

Fethi NADOOR¹, Abdallah CHABBI², Mohamed NADOOR³, Noureddine ZENATI⁴
Nabiha BELAHCENE⁵, Amal H. ALHARBI⁶, El-Sayed M. EL-KENAWY^{7,8}
and Abdelaziz RABEHI^{9*}

INTEGRATING CCME-WQI AND MACHINE LEARNING FOR GROUNDWATER QUALITY IN THE SEDRATA PLAIN

Abstract: The quality of groundwater in the Sedrata plain in northeastern Algeria is under increasing pressure due to semi-arid climatic conditions, geological mineralisation and growing agricultural demand. This study provides a thorough physicochemical characterisation of phreatic aquifer waters collected from 25 open wells over four sampling periods (February, May and December 2023 and May 2024; $n = 100$), combining a traditional Water Quality Index (WQI) analysis with six machine learning (ML) prediction models. The CCME-WQI framework was applied using 12 physicochemical parameters. Six ML algorithms, Gradient Boosting (GB), Random Forest (RF), Support Vector Regression (SVR), Bayesian Regression (BR), AdaBoost (AdB) and K-Nearest Neighbours (K-NN) were trained and evaluated across 12 nested input-variable combinations, using the Recursive Feature Elimination with Linear Cross-Validation (RFE-L) methodology. Gradient Boosting achieved the highest predictive accuracy on the full variable set ($R^2 = 0.88$ and RMSE = 4.89 in the test set), while SVR demonstrated the best generalisation ratio ($R^2 = 0.83$ in the test set with only two variables removed). Best-subset regression identified a parsimonious nine-variable model (T , EC , $Turb$, HCO_3^- , Ca^{2+} , Cl^- , K^+ , Na^+ ; Adj. $R^2 = 0.812$) as the recommended input for operational WQI prediction. Chloride, calcium and electrical conductivity exhibited the strongest negative correlations with WQI, reflecting evaporite dissolution and ion exchange processes. These findings provide actionable guidance for the management of groundwater in semi-arid regions facing analogous hydrochemical and climatic constraints.

Keywords: groundwater quality, hydrochemical characterisation, CCME-WQI, machine learning, semi-arid aquifer

¹ Laboratory of Modeling and Socio-Economic Analysis in Water Science - MASESE - Mohamed Cherif Messaadia University, Souk Ahras, 41000, Algeria, email: F.nadour@Univ-SoukAhras.dz, ORCID: 0009-0008-2380-2512

² Laboratory of Science and Techniques for Living, Souk Ahras University, Algeria, email: A.chabbi@Univ-SoukAhras.dz, ORCID: 0000-0003-1198-022X

³ Applied Automation and Industrial Diagnostics Laboratory, Faculty of Science and Technology, University of Djelfa 17000 DZ, Algeria, email: m.nadour@univ-djelfa.dz, ORCID: 0000-0002-6607-1261

⁴ Laboratory Science and Technical of Water and Environment, Souk Ahras University, 41000 Algeria, email: N.zenati@univ-soukahras.dz, ORCID: 0000-0002-1096-559X

⁵ Laboratory of Life Sciences and Techniques, University of Mohamed Cherif Messaadia, Souk-Ahras, 21000 Algeria, email: N.belahcene@Univ-SoukAhras.dz, ORCID: 0000-0002-2647-2968

⁶ Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh 11671, Saudi Arabia, email: ahalharbi@pnu.edu.sa, ORCID: 0000-0002-7332-4941

⁷ Delta Higher Institute of Engineering and Technology, Department for Communications and Electronics, Mansoura 35511, Egypt, email: skenawy@ieee.org, ORCID: 0009-0006-4318-5481

⁸ Jadara Research Center, Jadara University, Irbid 21110, Jordan

⁹ Laboratory of Telecommunication and Smart Systems (LTSS), Faculty of Science and Technology University of Djelfa, PO Box 3117, Djelfa 17000, Algeria, email: abdelaziz.rabehi@univ-djelfa.dz, ORCID: 0000-0001-8684-4754

* Corresponding author: abdelaziz.rabehi@univ-djelfa.dz

Introduction

Access to safe groundwater is fundamental to human health, food security and sustainable development, particularly in semi-arid regions where surface water resources are scarce, highly seasonal or unreliable [1]. Northeastern Algeria, including the wilaya of Souk Ahras, is an example of this: annual rainfall rarely exceeds 600 mm, and long periods of drought are becoming more common in the current climate. Phreatic aquifer systems must meet the demands of domestic use, livestock and irrigation simultaneously [2, 3]. In this context, groundwater quality is shaped by the interplay of natural processes such as mineralisation from carbonate and evaporite lithologies, ion exchange on clay minerals and evapotranspirative concentration, as well as anthropogenic pressures including irrigated agriculture, limited wastewater treatment infrastructure and direct aquifer recharge from unlined irrigation channels. Elevated chloride, sulphate and calcium concentrations have been well documented in shallow aquifers formed in Miocene-Pliocene detrital deposits overlying Triassic evaporite sequences [4, 5]. However, systematic, spatially resolved monitoring data collected over multiple hydrological seasons are still lacking for the Sedrata sub-basin. The Water Quality Index (WQI) is an internationally recognised tool for condensing multi-parameter water quality data into a single, policy-relevant metric. This study uses the Canadian Council of Ministers of the Environment (CCME) WQI [6], an open index whose scope-frequency-amplitude structure makes it particularly well-suited to monitoring programmes in which the list of parameters and the criteria for exceeding them may evolve over time [1]. The CCME-WQI has been applied in various hydrogeological settings and has demonstrated its effectiveness in identifying spatial and temporal variations in the quality of both surface water and groundwater [7]. Although the conventional computation of the WQI is feasible for small datasets, the management of water resources increasingly requires the rapid and cost-effective prediction of water quality at locations where physical sampling is infrequent or impractical. In this context, machine learning (ML) models have emerged as powerful tools: ensemble methods such as Random Forest (RF) and Gradient Boosting (GB) can capture complex, nonlinear relationships among physicochemical predictors, while wrapper-based feature selection algorithms can identify parsimonious input sets that minimise monitoring costs without sacrificing predictive accuracy [8, 9]. Employing AdaBoost, K-NN, GB, RF, SVR and Bayesian regression against 12 water-quality predictors with recursive feature elimination provides the methodological template adopted in the present study.

The objectives of this investigation are threefold: (1) characterising the seasonal and spatial variability of physicochemical parameters in the Sedrata phreatic aquifer across four sampling campaigns, (2) computing CCME-WQI values and mapping the hydrochemical controls on groundwater quality, and (3) evaluating and comparing six ML algorithms for WQI prediction using nested input-variable combinations, and identifying the most parsimonious yet accurate model for operational deployment.

Although the integration of CCME-WQI and machine learning has been reported previously, many published studies in comparable hydrogeological settings have focused on a single predictive model, a single train/test split, or a single sampling campaign, leaving three aspects insufficiently resolved for semi-arid Maghrebien aquifers: which minimal subset of physicochemical parameters can be used for reduced-cost monitoring without materially sacrificing predictive accuracy; how reliably the reported predictive performance of tree-based ensembles generalises beyond a single data split; and how strongly seasonal

recharge and irrigation-return-flow dynamics modulate the predictors of water quality. To address these gaps, the present study makes four complementary contributions. First, a systematic, nested-input-variable comparison of six ML algorithms combined with Recursive Feature Elimination with Linear Cross-Validation (RFE-L) across 12 candidate variable combinations identifies the smallest subset of physicochemical predictors while maintaining comparable predictive performance. Second, an exhaustive best-subset multiple regression analysis, validated against AIC, BIC and Mallows' C_p together with a formal VIF-based multicollinearity check, translates the feature-selection results into concrete, quantified guidance for reduced-cost operational monitoring. Third, an explicit quantification, via k-fold cross-validation and hyperparameter tuning described in the hyperparameter optimisation section, of the generalisation gap of tree-based ensembles provides quantitative evidence that single train/test splits may overestimate predictive performance, and identifies a regularised configuration suited to operational deployment. Fourth, a four-season sampling design capturing both winter recharge and summer irrigation-return-flow dynamics enabled the machine learning models to capture seasonal variability in groundwater quality, thereby improving the robustness and environmental relevance of the predictive framework under semi-arid hydrogeological conditions.

Materials and methods

Study area

Sedrata is a commune located 56 km southwest of Souk Ahras city in the northeastern portion of Algeria (36° 7' 42" N, 7° 31' 53" E; elevation 804 m - 811 m), approximately 85 km from the Mediterranean coastline (Fig. 1). With a resident population of approximately 80,197 inhabitants, the commune supports diverse agricultural activities, principally cereal cultivation, olive orchards, and market gardening, irrigated predominantly from phreatic wells [10].

