55	97.44
Prevalence	15.38	7.69	48.72	25.64	2.56
Balanced Accuracy	98.48	83.33	97.50	93.28	50.00

accuracy was 97.56% with a kappa test of 96.21%, confirming the validity of the LDA algorithm in predicting water quality. Prevalence of the water samples was 14.63%, 7.32%, 51.22%, 21.95%, and 2.44% for "Excellent", "Good", "Medium", "Poor" and "Extremely poor" classes, respectively. This is almost the same results obtained from the experiments (Fig. 11).

4.5.4 k nearest neighbour (kNN)

Variables used for the kNN regression model were first centred and scaled. The cured dataset included 44 groundwater samples. Tenfold cross validation test, RMSE, and R^2 were used to check the accuracy of the kNN model. It can be seen, from Fig. 12, that a high correlation between the measured and kNN-predicted data ($R^2 = 0.9359$, MAE = 6.49, MSE = 79.00). The best-predicting model was the 5-nearest neighbour regression model (k = 5; RMSE = 8.8887), further

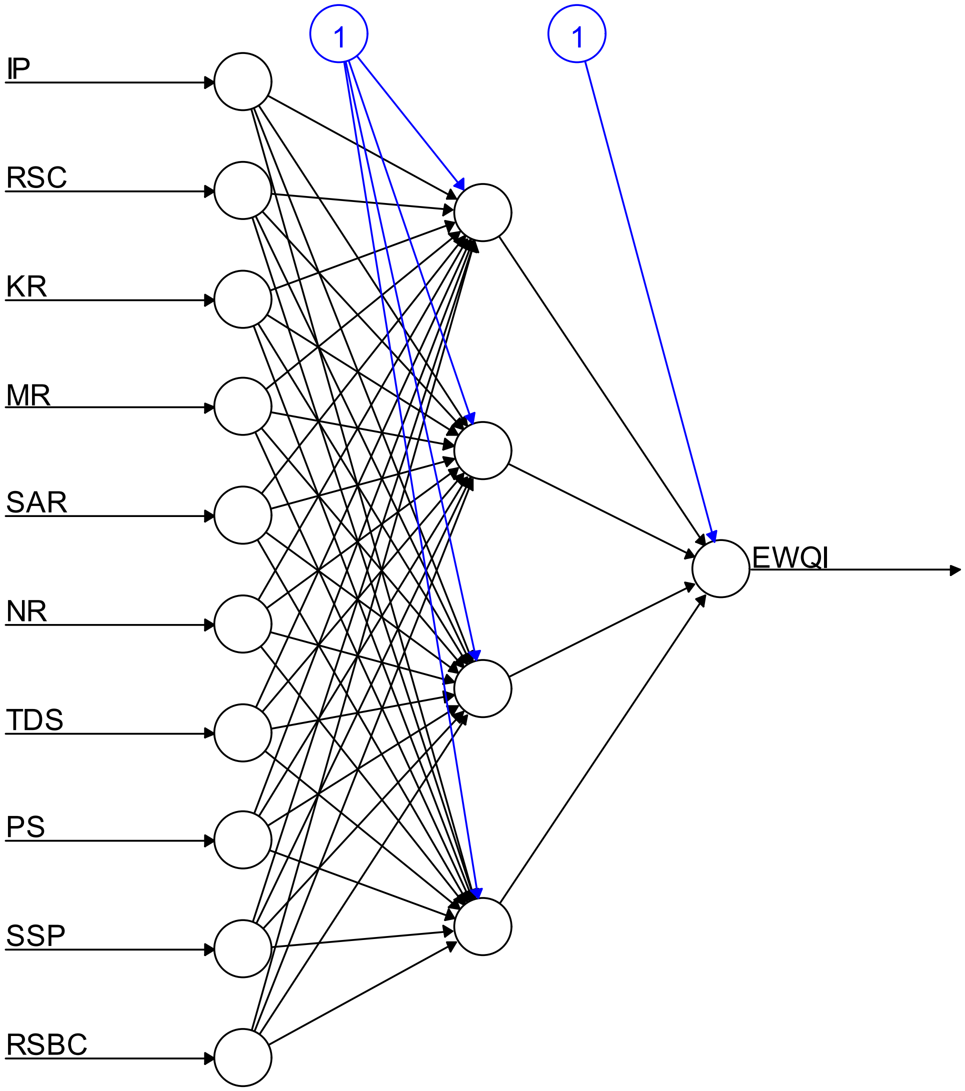
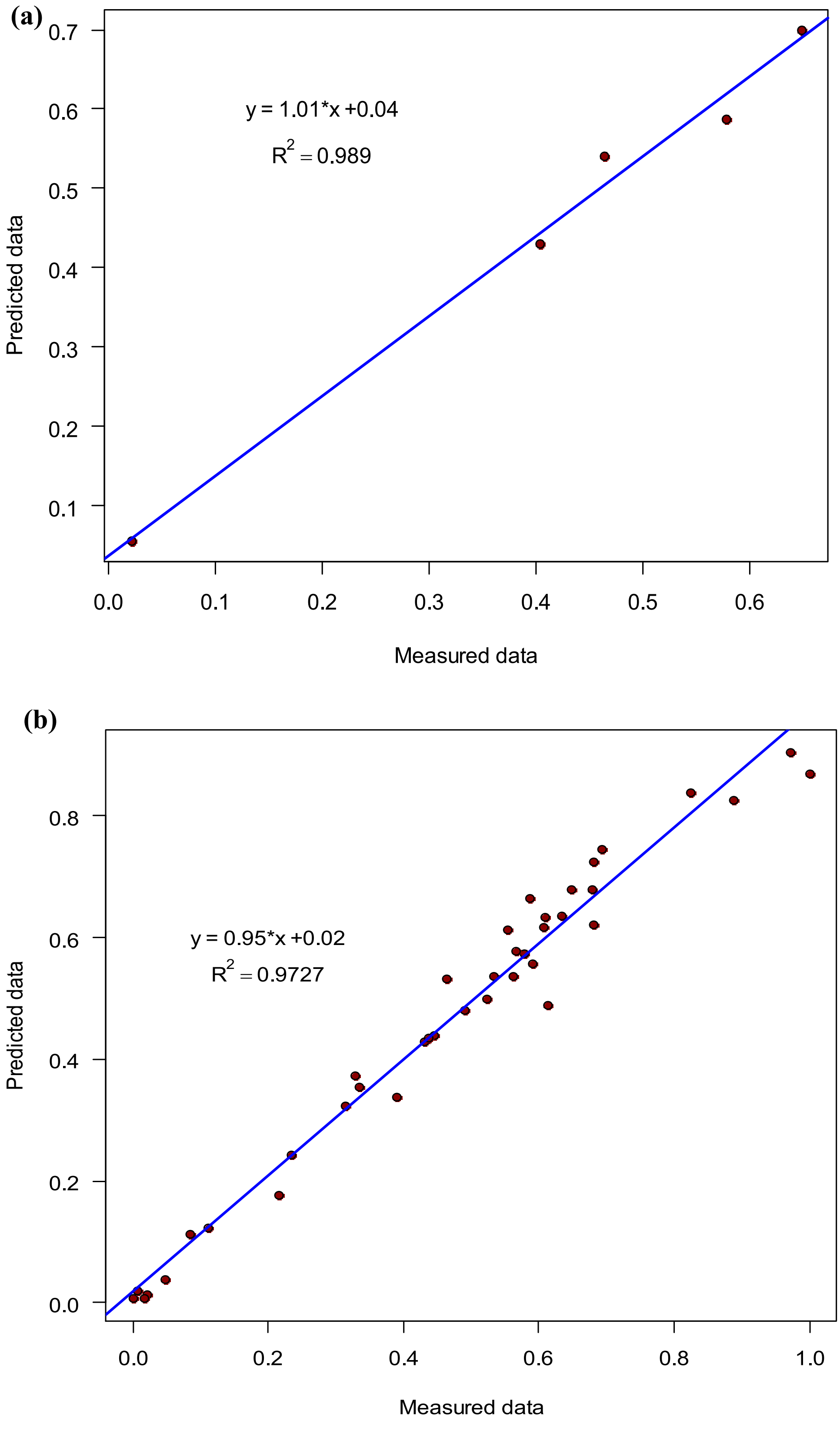


Fig. 9 ANN structure adopted to predict EWQI

Fig. 10 Fitting the predicted and measured data based on the adopted ANN model (a. testing data; b. training data)



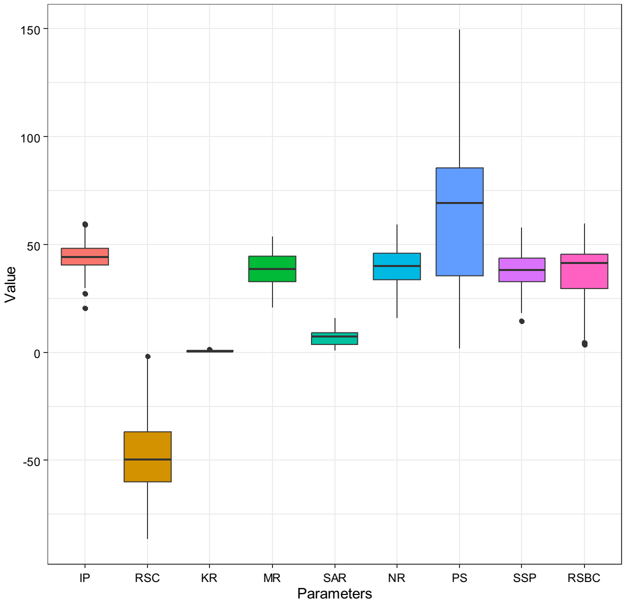


Fig. 11 Boxplots of the studied groundwater parameters after curation (i.e., elimination of outliers)

confirming the satisfactory fitting of the kNN model. Tang et al. (2019) developed a soft method for groundwater level prediction using machine-learning models. They constructed a least square support vector machine (LS-SVM) to predict the variation of groundwater level variation; the model was compared with kNN. They confirmed that LS-SVM and classical SVM offered the best performances in all cases. In addition, the kNN approach performed well, though to a somewhat lower extent. This is a confirmation of the data presented in this work. kNN classifier has also performed well with regards to water classification based on the EWQI index.