The climate is semi-arid, with mean annual precipitation ranging from 267.1 mm to 550.8 mm, high inter-annual variability, and a pronounced summer dry season. Temperatures are continental, with hot, dry summers ($> 35^{\circ}\text{C}$) and cold winters; snowfall is occasional above 800 m. The combination of high potential evapotranspiration and low, episodic recharge concentrates solutes in shallow groundwater, particularly during drought years.

Geological and hydrogeological setting

The Sedrata region is located at the tectonic boundary between two significant structural domains [4, 5, 11-15]. The autochthonous and para-autochthonous domain comprises the Sellaoua and North Aures units and consists of Lower to Upper Cretaceous (Barremian-Maastrichtian) carbonate sequences interbedded with marls of varying carbonate content. These sequences are locally disrupted by Triassic diapirs composed of argillaceous-, gypsiferous-, calcareous-, and dolomitic facies, principally gypsum and halite. These basement units are overthrust from the north by the allochthonous Tellian nappe (comprising marls and bioclastic limestones from the Palaeocene to the Lutetian) and the Numidian nappe (consisting of Oligocene to Miocene quartzose sandstones and green clays). The Sedrata plain, the main sampling area, is underlain by Mio-Pliocene conglomerates and red clays deposited after tectonic activity, a Villafranchian calcareous crust and Quaternary fluvial and colluvial deposits. This detrital sequence constitutes the

regional phreatic aquifer, characterised by an average transmissivity of around $3 \cdot 10^{-4}$ m²/s, a hydraulic conductivity of around $1.2 \cdot 10^{-5}$ m/s, and a storage coefficient of around $2.3 \cdot 10^{-4}$ [16]. Well depths across the sampling network range from 20 m to 50 m. The proximity of Triassic evaporite outcrops, particularly gypsum and halite primarily control elevated SO_4^{2-} and Cl^- concentrations in phreatic water throughout the plain.

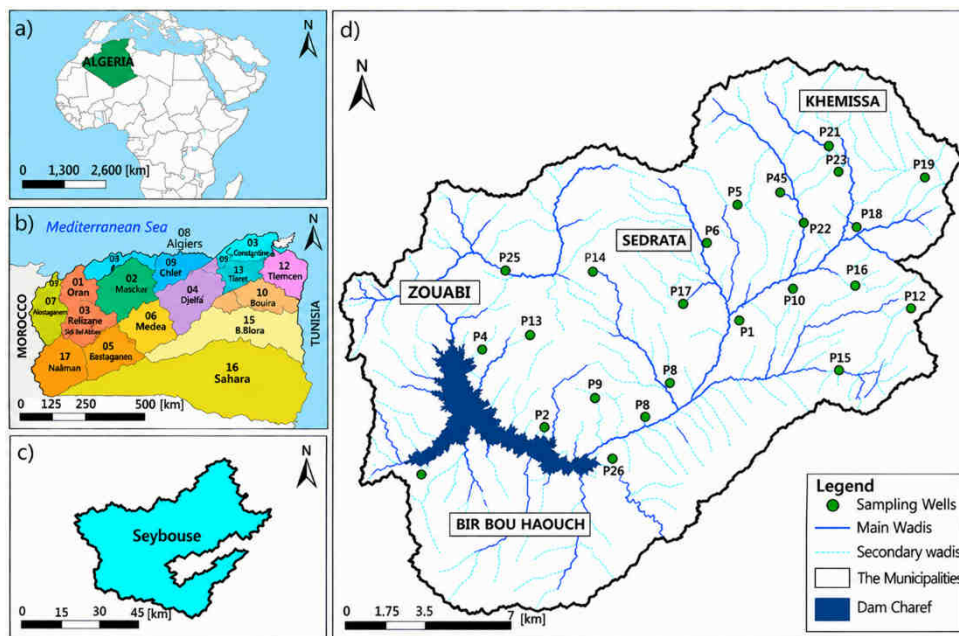


Fig. 1. Geographical location of the study area and the groundwater samples: a) Africa, b) Greater Algiers Watershed, c) Seybouse Watershed, d) study area

Data collection and sampling

Groundwater samples were collected during four campaigns representing contrasting hydrological conditions: February 2023 (winter recharge season), May 2023 (spring, pre-irrigation peak), December 2023 (post-autumn recharge), and May 2024 (spring repeat). Samples were drawn from 25 open phreatic wells (3 replicates per well per campaign would have yielded 300 samples; the study reports $N = 100$ composite samples, corresponding to 25 wells $\cdot$ 4 campaigns), used exclusively for irrigation.

Samples were collected in 1.5-litre high-density polyethylene bottles, labelled with well ID, sampling date, and aquifer depth, and stored at 4 °C during transport to the Algerian Water Laboratory (ADE), following APHA [17] protocols. Field parameters, electrical conductivity (EC), pH, temperature, and turbidity, were measured in situ using calibrated portable instruments (HANNA HI 8915 ATC pH meter; HANNA 54500-18 conductivity meter; HACH 2100N turbidity meter). Dissolved oxygen (O_2) was measured in situ using a pre-calibrated DO meter with bubble-free immersion. Major cations (Ca^{2+} , Mg^{2+} , K^+ , Na^+) and anions (Cl^- , SO_4^{2-}) were quantified by ICP-OES; HCO_3^- was determined by potentiometric titration following standard analytical methods [17, 18].

The ionic balance method (IB) was applied to verify the accuracy of the chemical analysis for each water sample.

All samples remained within the acceptable limit of $\pm 10\%$ tolerance, confirming analytical reliability [18-21].

Water quality index (WQI) calculation

The Water quality index (WQI) summarises extensive water quality data in the form of a single numerical score [22]. This study employed the CCME WQI, which was developed by the Canadian Council of Ministers of the Environment (CCME) in 2001. The CCME WQI is an open index, meaning it does not specify which parameters or standards must be used. Instead, users select the variables and criteria that match the water body and monitoring objectives. While this flexibility allows the index to be adapted for different monitoring programmes, consistent use of the same parameters and criteria is required to make valid comparisons over time or between sites. The CCME-WQI is calculated from three factors: scope (F_1), frequency (F_2) and amplitude (F_3). Scope (Eq. (1)) is the percentage of measured variables that exceed guideline values at least once; frequency (Eq. (2)) is the percentage of all measurements that violate the guidelines; and amplitude (Eq. (3)) quantifies the magnitude by which violating measurements fall outside the guideline values. The factors are defined as

$$F_1 = \frac{V_f}{V} \cdot 100 \quad (1)$$

$$F_2 = \frac{N_f}{N} \quad (2)$$

$$F_3 = \frac{NSE}{0.01NSE + 0.01} \quad (3)$$

where V is the total number of variables, V_f is the number of failed variables, N is the total number of tests, N_f is the number of failed tests, and NSE is the normalised sum of excursions. The excursion E_i for each failed test i is calculated as

$$E_i = \frac{x_i}{O_i} - 1 \quad \text{for } x_i > O_i \quad (4)$$

$$E_i = \frac{O_i}{x_i} - 1 \quad \text{for } O_i > x_i \quad (5)$$

where x_i is the value of the i -th failed test and O_i is the corresponding objective or guideline value. The normalised sum of excursions is then

$$NSE = \frac{1}{N} \sum_{i=1}^{N_f} E_i \quad (6)$$

Finally, the CCME-WQI is computed as [23]

$$\text{CCME-WQI} = 100 - \frac{\sqrt{F_1^2 + F_2^2 + F_3^2}}{1.732} \quad (7)$$

The resulting CCME-WQI values were classified into five categories as shown in (Table 1), WQI was computed using CCME Calculator MATLAB v2022.

Table 1

Classification of CCME-WQI and corresponding water quality status [23]

Water quality class	Index range	Water status
Excellent	95-100	Very good quality water
Good	80-94	Good quality water
Fair	65-79	Acceptable quality water
Marginal	45-64	Poor quality water
Poor	0-44	Very poor quality water

Machine learning algorithms and model training

Six machine-learning algorithms were evaluated for their ability to predict the CCME Water Quality Index using a 12-parameter physicochemical dataset. Gradient Boosting (GB) involves building successive decision trees that correct predecessor residuals iteratively, thereby enabling the modelling of complex nonlinear interactions [23, 24]. Random Forest (RF) uses bagging to aggregate uncorrelated trees, reducing overfitting and providing inherent feature importance measures [25, 26]. Bayesian regression (BR) incorporates prior information within a probabilistic framework to produce calibrated uncertainty estimates, which are valuable when data are limited [27]. AdaBoost (AdB) sequentially reweights training instances to focus learning on hard-to-predict observations [28]. K-Nearest Neighbours (K-NN) predicts by averaging the K closest samples in feature space and performs well when the local data structure is representative [29]. Support Vector Regression (SVR) fits an optimal regression hyperplane in a kernel-induced space, providing robustness to outliers and flexible modelling of non-linearity [30, 31]. All 100 samples were used for model development and were partitioned into training (80 samples, 80 %) and testing (20 samples, 20 %) sets using stratified random splitting, preserving seasonal representation across the subsets in line with hydrogeological ML practice [32]. Input variables were min-max normalised to the [0.1, 0.9] range to mitigate scale-induced bias among parameters with disparate magnitudes. Feature selection employed recursive feature elimination with linear cross-validation (RFE-L), a backward elimination approach in which an initial model is trained on all 12 variables, the features are ranked by importance, the least informative variable is removed and the model is refitted iteratively until the removal of a further variable causes the R^2 or RMSE to exceed the predetermined threshold.

Table 2

Nested input-variable combinations tested across all six ML algorithms

Model	Variables
1	pH, T, O ₂ , EC, Turb, HCO ₃ ⁻ , SO ₄ ²⁻ , Ca ²⁺ , Mg ²⁺ , Cl ⁻ , K ⁺ , Na ⁺
2	pH, T, O ₂ , EC, Turb, HCO ₃ ⁻ , SO ₄ ²⁻ , Ca ²⁺ , Mg ²⁺ , Cl ⁻ , K ⁺
3	pH, T, O ₂ , EC, Turb, HCO ₃ ⁻ , SO ₄ ²⁻ , Ca ²⁺ , Mg ²⁺ , Cl ⁻
4	pH, T, O ₂ , EC, Turb, HCO ₃ ⁻ , SO ₄ ²⁻ , Ca ²⁺ , Mg ²⁺
5	pH, T, O ₂ , EC, Turb, HCO ₃ ⁻ , SO ₄ ²⁻ , Ca ²⁺
6	pH, T, O ₂ , EC, Turb, HCO ₃ ⁻ , SO ₄ ²⁻
7	pH, T, O ₂ , EC, Turb, HCO ₃ ⁻
8	pH, T, O ₂ , EC, Turb
9	pH, T, O ₂ , EC
10	pH, T, O ₂
11	pH, T
12	pH

Evaluation metrics

Several statistical measures were applied to assess model performance: Root Mean Squared Error (*RMSE*), Mean Absolute Error (*MAE*), and the coefficient of determination (R^2) [33]. With optimal configurations presented in (Table 3).

Statistical metrics used for machine learning model evaluation

Table 3

$RMSE = \sqrt{\frac{1}{n} \sum (O_i - P_i)^2}$	The Root Mean Squared Error (<i>RMSE</i>) quantifies the average magnitude of prediction errors, giving higher weight to larger deviations Where O_i and P_i represent the observed and predicted WQI values, respectively, and n is the total number of samples.
$MAE = \frac{1}{n} \sum O_i - P_i $	Mean Absolute Error (<i>MAE</i>) measures the average absolute differences between predicted and actual values.
$R^2 = 1 - \frac{\sum (O_i - P_i)^2}{\sum (O_i - \bar{O}_i)^2}$	The coefficient of determination (R^2) measures the extent of IWQI variability captured by the independent variables Where $\bar{O}_i$ is the mean of the observed values, R^2 values closer to 1 indicate a stronger model fit, With a value of 1 representing perfect prediction

Feature importance analysis

Machine learning models assess input variables by assigning importance scores or weights that quantify each feature's influence on predictions [34]. Different algorithmic architectures and optimisation procedures produce distinct importance allocations: some models emphasise particular predictors, while others highlight alternative variables. Comparative analysis reveals these divergences and can guide model selection [35].

Selecting relevant and informative features directly affects model stability and robustness, improving predictive consistency across conditions [35–38].

As importance measures are model-dependent, comparing the normalised scores across models helps to identify predictors that are universally informative versus those that are algorithm-specific.

Best-subset regression analysis

To identify a statistically optimal, parsimonious predictor set for operational WQI monitoring, an exhaustive best subset multiple regression analysis was conducted across all 12 nested variable subsets. Six model-selection criteria were computed: Mean Squared Error (*MSE*), R^2 , adjusted R^2 , Akaike Information Criterion (*AIC*; [39], Bayesian Information Criterion (*BIC*; [40]), and Mallows' C_p [41]. The recommended model was selected as that minimising *AIC* and Mallows' C_p simultaneously while achieving the maximum adjusted R^2 , following established variable-selection practice [42, 43].

Hyperparameter optimisation using grid search and 5-fold cross-validation

To assess the sensitivity of model performance to hyperparameter choice and to identify robust model configurations, a 5-fold cross-validated grid search (scoring = R^2) was performed for the three best-performing algorithms (Gradient Boosting, Random Forest and Support Vector Regression) on the full 12-variable input set. Grid-search optimisation was implemented using GridSearchCV with five-fold cross-validation ($cv = 5$) and R^2 as the optimisation metric. The optimal hyperparameter combination was selected according to the highest mean cross-validated R^2 obtained from the five validation folds.

The parameter grids explored and the resulting optimal configurations are reported in Table 4.

Hyperparameter optimisation was performed only for the three best-performing algorithms (GB, RF and SVR), whereas the remaining models were retained with their default configurations to preserve a consistent baseline comparison, as their predictive performance was substantially lower and further optimisation was unlikely to alter the overall comparative conclusions.

Table 4

Grid-search hyperparameter ranges and optimal configurations (5-fold CV, scoring = R^2)

Algorithm	Hyperparameter grid searched	Optimal configuration	Best CV R^2
Gradient boosting	n_estimators $\in \{100, 200, 300\}$; max_depth $\in \{2, 3, 4\}$; learning_rate $\in \{0.01, 0.05, 0.1\}$; min_samples_leaf $\in \{2, 5, 10\}$	n_estimators = 300, max_depth = 2, learning_rate = 0.1, min_samples_leaf = 5	0.847
Random forest	n_estimators $\in \{200, 400, 600\}$; max_depth $\in \{3, 5, 8, \text{None}\}$; min_samples_leaf $\in \{1, 2, 4\}$	n_estimators = 200, max_depth = 8, min_samples_leaf = 1	0.689
Support vector regression	C $\in \{1, 10, 50, 100\}$; epsilon $\in \{0.5, 1, 2\}$; kernel = RBF; gamma $\in \{\text{scale}, \text{auto}\}$	C = 100, epsilon = 2, gamma = 'auto'	0.749

All input variables were min-max normalised to [0.1, 0.9] prior to grid search, consistent with the machine learning algorithms and model training section. The shallow optimal tree depth selected for Gradient Boosting (max_depth = 2) and the associated minimum leaf size (5) act as explicit regularisers and are discussed further in relation to the overfitting behaviour identified in the model comparison discussion.

The tuned configurations reported in Table 4 were subsequently adopted in the independent 5-fold cross-validation analysis to evaluate model robustness and generalisation performance.

Results

Physicochemical characterisation of the Sedrata phreatic aquifer

The descriptive statistics for the 12 measured parameters (Table 5) suggest that there is substantial variability across the dataset. Several of the variables exhibit a clear positive skew, which is consistent with point source or episodic contamination. The range of groundwater pH was from 6.43 to 9.22 (mean 7.34 ± 0.38), reflecting bicarbonate buffering in the calcareous aquifer. Temperature varied from 11.0 °C to 19.8 °C (mean 15.6 °C), which is consistent with seasonal recharge dynamics. Dissolved oxygen showed wide dispersion (0.39–9.32 mg/L; mean 4.93 mg/L), with low values potentially linked to stagnant wells or the degradation of organic matter. Electrical conductivity was the most variable parameter in terms of diagnosis ((400–5,160) $\mu\text{S}\cdot\text{cm}^{-1}$; mean 1,570 $\mu\text{S}\cdot\text{cm}^{-1}$; skewness 1.49; kurtosis 3.31), indicating a subset of highly mineralised samples from evaporite-influenced flow paths or evaporative concentration. Turbidity was similarly skewed ((0.13–39.3) NTU (nephelometric turbidity unit); mean 4.72 NTU; skewness 3.25), with extreme values tied to storm recharge or sampling disturbance. The major ion composition showed a dominance of HCO_3^- (mean 310 mg/L) and Ca^{2+} (mean 142 mg/L),

as well as elevated Cl^- ((142-710) mg/L; mean 351 mg/L), which was often above guideline levels. There was also a high degree of skewness in K^+ (skewness 4.49), which suggests localised agricultural inputs.