4.5.5 *Naïve Bayes (NB)*

The naive bayes classifier was used to distinguish the possible groundwater classes due to its explicit and sound theoretical basis that allows an optimal and robust assumption (Taheri and Mammadov 2013). The obtained classification fitted the experimental data to a high accuracy value, though somewhat lower than SVM and ANN. A glance at Table 4 may shed light on the high accuracy of the naïve Bayes algorithm in predicting the studied water classes. It appeared that, out of 40 water samples, 6 samples were “Excellent”, 3 samples as “Good”, 21 were “Medium” class samples, 9 were “Poor” quality, and one was “Extremely poor” (Fig. 13).

5 Machine learning models limitations

While machine learning holds great promise for EWQI determination, it shows variable shortcomings and limitations (Mashaly and Alazba 2019; Uddin et al. 2021; Thirumalai Raja et al. 2022; Adnan et al. 2023). One can recognize ML algorithms by their weak generalizability, overfitting, and interpretability (Parkavi et al. 2023). A brief description of the main controversial points is given below:

- Dependence on data quality since ML models are reflected by the quality of the data they were trained on. Poor quality or incomplete data can lead to inaccurate and unreliable EWQI predictions;
- Large amounts of high-quality data are often required to train an ML model effectively. This can be a challenge, especially for hardly accessible water bodies where data collection is limited;
- Model limitations included generalizability (models are generally trained on specific water body, but not generalized well to other locations with different environmental conditions);
- Overfitting the training data results in poor performance on new data;
- Implementing and maintaining complex ML models is an expensive approach; it requires specialized expertise.

Other limitations related to EWQI may include simplicity, subjectivity, and specificity (Islam Khan et al. 2022).

Despite these limitations, ML tools are valuable for WQI determination when used cautiously and in conjunction with other approaches. Continuous research and development are addressing these limitations, making ML increasingly valuable for water quality monitoring and management (Abuzir and Abuzir 2022; Islam and Mostafa 2022; Georgescu et al. 2023; Verma and Chaurasia 2023; Kraiem et al. 2024).

6 Conclusions

This study was performed for a comprehensive assessment of water quality through the application of several machine learning algorithms. Physico-chemical properties as well as several water quality indices have been measured accordingly. Traditional characterization of those water samples has been carried out to assess the main potential application of those natural resources. Machine learning methodology has been applied in this study as a handy tool to predict/evaluate the useability of the available water resources of the shallow aquifer extending below the Kebili area. Artificial intelligence has the potential to revolutionize groundwater quality assessment. This paves the way for developing new