Table 5
Descriptive statistics (mean, standard deviation (SD), skewness, kurtosis) of groundwater quality parameters
[$N = 100$]

Parameter	Min.	Max.	Mean	SD	Skewness	Kurtosis	WHO Standard 2017	Alg Standards 2011
pH [-]	6.43	9.22	7.34	0.38	1.36	5.28	6.5-9	6.5-9
T [°C]	11.0	19.8	15.6	1.9	-0.16	-0.61	25	25
O_2 [mg/L]	0.4	9.3	4.9	2.7	-0.10	-1.26	/	/
EC [$\mu\text{S}/\text{cm}$]	400	5160	1570	820	1.49	3.31	1500	2800
Turb [NTU]	0.1	39.3	4.7	5.8	3.25	14.0	≤ 5	≤ 5
HCO_3^- [mg/L]	91	945	310	124	2.05	7.36	300	/
SO_4^{2-} [mg/L]	22	691	117	110	2.99	10.8	250	400
Ca^{2+} [mg/L]	27	680	142	100	2.49	8.63	120	200
Mg^{2+} [mg/L]	0.24	156	28	33	1.79	2.98	50	150
Cl^- [mg/L]	142	710	351	135	0.82	0.32	200	500
K^+ [mg/L]	0.13	15	14	29	4.49	19.38	12	12
Na^+ [mg/L]	37	414	167	52	1.28	4.86	150	200

Pearson correlation analysis

The Pearson correlation matrix (Fig. 2) reveals the hydrochemical associations that structure the Sedrata groundwater system, as well as their relationship with the CCME-WQI. The strongest positive inter-parameter correlations among the physicochemical parameters are observed between Ca^{2+} and SO_4^{2-} ($r = 0.79$), Ca^{2+} and HCO_3^- ($r = 0.62$), and SO_4^{2-} and HCO_3^- ($r = 0.59$). This is consistent with the dissolution of carbonate and evaporite minerals (e.g. calcite, dolomite and gypsum) occurring simultaneously in the Triassic and Cretaceous sequences. Electrical conductivity correlates positively with Ca^{2+} ($r = 0.60$), SO_4^{2-} ($r = 0.56$), HCO_3^- ($r = 0.44$) and Cl^- ($r = 0.45$), which confirms that total dissolved solids are controlled by the same mineralogical sources. The Cl^- versus Na^+ covariation ($r = 0.40$) suggests halite dissolution as an additional source of salinity, while the moderate $\text{Mg}-\text{Cl}$ ($r = 0.35$) and $\text{Mg}-\text{Na}$ ($r = 0.22$) correlations point to dolomite and exchange processes. Regarding WQI, the strongest negative correlations are with Cl^- ($r = -0.57$), Ca^{2+} ($r = -0.55$), EC ($r = -0.52$), K^+ ($r = -0.47$), SO_4^{2-} ($r = -0.44$), turbidity ($r = -0.35$) and HCO_3^- ($r = -0.33$). This confirms that these are the main parameters causing index exceedances and degrading groundwater quality. The weakest WQI correlations are for Na^+ ($r = -0.08$), O_2 ($r = -0.07$) and pH ($r = +0.17$), indicating that they contribute less to variation in quality across the sampled wells.

CCME-WQI assessment

The CCME-WQI scores derived from the analyses of 100 samples indicate that the quality of the groundwater across the Sedrata phreatic aquifer (Table 6) is predominantly poor to marginal. Of the 25 wells sampled, none achieved an 'Excellent' classification ($\text{WQI} > 95$), and only three wells (12 %) attained 'Good' status ($\text{WQI} 80-94$). These three wells are all located in the less-irrigated, upstream northwestern sector of the plain. Seven wells (28 %) were classified as fair ($\text{WQI} 65-79$), while the majority (44 %) fell into the

marginal category (WQI 45-64). Four wells (16 %), clustered in the southeastern sector adjacent to the deepest Triassic outcrops, were classified as poor (WQI < 44). The parameters that most frequently exceeded WHO objectives were SO_4^{2-} , Cl^- , EC , Ca^{2+} and K^+ , which confirms the primary role of evaporite dissolution and anthropogenic salinisation in quality degradation. A seasonal pattern was also evident: WQI scores were generally lower (indicating poorer quality) during the May sampling campaigns, when irrigation demand peaks and return flows contribute concentrated solutes to the shallow water table. This is in contrast to the February and December campaigns, when the aquifer receives diluting recharge from winter rainfall. This temporal variability highlights the importance of multi-season monitoring for a representative quality assessment.

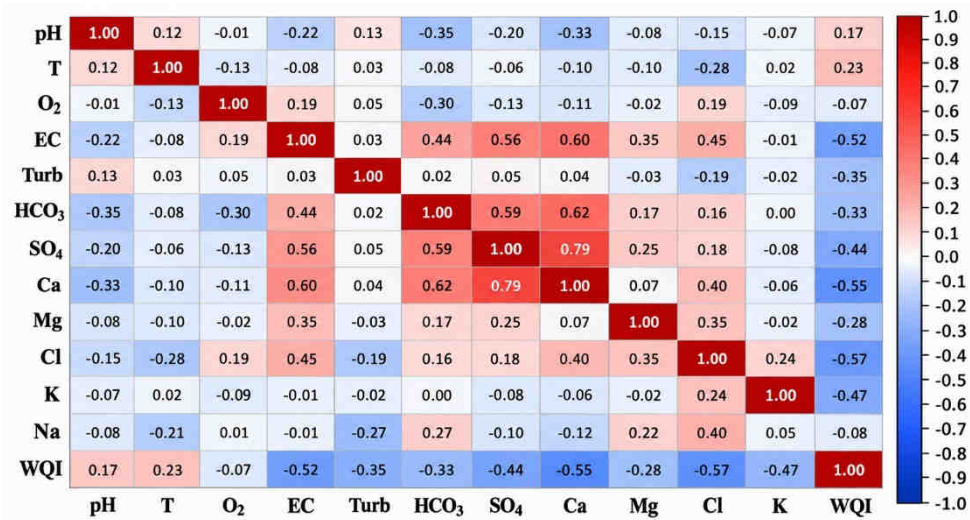


Fig. 2. Pearson correlation matrix of all studied variables and the computed WQI (N = 100)

Table 6

CCME-WQI classification of Sedrata groundwater wells

WQI Class	Index range	Water status	Sedrata wells
Excellent	95-100	Very good quality	0 wells (0 %)
Good	80-94	Good quality	3 wells (12 %)
Fair	65-79	Acceptable quality	7 wells (28 %)
Marginal	45-64	Poor quality	11 wells (44 %)
Poor	0-44	Very poor quality	4 wells (16 %)

Machine learning model performance and comparative evaluation

The predictive performance of six machine-learning algorithms across all 12 combinations of input variables was evaluated using separate training and testing datasets. Model performance was assessed using the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), relative absolute error (RAE) and root relative squared error (RRSE) (Table 7 and Fig. 3). Gradient Boosting (GB) achieved a perfect training fit ($R^2 = 1.000$) with negligible prediction error (RMSE = 0.150;

$MAE = 0.110$) and the highest testing performance ($R^2 = 0.880$; $RMSE = 4.890$). However, the decline in performance from training to testing indicates clear evidence of overfitting, further reflected by the increase in $RMSE$ from 0.150 to 4.890. Nevertheless, GB remains the most accurate model in terms of predictive performance on unseen data. This interpretation is further supported by the 5-fold cross-validation analysis. Support Vector Regression (SVR, Model 1) achieved a training R^2 of 0.950 and a testing R^2 of 0.830, corresponding to the same train-test R^2 difference (0.12) as GB but with substantially lower training accuracy, suggesting less severe overfitting. Its $RMSE$ increased moderately (2.910–5.810), while its RAE (38.10 %) and $RRSE$ (41.80 %) remained competitive, indicating good predictive stability. Bayesian Regression (BR, Model 2, 11 variables) exhibited the strongest generalisation ability. It is testing R^2 (0.820) slightly exceeded its training R^2 (0.810), indicating no apparent evidence of overfitting and excellent model stability. BR also showed the smallest train-test discrepancy among all models, while maintaining consistent prediction errors ($MAE = 4.320$ for training and 4.520 for testing), although its overall predictive accuracy remained slightly lower than that of the tuned GB model. Random Forest (RF, Model 2, 11 variables) demonstrated strong training performance ($R^2 = 0.950$), but its testing R^2 decreased to 0.800, accompanied by an increase in $RMSE$ from 2.710 to 6.230. Likewise, its RAE increased from 19.18% to 36.33 %, indicating moderate overfitting. Overall, RF provides robust predictive performance but exhibits weaker generalisation than BR and the tuned GB model.

Table 7
Comparative performance of machine learning regression models on training and testing datasets

Model	Algorithm	R^2 [-]	$RMSE$ [-]	MAE [-]	RAE [%]	$RRSE$ [%]	R^2 [-]	$RMSE$ [-]	MAE [-]	RAE [%]	$RRSE$ [%]
Training						Testing					
1	Gradient Boosting (GB)	1.00	0.15	0.11	1.13	1.17	0.88	4.89	3.10	25.31	35.21
2	Random Forest (RF)	0.95	2.71	1.94	19.18	21.76	0.80	6.23	4.45	36.33	44.85
1	Support Vector Regression (SVR)	0.95	2.91	1.95	19.28	23.36	0.83	5.81	4.66	38.10	41.80
2	Bayesian Regression (BR)	0.81	5.45	4.32	42.76	43.74	0.82	5.93	4.52	36.92	42.66
11	AdaBoost (AdB)	0.37	9.93	8.04	79.55	79.65	0.71	7.44	5.95	48.63	53.52
3	K-Nearest Neighbours (KNN)	0.50	8.82	6.76	66.87	70.69	0.48	9.98	7.82	63.88	71.81

The AdaBoost model (Model 11), developed using only pH and temperature, and was evaluated with the smallest predictor subset. Consequently, its training R^2 was relatively low (0.370), whereas it is testing R^2 reached 0.710. This inverted train-test pattern is most likely attributable to the small size of the testing subset ($n = 20$). With only 20 observations, the specific composition of the randomly selected test subset may differ from that of the training data, producing a testing R^2 that reflects the characteristics of those samples rather than true generalisation performance. This result should therefore be interpreted cautiously rather than as evidence of superior predictive ability. Nevertheless, it suggests that these two variables retain meaningful predictive information. However, the relatively high error values ($RAE = 48.63$ % and $RRSE = 53.52$ %) limit the practical applicability of this simplified model. K-Nearest Neighbours (KNN) exhibited the weakest predictive

performance, with the lowest testing R^2 (0.480) and the highest testing errors ($RMSE = 9.98$, $MAE = 7.82$, $RAE = 63.88\%$ and $RRSE = 71.81\%$). Even with ten input variables, the model failed to generalise effectively. Its relatively low training R^2 (0.500) further indicates limited fitting capability. This behaviour is likely attributable to the combination of correlated physicochemical variables, the limited sample size ($N = 100$), and heterogeneous, skewed data distributions, as documented in the descriptive statistics.

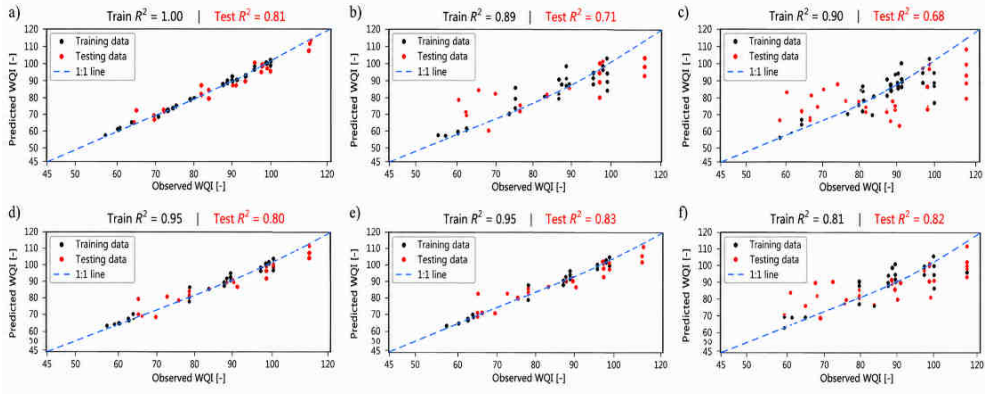


Fig. 3. Observed vs. predicted WQI for the best input combination of each machine-learning algorithm: a) Gradient Boosting, b) AdaBoost, c) K-Nearest Neighbors, d) Random Forest, e) Support Vector Regression, and f) Bayesian Regression; Black and red circles denote training and testing datasets, respectively, the dashed line represents the 1:1 line, and the R^2 values are given in each panel

Cross-validation and hyperparameter sensitivity

To verify whether the single 80/20 split reported in (Table 7) provides a reliable estimate of generalisation performance, a 5-fold cross-validation was additionally performed for Gradient Boosting, Random Forest and Support Vector Regression, both with default (scikit-learn) hyperparameters and with the tuned configurations identified in the hyperparameter optimisation section and summarised in Table 8.

Table 8

5-fold cross-validated R^2 and $RMSE$ for default vs. tuned models (full 12-variable input set)

Algorithm	Configuration	CV R^2 (mean $\pm$ SD)	CV $RMSE$ (mean $\pm$ SD)
Gradient boosting	Default	0.77 $\pm$ 0.06	6.63 $\pm$ 1.1
Gradient boosting	Tuned	0.85 $\pm$ 0.03	5.3
Random forest	Default	0.68 $\pm$ 0.08	7.7 $\pm$ 1.2
Support vector regression	Default	0.38 $\pm$ 0.04	10.8 $\pm$ 1.1
Support vector regression	Tuned	0.75	-

With default hyperparameters, Gradient Boosting's mean cross-validated R^2 (0.77 $\pm$ 0.06) is materially lower than the single-split testing R^2 of 0.88 reported in Table 7, confirming that the original 80/20 split was, to some extent, an optimistic estimate of generalisation performance. Following hyperparameter tuning, the cross-validated R^2 for Gradient Boosting improves to 0.85 $\pm$ 0.03, with both a higher mean and substantially

reduced variance across folds, indicating that the shallower trees and larger minimum leaf size identified by the grid search meaningfully mitigate the overfitting noted in the model comparison discussion. Random Forest and Support Vector Regression show a smaller default-to-tuned gap, consistent with their comparatively stable train-test performance in Table 7.

Best-subset regression analysis

The best subset multiple regression results for all 12 parameter combinations are summarised in Table 9. Of the 12 candidate models, the nine-variable combination comprising temperature (T), electrical conductivity (EC), turbidity ($Turb$), bicarbonate (HCO_3^-), magnesium (Mg^{2+}), calcium (Ca^{2+}), chloride (Cl^-), potassium (K^+) and sodium (Na^+) achieved the lowest values of both AIC (636.75) and Mallows' C_p (7.91), while also achieving the highest adjusted R^2 value (0.812) and an MSE of 27.93.

Adding the remaining three variables (pH, dissolved oxygen and sulphate (SO_4^{2-})) to create the full 12-variable model increased R^2 by only 0.002 (from 0.812 to 0.814), which is a marginal gain that does not justify the additional measurement costs. This is consistent with the principle of parsimony. The nine-variable model is therefore recommended as the operational input configuration for real-time WQI prediction in the Sedrata aquifer monitoring programme.

Table 9
Best-subset regression results for WQI modelling in the Sedrata groundwater system [$N = 100$].
Bold row indicates the recommended parsimonious model

k	Variables	MSE	R^2	Adj R^2	AIC	BIC	Mallows C_p
1	Cl^-	109.44	0.33	0.32	757.33	762.54	248.50
2	Ca^{2+}, K^+	73.38	0.55	0.54	719.36	727.18	137.00
3	$Turb, Ca^{2+}, K^+$	55.26	0.66	0.65	693.00	703.42	81.96
4	$Turb, Ca^{2+}, Cl^-, K^+$	36.65	0.77	0.76	653.93	666.96	25.37
5	$Turb, Ca^{2+}, Mg^{2+}, Cl^-, K^+$	33.17	0.79	0.78	645.97	661.61	16.43
6	$T, Turb, Ca^{2+}, Mg^{2+}, Cl^-, K^+$	31.38	0.80	0.79	642.41	660.65	12.78
7	$T, EC, Turb, Ca^{2+}, Mg^{2+}, Cl^-, K^+$	29.96	0.81	0.80	639.78	660.62	10.31
8	$T, EC, Turb, HCO_3^-, Ca^{2+}, Mg^{2+}, K^+, Na^+$	29.18	0.82	0.80	639.15	662.60	9.86
9	$T, EC, Turb, HCO_3^-, Ca^{2+}, Mg^{2+}, Cl^-, K^+, Na^+$	27.92	0.82	0.81	636.75	662.80	7.91
10	$T, EC, Turb, HCO_3^-, SO_4^{2-}, Ca^{2+}, Mg^{2+}, Cl^-, K^+, Na^+$	27.799	0.83	0.81	638.29	666.94	9.50
11	$T, O_2, EC, Turb, HCO_3^-, SO_4^{2-}, Ca^{2+}, Mg^{2+}, Cl^-, K^+, Na^+$	27.693	0.83	0.80	639.90	671.17	11.17
12	$pH, T, O_2, EC, Turb, HCO_3^-, SO_4^{2-}, Ca^{2+}, Mg^{2+}, Cl^-, K^+, Na^+$	27.639	0.83	0.80	641.71	675.58	13.00

Multicollinearity assessment

The variance inflation factors (VIFs) for all 12 predictor variables are presented in Table 10. All VIF values fall below the conventional threshold of 10, confirming the absence of severe multicollinearity. Calcium has the highest VIF value (7.08), as expected given its covariation with EC , SO_4^{2-} and HCO_3^- through coupled carbonate-evaporite dissolution. However, it does not breach the critical threshold. All the other variables have VIF values below 4.0, with pH (1.31), turbidity (1.17) and temperature (1.14) showing near-orthogonal variation.