Fig. 12 Fitting the predicted and measured data based on the adopted kNN model

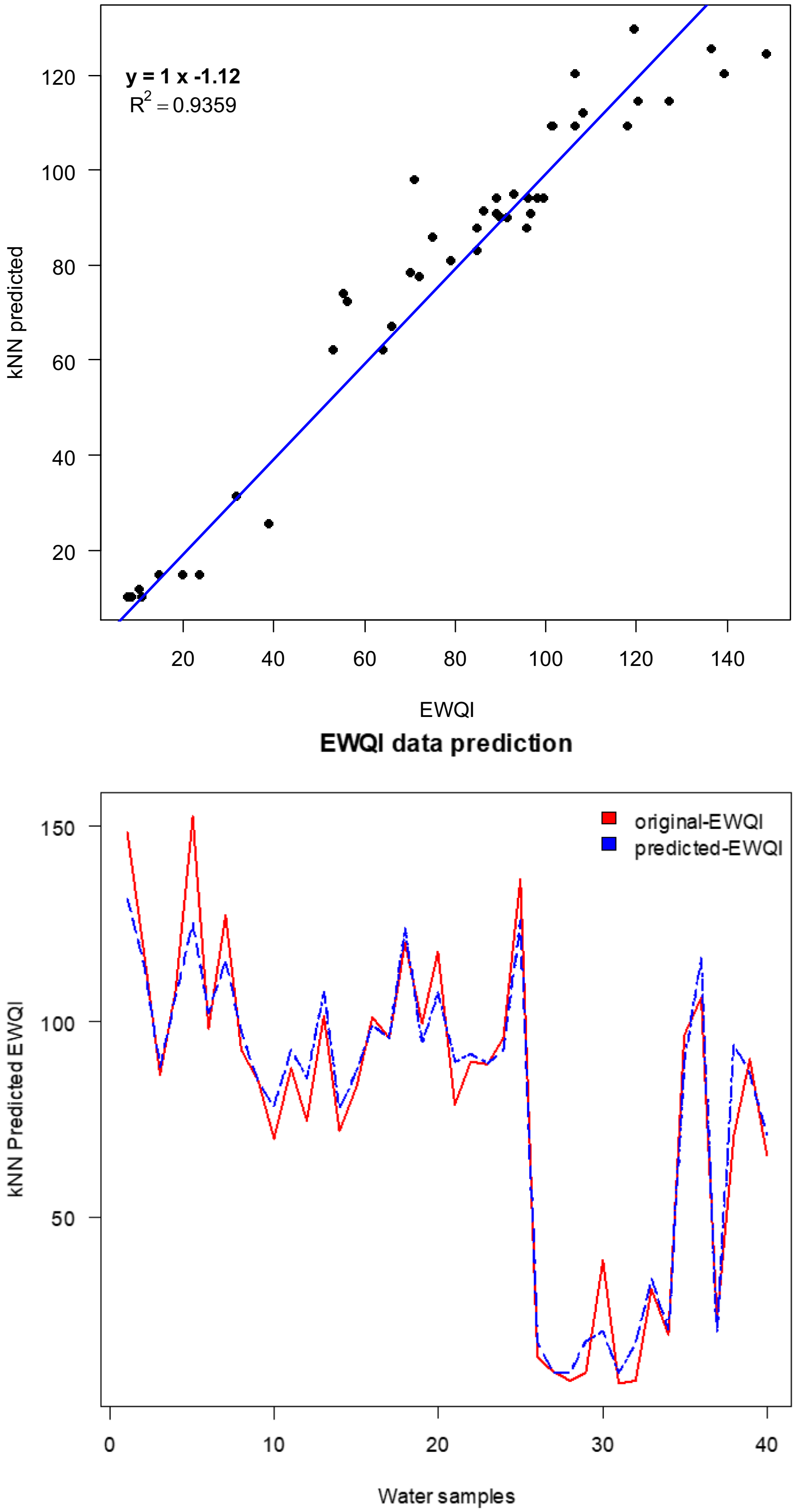


Table 4 Classification of water status versus EWQI range (Kumar and Augustine 2022)

N	EWQI	Status	Water samples		Naïve Bayes		LDA		SVM		kNN	
			Number	%	*n*	%	Pred	%	Pred	%	Pred	%
1	<25	Excellent	8	15.39	6	15	6	15.38	6 (1)	15.38	7	19.44
2	25–50	Good	3	5.77	3	7.5	3	7.69	2 (1)	5.13	3	8.33
3	50–100	Medium	27	51.92	21	52.5	21	53.85	19 (1)	48.72	18	50
4	76–100	Poor	13	25.00	9	22.5	8	20.51	9	23.07	8	22.22
5	>150	Extremely Poor	1	1.92	1	2.5	1	2.56	0	0	–	–
Total			52	100	40	100	39	100	39 (3)	92.3 (+7.7)	36	99.99

(3) False negative samples

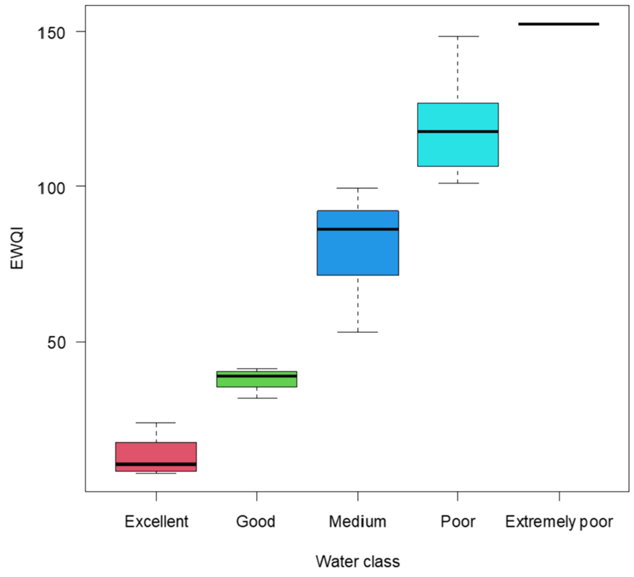


Fig. 13 Classification plot of the studied groundwater parameters based on EWQI predicted by Naïve Bayes algorithm

tools and techniques for assessing the risk of groundwater contamination and for more effective management. Machine learning models (i.e., SVM, KNN, Naïve Bayes, and LDA) were used to predict groundwater quality for a comprehensive evaluation of the likelihood of contamination in the southwestern Tunisian Djerid and Kebili shallow aquifers. The main findings of this study can be summarized in the following points:

- ML algorithms were used as handy tools to identify low-quality water points and to develop mitigation strategies;
- Classical multivariate analyses (descriptive statistics, PCA, and HCA) were first used for a systematic classification of the available dataset (i.e., chemical composition of 52 groundwater samples);
- Saturation indices were addressed accordingly to highlight the geochemical properties of the studied samples;
- An in-depth assessment of groundwater quality was carried out by the development of machine learning models. These models were trained on 52 groundwater samples to identify patterns and relationships;
- ML model results indicated high accuracy of all trained algorithms tested (high fittings/classification efficiencies of the proposed models) as confirmed by cross validation tests.

These models are fundamentally important for the prediction of groundwater quality in the well-known PQ shallow aquifer of the Kebili district, Tunisia. The study highlights the importance of these models for predicting water quality in the Kebili PQ aquifer. This is necessary for the groundwater quality monitoring approach. However, a more detailed assessment is still required for sustainable water quality management.

7 Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the first author used Google AI service to generate descriptions and polished text. After using this tool, the author reviewed and edited the content as needed and took full responsibility for the content of the publication.

Acknowledgements The authors extend their appreciation to Professor Md Galal Uddin from Civil Engineering, College of Science and Engineering, University of Galway, Ireland, for his careful and active contribution to the reviewing of the manuscript. His thoughtful comments and suggestions have undoubtedly improved the document submitted for evaluation.

Author contributions ZK: Formal analysis, Writing—original draft, Writing—review and editing. Conceptualization, Methodology, Supervision, Data Curation, Visualization, Validation; KZ, RT: Conceptualization, Methodology, Funding acquisition, Validation.

Funding This research received no external funding.

Data availability Data are available upon request.

Declarations

Conflict of Interest The authors declare no conflict of interest.

Informed consent Not applicable.