Table 10

Variance Inflation Factor (VIF) statistics for the 12 physicochemical parameters

Variable	VIF	1/VIF
Ca ²⁺	7.08	0.1412
SO ₄ ²⁻	3.97	0.2521
Cl ⁻	3.56	0.2807
HCO ₃ ⁻	3.15	0.3175
EC	2.36	0.4242
Na ⁺	2.26	0.4416
Mg ²⁺	1.73	0.5770
O ₂	1.50	0.6673
pH	1.31	0.7632
K ⁺	1.29	0.7773
Turb	1.17	0.8569
T	1.14	0.8740

Feature importance

Permutation importance (30 repeats, scored on held-out R^2) was computed for every model using the full 12-variable input combination, providing a feature-ranking method applicable uniformly across all six algorithm types, including SVR and K-NN, which have no native importance score (Table 11 and Fig. 4).

Table 11

Permutation-importance scores (ΔR^2) of the predictive models

Variable	GB	RF	BR	AdB	KNN	SVR
pH	0.0014	0.0044	0.0000	0.0015	0.0000	0.0000
T	0.0014	0.0018	0.0000	0.0053	0.0000	0.0000
O ₂	0.0024	0.0036	0.0063	0.0084	0.0000	0.0322
EC	0.0479	0.1056	0.0686	0.0480	0.0248	0.0723
Turb	0.2096	0.1864	0.3100	0.1935	0.0441	0.2906
HCO ₃ ⁻	0.0065	0.0072	0.0100	0.0000	0.0000	0.0038
SO ₄ ²⁻	0.0341	0.0260	0.0099	0.0228	0.0000	0.0319
Ca ²⁺	0.2727	0.1314	0.2719	0.0734	0.0013	0.1663
Mg ²⁺	0.0003	0.0000	0.0428	0.0016	0.0000	0.0003
Cl ⁻	0.1100	0.1577	0.0852	0.0884	0.1665	0.2193
K ⁺	0.2020	0.1612	0.0907	0.1181	0.0045	0.0769
Na ⁺	0.0467	0.0292	0.0000	0.0041	0.0000	0.0086

Turbidity, chloride, calcium, potassium and electrical conductivity emerged as the dominant predictors across all six algorithms, consistent with the parameters that most frequently exceeded guideline objectives in the WQI calculation itself (Table 4). Magnesium, bicarbonate, pH, dissolved oxygen and temperature contributed little to no predictive power in any model, in line with their weak correlation with the computed WQI (Table 5) and their exclusion (T, O₂, HCO₃⁻ partly) or minor weighting in the index formula.

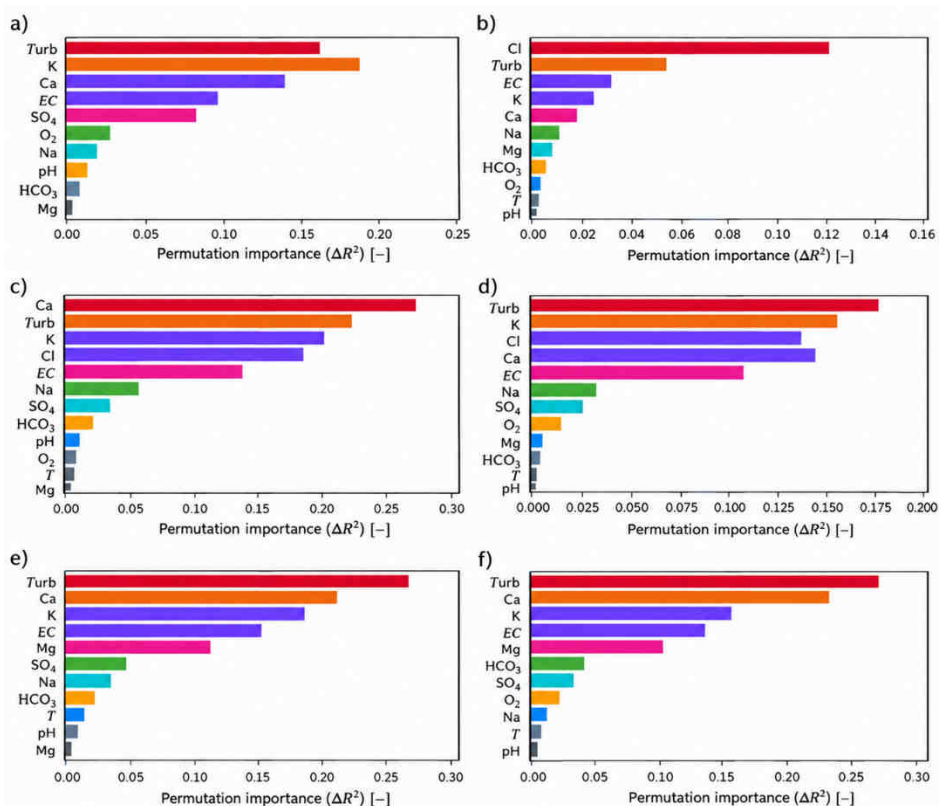


Fig. 4. Permutation feature importance (ΔR^2) of predictors across the six WQI models: a) AdaBoost, b) K-Nearest Neighbors, c) Gradient Boosting, d) Random Forest, e) Support Vector Regression, and f) Bayesian Regression

Discussion

Hydrochemical controls on groundwater quality

Physicochemical characterisation of the Sedrata phreatic aquifer reveals considerable stress in the groundwater system due to natural mineralisation and anthropogenic pressures. The wide range of electrical conductivities (400-5,160) $\mu\text{S}\cdot\text{cm}^{-1}$ and major ion concentrations is consistent with the Souk Ahras region's heterogeneous lithological setting, in which Triassic evaporite sequences contribute gypsum, halite and associated salts to shallow groundwater through congruent dissolution [4, 5]. The strong positive correlations observed between Ca^{2+} and SO_4^{2-} ($r = 0.79$) and between Ca^{2+} and HCO_3^- ($r = 0.62$) are indicative of gypsum and calcite dissolution, respectively, and corroborate petrographic evidence of evaporite and carbonate lithologies in the recharge area. The co-enrichment of Na^+ and Cl^- ($r = 0.40$) further implicates halite dissolution as a secondary salinity pathway, in line with findings from similar semi-arid aquifers in North Africa [7, 22]. The CCME-WQI results confirm that the quality of groundwater in the Sedrata plain is predominantly poor: 60 % of sampled wells fall into the marginal or poor categories, with

only 12 % achieving good status. These findings are consistent with previous basin-scale evaluations which identified elevated sulphate, chloride and calcium concentrations as the primary causes of quality degradation in shallow Algerian aquifers [2]. The pronounced seasonal signal and lower WQI scores in May, compared to February and December, reflect the dual mechanisms of evaporative concentration and agricultural return flow. During peak irrigation, water abstraction lowers the water table and concentrates solutes, while return flows from unlined channels reintroduce agrochemical loads, particularly potassium (skewness = 4.49), into the shallowest portion of the aquifer. This seasonal pattern highlights the importance of multi-campaign monitoring protocols rather than single-season assessments, as the latter are at risk of systematically over- or underestimating the true quality of the aquifer [7].

Machine learning (model comparison)

Of the six machine-learning algorithms evaluated, Gradient Boosting (GB) achieved the highest absolute predictive accuracy on the full 12-variable input set (testing $R^2 = 0.880$, $RMSE = 4.890$). This is consistent with its well-documented ability to identify complex, non-linear relationships between predictors and responses in hydrogeological datasets [23, 24]. However, the perfect training R^2 value of 1.000, coupled with the moderate test R^2 value, reveals a degree of overfitting that warrants caution when deploying the algorithm in practice without regularisation tuning or ensemble averaging. Support Vector Regression (SVR) and Bayesian Regression (BR) were the most generalisable algorithms: SVR achieved a testing R^2 of 0.830 with only two variables removed, while BR yielded training and testing R^2 values of 0.810 and 0.820, respectively, indicating minimal variance inflation and stable performance across the training-testing split. These results corroborate the comparative analysis of Ahmed et al. [8], who also identified ensemble methods and kernel-based regressors as the most effective algorithms for predicting WQI in semi-arid basins. The substantially weaker performance of K-Nearest Neighbours (testing $R^2 = 0.480$, $RRSE = 71.81\%$) can be attributed to its sensitivity to the high-dimensional, heterogeneous input space, as well as the limited sample size ($N = 100$). Both of these factors impair neighbourhood distance calculations [44].

The 5-fold cross-validation analysis quantitatively confirms the overfitting tendency of the default Gradient Boosting (GB) model. Its mean cross-validated R^2 (0.77 ± 0.06) is substantially lower than the single-split testing R^2 (0.88), indicating that the latter provides an optimistic estimate of generalisation performance. Following hyperparameter tuning (shallower trees and larger leaf size; see the hyperparameter optimisation section), the cross-validated R^2 increased to 0.85 ± 0.03 , with markedly reduced variance across folds. Although the tuned GB model achieved the highest predictive accuracy, Support Vector Regression (SVR) and particularly Bayesian Regression (BR) exhibited more stable train-test performance (Table 7), suggesting stronger generalisation despite their slightly lower predictive accuracy.

The present results are consistent with recent studies conducted in semi-arid regions of Algeria. Azlaoui et al. [45] reported excellent IWQI prediction using XGBoost combined with SHAP in the Ain Oussera Plain, whereas Nefla et al. [46] successfully applied Random Forest, XGBoost and AdaBoost with SHAP to irrigation water quality assessment in the Skikda aquifer. However, unlike the present study, these investigations did not explicitly evaluate the generalisation gap through k-fold cross-validation. Likewise, Farhi et al. [47] demonstrated that spatial data augmentation can improve predictive performance in

the data-scarce Tabelbala region, highlighting a promising strategy for addressing small-sample limitations.

Nevertheless, the relatively small dataset (100 samples from 25 wells) remains a limitation of the present study. Although sampling over four independent seasonal campaigns improves the representativeness of the dataset, parameter-rich algorithms such as GB and Random Forest remain susceptible to overfitting, while instance-based methods such as K-Nearest Neighbours are more sensitive to limited sample size. Future studies should therefore expand the groundwater monitoring network over multiple years and evaluate spatial data augmentation as a complementary strategy for improving model robustness and generalisation.

Optimal input combination and operational recommendations

Best-subset regression analysis identified a parsimonious nine-variable model (T , EC , $Turb$, HCO_3^- , Ca^{2+} , Mg^{2+} , Cl^- , K^+ and Na^+ ; adj. $R^2 = 0.812$, $AIC = 636.75$, Mallows' $C_p = 7.91$) as the preferred operational input configuration. The marginal increase in R^2 achieved by incorporating pH, dissolved oxygen and sulphate into the complete 12-variable model (0.831 versus 0.829) is insufficient to justify the extra analytical workload. This is consistent with the principle of parsimony and the observation that pH and O_2 exhibit weak correlations with WQI in this carbonate-buffered, aerobic system ($r = +0.17$ and -0.07 , respectively). From a water-quality monitoring perspective, this result is practically significant: excluding three parameters from routine laboratory analyses reduces per-sample costs without meaningfully compromising model accuracy. This enables more frequent or spatially extensive sampling within fixed budgets. It is noteworthy that SO_4^{2-} was excluded from the recommended model, given its high exceedance frequency. This can be explained by the strong collinearity between SO_4^{2-} and Ca^{2+} ($r = 0.79$), which renders SO_4^{2-} statistically redundant once Ca^{2+} is included. This pattern is also evident in the relatively modest VIF of SO_4^{2-} (3.97) compared to that of Ca^{2+} (VIF = 7.08).

The VIF analysis confirms that multicollinearity does not invalidate the regression framework, with all VIF values remaining below the critical threshold of 10 and the majority falling below 4.0. The near-orthogonal variation of temperature (VIF = 1.14), turbidity (VIF = 1.17) and pH (VIF = 1.31) reflects the distinct and largely independent sources of variability associated with seasonal thermal cycles, event-driven turbidity pulses and carbonate equilibrium, respectively. The moderate Ca^{2+} VIF value (7.08) indicates the well-known geochemical coupling between carbonate, sulphate and bicarbonate enrichment. However, it falls short of the threshold associated with regression instability [48].

Management implications

From a water resource management perspective, the spatial clustering of poor-quality wells in the south-eastern sector, adjacent to the deepest Triassic evaporite outcrops, suggests that remediation efforts should be prioritised in this area. Targeted interventions could include installing lining in irrigation return flow channels to reduce solute recharge, promoting deficit irrigation scheduling from May onwards to minimise evaporative concentration and substituting phreatic well water with piped supplies for domestic use in the most affected areas. The GB and SVR models trained in this study could be incorporated into a decision-support dashboard providing near-real-time WQI estimates

from routine field measurements of the nine recommended parameters. This would facilitate proactive management responses to quality deterioration events.

Conclusion

Physicochemical monitoring of the Sedrata phreatic aquifer across four hydrological seasons reveals a groundwater system under sustained stress from evaporite dissolution and irrigation-induced salinisation. CCME-WQI classification confirms that 60 % of sampled wells fall into the Marginal or Poor categories, with Cl^- , Ca^{2+} , EC , K^+ and SO_4^{2-} most frequently exceeding WHO and Algerian standards. The pronounced deterioration observed during the May irrigation campaigns, relative to the February and December recharge periods, demonstrates that single-season assessments systematically misrepresent true aquifer quality, and that multi-campaign monitoring is a methodological necessity rather than a luxury in semi-arid agricultural settings.

Among the six machine learning algorithms evaluated, Gradient Boosting achieved the highest absolute predictive accuracy (test $R^2 = 0.88$, $RMSE = 4.89$) but exhibited significant overfitting on this 100-sample dataset. Bayesian Regression and Support Vector Regression offered superior operational value through their generalisation stability (test R^2 of 0.82 and 0.83 respectively, with near-zero train-test gap), and are recommended for deployment in real-time monitoring contexts. Exhaustive best-subset regression identified a parsimonious nine-variable input configuration (T , EC , $Turb$, HCO_3^- , Ca^{2+} , Mg^{2+} , Cl^- , K^+ and Na^+ ; adj. $R^2 = 0.812$) that sacrifices only 0.002 in R^2 relative to the full twelve-variable model while eliminating pH , O_2 and SO_4^{2-} from routine analysis, a cost reduction achievable without meaningful loss of predictive fidelity. Three of the nine recommended parameters are measurable in situ with portable instruments, enabling near-real-time WQI estimation between laboratory sampling cycles.

The integrated CCME-WQI and multi-algorithm ML framework demonstrated here is directly transferable to other semi-arid phreatic aquifer systems underlain by carbonate and evaporite sequences, particularly across the Maghreb, the Middle East and the wider Mediterranean basin. Priority future work should expand the monitoring network temporally and spatially, incorporate microbiological parameters into the WQI calculation, apply isotopic and hydrogeochemical modelling to better constrain solute residence times, and explore the addition of remote sensing proxies, land-surface temperature, vegetation indices and soil salinity estimates, as supplementary ML inputs to extend spatial prediction coverage in data-sparse regions.

Acknowledgment

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R120), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

References

- [1] Priyan K. Issues and challenges of groundwater and surface water management in semi-arid regions. In: Pande CB, Moharir KN, editors. Groundwater Resources Development and Planning in the Semi-Arid Region. Cham: Springer; 2021. p. 1-17. DOI: 10.1007/978-3-030-68124-1_1.
- [2] Fida M, Li P, Wang Y, Alam SMK, Nsabimana A. Water contamination and human health risks in Pakistan: a review. *Exposure Health*. 2023;15(3):619-39. DOI: 10.1007/s12403-022-00512-1.