References

Abuzir SY, Abuzir YS (2022) Machine learning for water quality classification. Water Qual Res J. 57:152–164. https://doi.org/10.2166/wqrj.2022.004

Adnan RM, Dai HL, Kuriqi A, Kisi O, Zounemat-Kermani M (2023) Improving drought modeling based on new heuristic machine learning methods. Ain Shams Eng J. https://doi.org/10.1016/j.asej.2023.102168

Ahmed U, Mumtaz R, Anwar H, Mumtaz S, Qamar AM (2020) Water quality monitoring: From conventional to emerging technologies. Water Supply. 20:28–45. https://doi.org/10.2166/ws.2019.144

Al-Mukhtar M (2021) Modeling the monthly pan evaporation rates using artificial intelligence methods: A case study in Iraq. Environ Earth Sci. 80:1–14. https://doi.org/10.1007/s12665-020-09337-0

Aouiti S, Hamzaoui Azaza F, El Melki F, Hamdi M, Celico F, Zammouri M (2021) Groundwater quality assessment for different uses using various water quality indices in semi-arid region of central Tunisia. Environ Sci Pollut Res. 28:46669–46691. https://doi.org/10.1007/s11356-020-11149-5

Bashar AM, Nozari H, Marofi S, Mohamadi M, Ahadiiman A (2023) Investigation of factors affecting rural drinking water consumption using intelligent hybrid models. Water Sci Eng. 16:175–183. https://doi.org/10.1016/J.WSE.2022.12.002

Behzad M, Asghari K, Eazi M, Palhang M (2009) Generalization performance of support vector machines and neural networks in runoff modeling. Expert Syst Appl. 36:7624–7629. https://doi.org/10.1016/J.ESWA.2008.09.053

Benkhaira S, Layeb A (2020) Face recognition using RLDA method based on mutated cuckoo search algorithm to extract optimal features. Int J Appl Metaheuristic Comput. 11:118–133. https://doi.org/10.4018/IJAMC.2020040106

Besser H, Hamed Y (2019) Causes and risk evaluation of oil and brine contamination in the Lower Cretaceous Continental Intercalaire aquifer in the Kebili region of southern Tunisia using chemical fingerprinting techniques. Environ Pollut. 253:412–423. https://doi.org/10.1016/j.envpol.2019.07.020

Besser H, Dhaouadi L, Hadji R, Hamed Y, Jemmali H (2021) Ecologic and economic perspectives for sustainable irrigated agriculture under arid climate conditions: An analysis based on environmental indicators for southern Tunisia. J Afr Earth Sci. 177:104134. https://doi.org/10.1016/j.jafrearsci.2021.104134

Bharani Baanu B, Jinesh Babu KS (2022) Smart water grid: A review and a suggestion for water quality monitoring. Water Supply. 22:1434–1444. https://doi.org/10.2166/ws.2021.342

Bradbury MH, Baeyens B (2009) Sorption modelling on illite Part I: titration measurements and the sorption of Ni, Co, Eu and Sn. Geochim Cosmochim Acta. 73:990–1003. https://doi.org/10.1016/j.gca.2008.11.017

Çakir M, Yilmaz M, Oral MA, Kazanci HÖ, Oral O (2023) Accuracy assessment of RFerns, NB, SVM, and kNN machine learning classifiers in aquaculture. J King Saud Univ Sci. 35:102754. https://doi.org/10.1016/J.JKSUS.2023.102754

Che Nordin NF, Mohd NS, Koting S, Ismail Z, Sherif M, El-Shafie A (2021) Groundwater quality forecasting modelling using artificial intelligence: A review. Groundw Sustain Dev. 14:100643. https://doi.org/10.1016/j.gsd.2021.100643

Dassi L (2011) Investigation by multivariate analysis of groundwater composition in a multilayer aquifer system from North Africa: A multi-tracer approach. Appl Geochem. 26:1386–1398. https://doi.org/10.1016/j.apgeochem.2011.05.012

Dhaouadi L, Besser H, Wassar F, Kharbout N, Brahim NB, Wahba MA, Kang YK (2020) Agriculture sustainability in arid lands of southern Tunisia: Ecological impacts of irrigation water quality and human practices. Irrig Drain. 69:974–996. https://doi.org/10.1002/ird.2492

Dhaouadi L, Besser H, karbout, N., Wassar, F., Alomrane, A.R. (2021) Assessment of natural resources in tunisian Oases: Degradation of irrigation water quality and continued overexploitation of groundwater. EuroMediterr J Environ Integr. 6:1–13. https://doi.org/10.1007/s41207-020-00234-3

Dhaouadi L, Besser H, Karbout N, Khaldi R, Haj-Amor Z, Maachia S, Ouassar F (2022) Environmental sensitivity and risk assessment in the Saharan Tunisian oasis agro-systems using the deepest water table source for irrigation: water quality and land management impacts. Environ Dev Sustain. https://doi.org/10.1007/s10668-021-01878-z

Gaaloul N (2011) Water resources and management in Tunisia. Int J Water. 6:92–116. https://doi.org/10.1504/IJW.2011.043319

Georgescu PL, Moldovanu S, Iticescu C, Calmuc M, Calmuc V, Topa C, Moraru L (2023) Assessing and forecasting water quality in the Danube River by using neural network approaches. Sci Total Environ. 879:162998. https://doi.org/10.1016/j.scitotenv.2023.162998

Goodarzi MR, Abedi MJ, Niknam ARR, Heydaripour M (2022) Groundwater quality status based on a modification of water quality index in an arid area, Iran. Water Supply. 22:6245–6261. https://doi.org/10.2166/ws.2022.225

Guasmi I, Kraiem Z, Hallouche B (2022) Integrated hydrogeochemical and qualitative assessments of the High Melrir watershed (NW Algeria) for sustainable water resource management. Arab J Geosci. 15:1163. https://doi.org/10.1007/s12517-022-10390-1

Haj-Amor Z, Tóth T, Ibrahimi MK, Bouri S (2017) Effects of excessive irrigation of date palm on soil salinization, shallow groundwater properties, and water use in a Saharan oasis. Environ Earth Sci. 76:1–13. https://doi.org/10.1007/s12665-017-6935-8

Hameed M, Sharqi SS, Yaseen ZM, Afan HA, Hussain A, Elshafie A (2017) Application of artificial intelligence (AI) techniques in water quality index prediction: A case study in tropical region. Malays Neural Comput Appl. 28:893–905.

Hassen I, Hamzaoui-Azaza F, Bouhlila R (2016) Application of multivariate statistical analysis and hydrochemical and isotopic investigations for evaluation of groundwater quality and its suitability for drinking and agriculture purposes: Case of Oum Ali-Thelepte aquifer, central Tunisia. Environ Monit Assess. 188:1–20. https://doi.org/10.1007/s10661-016-5124-7

Hirvasniemi J, Runhaar J, van der Heijden RA, Zokaeinikoo M, Yang M, Li X, Tan J, Rajamohan HR, Zhou Y, Deniz CM, Caliva F, Iriondo C, Lee JJ, Liu F, Martinez AM, Namiri N, Pedoia V, Panfilov E, Bayramoglu N, Nguyen HH, Nieminen MT, Saarakkala S, Tiulpin A, Lin E, Li A, Li V, Dam EB, Chaudhari AS, Kijowski R, Bierma-Zeinstra S, Oei EHG, Klein S (2023) The KNee OsteoArthritis Prediction (KNOAP2020) challenge: An image analysis challenge to predict incident symptomatic radiographic knee osteoarthritis from MRI and X-ray images. Osteoarthr Cartil. 31:115–125. https://doi.org/10.1016/J.JOCA.2022.10.001

Isaac Sajan R, Christopher VB (2023) A fuzzy inference system for enhanced groundwater quality assessment and index determination. Water Quality Res J. 58:230–246. https://doi.org/10.2166/wqrj.2023.031

Isaac R, Siddiqui S (2022) Application of water quality index and multivariate statistical techniques for assessment of water quality around Yamuna River in Agra Region, Uttar Pradesh, India. Water Supply. 22:3399–3418. https://doi.org/10.2166/WS.2021.395

Islam Khan MS, Islam N, Uddin J, Islam S, Nasir MK (2022) Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. J King Saud Univ Comput Inf Sci. 34:4773–4781. https://doi.org/10.1016/J.JKSUCI.2021.06.003

Islam MS, Mostafa MG (2022) Development of an Integrated Irrigation Water Quality Index (IIWQIndex) model. Water Supply. 22:2322–2337. https://doi.org/10.2166/ws.2021.378

Kachroud M, Trolard F, Kefi M, Jebari S, Bourrié G (2019) Water quality indices: Challenges and application limits in the literature. Water (switzerland). 11:1–26. https://doi.org/10.3390/w11020361

Keskin TE, Düğenci M, Kaçaroğlu F (2015) Prediction of water pollution sources using artificial neural networks in the study areas of Sivas, Karabük and Bartın (Turkey). Environ Earth Sci. 73:5333–5347. https://doi.org/10.1007/s12665-014-3784-6

Khalfi C, Tarki M, Dassi L (2021a) An appraisal of Chott El Jerid brine encroachment in the Tozeur-south shallow aquifer: Geoelectrical and hydrochemical approach. J Appl Geophy. 190:104341. https://doi.org/10.1016/j.jappgeo.2021.104341

Khalfi C, Tarki M, Dassi L (2021b) An appraisal of Chott El Jerid brine encroachment in the Tozeur-south shallow aquifer: Geoelectrical and hydrochemical approach. J Appl Geophys. 190:104341. https://doi.org/10.1016/j.jappgeo.2021.104341

Khalil A, Almasri MN, McKee M, Kaluarachchi JJ (2005) Applicability of statistical learning algorithms in groundwater quality modeling. Water Resour Res. 41:1–16. https://doi.org/10.1029/2004WR003608

Khullar S, Singh N (2021) Machine learning techniques in river water quality modelling: A research travelogue. Water Sci Technol Water Supply. https://doi.org/10.2166/ws.2020.277

Kraiem Z, Chkir N, Zouari K, Parisot JC, Agoun A, Hermitte D (2012) Tomographic, hydrochemical and isotopic investigations of the salinization processes in the oasis shallow aquifers, Nefzaoua region, southwestern Tunisia. J Earth Syst Sci. 121:1185–1200. https://doi.org/10.1007/s12040-012-0221-7

Kraiem Z, Zouari K, Chkir N, Agoune A (2014) Geochemical characteristics of arid shallow aquifers in Chott Djerid, south-western Tunisia. J Hydro-Environ Res. 8:460–473. https://doi.org/10.1016/j.jher.2013.06.002

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

Kumar PJS, Augustine CM (2022) Entropy-weighted water quality index (EWQI) modeling of groundwater quality and spatial mapping in Uppar Odai Sub-Basin, South India. Model Earth Syst Environ. 8:911–924. https://doi.org/10.1007/s40808-021-01132-5

Kumari M, Rai SC (2020) Hydrogeochemical evaluation of groundwater quality for drinking and irrigation purposes using water quality index in semi arid region of India. J Geol Soc India. 95:159–168. https://doi.org/10.1007/s12594-020-1405-4

Lee JH, Lee JY, Cha YS, Cho SJ, Kim TH, Cha YK, Koo JY (2022) Development of customized water quality index for medium- and long-term evaluation of urban river water quality. Water Supply. 22:6338–6355. https://doi.org/10.2166/ws.2022.238

Lin GC, Wang WJ, Wang CM, Sun SY (2010) Automated classification of multi-spectral MR images using linear discriminant analysis. Comput Med Imaging Graph. 34:251–268. https://doi.org/10.1016/J.COMPMEDIMAG.2009.11.001

Liu F, Zhang J, Wang S, Zou J, Zhen P (2023) Multivariate statistical analysis of chemical and stable isotopic data as indicative of groundwater evolution with reduced exploitation. Geosci Front. 14:101476. https://doi.org/10.1016/J.GSF.2022.101476

Lu Y, Gan L, Chen Y, Zheng N (2023) Research on the allocation of water resources engineering projects based on multi-objective optimization. Water Supply. 23:3297–3308. https://doi.org/10.2166/ws.2023.153

Mashaly AF, Alazba A (2019) Assessing the accuracy of ANN, ANFIS, and MR techniques in forecasting productivity of an inclined passive solar still in a hot, arid environment. Water SA. 45:239–250. https://doi.org/10.4314/wsa.v45i2.11

McHugh ML (2012) Interrater reliability: the kappa statistic. Biochem Med (zagreb). 22:276–282

Meegahakumbura SPS, Nanayakkara N (2023) Assessment of groundwater quality in Vavuniya and Mullaitivu, Sri Lanka using multivariate statistical techniques and a Water Quality Index. Water Supply. 23:867–883. https://doi.org/10.2166/ws.2023.028

Msaddek MH, Moumni Y, Ayari A, El May M, Chenini I (2022) Artificial intelligence modelling framework for mapping groundwater vulnerability of fractured aquifer. Geocarto Int. 37:10480–10510. https://doi.org/10.1080/10106049.2022.2037729

Najah Ahmed A, Binti Othman F, Abdulmohsin Afan H, Khaleel Ibrahim R, Ming Fai C, Shabbir Hossain M, Ehteram M, Elshafie A (2019) Machine learning methods for better water quality prediction. J Hydrol (amst). https://doi.org/10.1016/j.jhydrol.2019.124084

Oğuz A, Ertuğrul ÖF (2023) A survey on applications of machine learning algorithms in water quality assessment and water supply and management. Water Supply. 23:895–922. https://doi.org/10.2166/ws.2023.033

Parkavi R, Karthikeyan P, Sheik Abdullah A (2023) Predicting academic performance of learners with the three domains of learning data using neuro-fuzzy model and machine learning algorithms. J Eng Res. https://doi.org/10.1016/j.jer.2023.09.006

Parween S, Siddique NA, Mahammad Diganta MT, Olbert AI, Uddin MG (2022) Assessment of urban river water quality using modified NSF water quality index model at Siliguri city, West Bengal, India. Environ Sustain Indic. 16:100202. https://doi.org/10.1016/j.indic.2022.100202

Rashid A, Kumari S (2023) Performance evaluation of ANN and ANFIS models for estimating velocity and pressure in water distribution networks. Water Supply. https://doi.org/10.2166/WS.2023.224

Sahu S, Gogoi U, Nayak NC (2021) Groundwater solute chemistry, hydrogeochemical processes and fluoride contamination in phreatic aquifer of Odisha, India. Geosci Front. 12:101093. https://doi.org/10.1016/j.gsf.2020.10.001

Sajib AM, Diganta MTM, Rahman A, Dabrowski T, Olbert AI, Uddin MG (2023) Developing a novel tool for assessing the groundwater incorporating water quality index and machine learning approach. Groundw Sustain Dev. 23:101049. https://doi.org/10.1016/j.gsd.2023.101049

Sangare LO, Ba S, Toure A, Samake M, Zheng T (2023) Assessment of the water quality of the Niger River in Bamako, Mali, based on the Water Quality Index. Water Supply. 23:671–687. https://doi.org/10.2166/ws.2023.029

Satish Kumar V, Amarender B, Dhakate R, Sankaran S, Raj Kumar K (2016) Assessment of groundwater quality for drinking and irrigation use in shallow hard rock aquifer of Pudunagaram, Palakkad District Kerala. Appl Water Sci. 6:149–167. https://doi.org/10.1007/S13201-014-0214-6

Schreiber SG, Schreiber S, Tanna RN, Roberts DR, Arciszewski TJ (2022) Statistical tools for water quality assessment and monitoring in river ecosystems—A scoping review and recommendations for data analysis. Water Quality Res J. 57:40–57. https://doi.org/10.2166/wqrj.2022.028

Smida H, Tarki M, Gammoudi N, Dassi L (2023) GIS-based multicriteria and artificial neural network (ANN) investigation for the assessment of groundwater vulnerability and pollution hazard in the Braga shallow aquifer (Central Tunisia): A critical review of generic and modified DRASTIC models. J Contam Hydrol. 104:245. https://doi.org/10.1016/J.JCONHYD.2023.104245

Taheri S, Mammadov M (2013) Learning the naive bayes classifier with optimization models. Int J Appl Math Comput Sci. 23:787–795. https://doi.org/10.2478/amcs-2013-0059

Tang Y, Zang C, Wei Y, Jiang M (2019) Data-driven modeling of groundwater level with least-square support vector machine and spatial-temporal analysis. Geotech Geol Eng. 37:1661–1670. https://doi.org/10.1007/s10706-018-0713-6

Tao H, Hameed MM, Marhoon HA, Zounemat-Kermani M, Salim H, Sungwon K, Sulaiman SO, Tan ML (2022) Groundwater level prediction using machine learning models: A comprehensive review. Neurocomputing. 25:879.

Tarki M, Dassi L, Jedoui Y (2012) Groundwater composition and recharge origin in the shallow aquifer of the Djerid oases, southern Tunisia: Implications of return flow. Hydrol Sci J. 57:790–804. https://doi.org/10.1080/02626667.2012.681783

Thirumalai Raja K, Jayanthi N, Leta Tesfaye J, Nagaprasad N, Krishnaraj R, Kaushik VS (2022) Using an artificial neural network to validate and predict the physical properties of self-compacting concrete. Adv Mater Sci Eng. https://doi.org/10.1155/2022/1206512

Uddin MG, Nash S, Olbert AI (2021) A review of water quality index models and their use for assessing surface water quality. Ecol Indic. 122:107218. https://doi.org/10.1016/j.ecolind.2020.107218

Uddin MG, Nash S, Rahman A, Olbert AI (2022) A comprehensive method for improvement of water quality index (WQI) models for coastal water quality assessment. Water Res. 219:118532. https://doi.org/10.1016/j.watres.2022.118532

Uddin MG, Diganta MTM, Sajib AM, Hasan MA, Moniruzzaman M, Rahman A, Olbert AI, Moniruzzaman M (2023a) Assessment of hydrogeochemistry in groundwater using water quality index model and indices approaches. Heliyon. 9:e19668. https://doi.org/10.1016/j.heliyon.2023.e19668

Uddin MG, Jackson A, Nash S, Rahman A, Olbert AI (2023b) Comparison between the WFD approaches and newly developed water quality model for monitoring transitional and coastal water quality in Northern Ireland. Sci Total Environ. 901:165960. https://doi.org/10.1016/j.scitotenv.2023.165960

Uddin MG, Rahman A, Nash S, Diganta MTM, Sajib AM, Moniruzzaman M, Olbert AI (2023c) Marine waters assessment using improved water quality model incorporating machine learning approaches. J Environ Manag. 344:118368. https://doi.org/10.1016/j.jenvman.2023.118368

Uddin MG, Nash S, Rahman A, Dabrowski T, Olbert AI (2024) Data-driven modelling for assessing trophic status in marine ecosystems using machine learning approaches. Environ Res. 242:117755. https://doi.org/10.1016/j.envres.2023.117755

Verma SK, Chaurasia S (2023) A hydrogeochemical study of groundwater of district Ballia, U.P., India. Water Supply. 23:3221–3238. https://doi.org/10.2166/ws.2023.166

Wan H, Wang H, Scotney BW, Liu J, Wei X (2023) Global subclass discriminant analysis. Knowl Based Syst. 280:111010. https://doi.org/10.1016/J.KNOSYS.2023.111010

WHO (2017) Guidelines for drinking-water quality: First addendum to the fourth edition. World Health Organization, Geneva.

Xiao J, Jin Z, Wang J, Zhang F (2015) Major ion chemistry, weathering process and water quality of natural waters in the Bosten Lake catchment in an extreme arid region, NW China. Environ Earth Sci. 73:3697–3708. https://doi.org/10.1007/S12665-014-3657

Zaqoot HA, Hamada M, Miqdad S (2018) A comparative study of ann for predicting nitrate concentration in groundwater wells in the Southern Area of Gaza Strip. Appl Artif Intell. 32:727–744. https://doi.org/10.1080/08839514.2018.1506970

Zhang L, Jiang L, Li C, Kong G (2016) Two feature weighting approaches for naive Bayes text classifiers. Knowl Based Syst. 100:137–144. https://doi.org/10.1016/j.knosys.2016.02.017

Zhen J, Nan J, Hou J (2023) Quantitative analysis of regional water resources and secure development of economic and social coupling systems. Water Supply. 23:3192–3204. https://doi.org/10.2166/ws.2023.198

Zhu F, Gao J, Yang J, Ye N (2022a) Neighborhood linear discriminant analysis. Pattern Recognit. 123:108422. https://doi.org/10.1016/j.patcog.2021.108422

Zhu M, Wang J, Yang X, Zhang Y, Zhang L, Ren H, Wu B, Ye L (2022b) A review of the application of machine learning in water quality evaluation. Eco-Environ Health. 1:107–116. https://doi.org/10.1016/j.eehl.2022.06.001