- [3] McGrath TF, Shishkina N, Theobald R, Rodriguez C. Review of the regulatory requirements for the maintenance of aerobic treatment units and greywater treatment systems in Western Australia. *J Water Resour Prot.* 2015;7(17):1502-15. DOI: 10.4236/jwarp.2015.717124.
- [4] Chabbi A, Chermiti A, Brusset S. Stratigraphy and structural style of Souk Ahras foreland fold-thrust belt in northeastern Algeria. *Arab J Geosci.* 2021;14(17):1789. DOI: 10.1007/s12517-021-08170-4.
- [5] Leprêtre R, Frizon de Lamotte D, Combier V, Gimeno-Vives O, Mohn G, Eschard R. The Tell-Rif orogenic system (Morocco, Algeria, Tunisia) and the structural heritage of the southern Tethys margin. *BSGF - Earth Sci Bull.* 2018;189(2):10. DOI: 10.1051/bsgf/2018009.
- [6] CCME. Canadian water quality guidelines for the protection of aquatic life: CCME water quality index 1.0, technical report. Winnipeg: Canadian Council of Ministers of the Environment; 2001.
- [7] Uddin MG, Nash S, Rahman A, Olbert AI. A comprehensive method for improvement of water quality index (WQI) models for coastal water quality assessment. *Water Res.* 2022;219:118532. DOI: 10.1016/j.watres.2022.118532.
- [8] Ahmed AA, Sayed S, Abdoulhalik A, Moutari S, Oyedele L. Applications of machine learning to water resources management: a review of present status and future opportunities. *J Clean Prod.* 2024;441:140715. DOI: 10.1016/j.jclepro.2024.140715.
- [9] Huan S, Liu X. Assessing the nonlinear relationship between consumer goods and water pollution in different seasons with machine learning models: a case study in the Yangtze River Economic Belt. *J Clean Prod.* 2024;444:141254. DOI: 10.1016/j.jclepro.2024.141254.
- [10] Bellazi MA. Diagnostic de la pollution de certains milieux aquatiques de la Région de Souk Ahras [Diagnosis of the pollution of certain aquatic environments in the Souk Ahras Region] [PhD thesis]. Souk Ahras: University of Souk Ahras; 2021. Available from: <https://univ-soukahras.dz/fr/publication/article/2811>.
- [11] Dubourdieu G. Étude géologique de la région de l'Ouenza (confins algéro-tunisiens) [Geological study of the Ouenza region (Algerian-Tunisian borders)]. Algiers: Publ Serv Carte Géol Algérie; 1956. Bulletin No. 10, Vol. 1, 659 p. Available from: <https://books.google.com/books/about/?id=dTcjnQAACAAJ>.
- [12] Nadour F, Adjili M, Chabbi A, Nadour M, Zenati N, Belahcene N, et al. Explainable ML for Irrigation Water Quality Prediction in the Sedrata Aquifer, Plain (Algeria). *Water.* 2026;18:2215. DOI: 10.3390/w18172215.
- [13] David L. Étude géologique des monts de la haute Medjerda [Geological study of the mounts of the high Medjerda]. Algiers: Service de la carte géologique de l'Algérie; 1956. 304 p. Available from: <https://search.worldcat.org/title/80247012>.
- [14] Khelil M, Khomsi S, Roure F, Zargouni F. The Eastern Tell of North Africa: styles, wedge geometry and tectonics. *Mediterr Geosci Rev.* 2023. DOI: 10.1007/s42990-023-00113-3.
- [15] Hinze WJ, von Frese RRB, Saad AH. Gravity and Magnetic Exploration: Principles, Practices, and Applications. Cambridge: Cambridge University Press; 2013. DOI: 10.1017/CBO9780511740012.
- [16] Malek A, Kahoul M, Bouguerra H. Groundwater's physicochemical and bacteriological assessment: case study of well water in the region of Sedrata, North-East of Algeria. *J Water Land Dev.* 2019;41(IV-VI):91-100. DOI: 10.2478/jwld-2019-0032.
- [17] American Public Health Association, American Water Works Association, Water Environment Federation. (1995). Standard methods for the examination of water and wastewater. Eaton AD, Clesceri LS, Greenberg AE, Eds.; 19th ed.). Washington, DC: American Public Health Association. ISBN: 9780875532233.
- [18] Randazzo G, Stanley DJ, Di Geronimo SI, Amore C. Human-induced sedimentological changes in Manzala Lagoon, Nile Delta, Egypt. *Environ Geol.* 1998;36(3-4):235-58. DOI: 10.1007/s002540050340.
- [19] Fritz SJ. A survey of charge-balance errors on published analyses of potable ground and surface waters. *Groundwater.* 1994;32(4):539-46. DOI: 10.1111/j.1745-6584.1994.tb00888.x.
- [20] Hounslow A. Water Quality Data: Analysis and Interpretation. Boca Raton: CRC Press; 2018. DOI: 10.1201/9780203734117.
- [21] Shawn MM, Abdel Hamed AR. Chemical constituents of some wells water in Sinai (Egypt) and their suitability of quality to different use activities. *Fayoum J Agric Res Dev.* 2016;30(2):134-48. DOI: 10.21608/fjard.2016.191752.
- [22] Parween S, Siddique NA, Diganta MTM, Olbert AI, Uddin MG. Assessment of urban river water quality using modified NSF water quality index model at Siliguri city, West Bengal, India. *Environ Sust Indic.* 2022;16:100202. DOI: 10.1016/j.indic.2022.100202.
- [23] Khoi DN, Quan NT, Linh DQ, Nhi PTT, Thuy NTD. Using machine learning models for predicting the water quality index in the La Buong River, Vietnam. *Water.* 2022;14(10):1552. DOI: 10.3390/w14101552.
- [24] Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat.* 2001;29(5):1189-232. DOI: 10.1214/aos/1013203451.
- [25] Rigatti SJ. Random forest. *J Insur Med.* 2017;47(1):31-9. DOI: 10.17849/insm-47-01-31-39.1.

- [26] Wang F, Wang Y, Zhang K, Hu M, Weng Q, Zhang H. Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation. *Environ Res.* 2021;202:111660. DOI: 10.1016/j.envres.2021.111660.
- [27] Sharma A, Goyal MK. A comparison of three soft computing techniques, Bayesian regression, support vector regression, and wavelet regression, for monthly rainfall forecast. *J Intell Syst.* 2017;26(4):641-55. DOI: 10.1515/jisys-2016-0065.
- [28] Bourel M, Segura AM. Multiclass classification methods in ecology. *Ecol Indic.* 2018;85:1012-21. DOI: 10.1016/j.ecolind.2017.11.031.
- [29] Sakizadeh M, Mirzaei R. A comparative study of performance of K-nearest neighbors and support vector machines for classification of groundwater. *J Min Environ.* 2016;7(2):149-64. DOI: 10.22044/jme.2016.534.
- [30] Ahmed U, Mumtaz R, Anwar H, Shah AA, Irfan R, Garcia-Nieto J. Efficient water quality prediction using supervised machine learning. *Water.* 2019;11(11):2210. DOI: 10.3390/w11112210.
- [31] Liu S, Tai H, Ding Q, Li D, Xu L, Wei Y. A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. *Math Comput Model.* 2013;58(3-4):458-65. DOI: 10.1016/j.mcm.2011.11.021.
- [32] Nguyen QH, Ly HB, Ho LS, Al-Ansari N, Le HV, Tran VQ, et al. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Math Probl Eng.* 2021;2021:4832864. DOI: 10.1155/2021/4832864.
- [33] Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)? *Geosci Model Dev.* 2014;7:1247-50. DOI: 10.5194/gmd-7-1247-2014.
- [34] Siham A, Sara S, Abdellah A. Feature selection based on machine learning for credit scoring: an evaluation of filter and embedded methods. In: 2021 Int Conf Innovat Intelligent Systems Applications (INISTA). IEEE; 2021:1-6. DOI: 10.1109/INISTA52262.2021.9548410.
- [35] Singha S, Pasupuleti S, Singha SS, Singh R, Kumar S. Prediction of groundwater quality using efficient machine learning technique. *Chemosphere.* 2021;276:130265. DOI: 10.1016/j.chemosphere.2021.130265.
- [36] Hameed M, Sharqi SS, Yaseen ZM, Afan HA, Hussain A, Elshafie A. Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia. *Neural Comput Appl.* 2017;28(S1):893-905. DOI: 10.1007/s00521-016-2404-7.
- [37] Sakar BGBE. Variable importance analysis in default prediction using machine learning techniques. 2018. Available from: <https://www.scitepress.org/papers/2018/68724/68724.pdf>.
- [38] Wong WY, Al-Ani AKI, Hasikin K, Khairuddin ASM, Razak SA, Hizaddin HF, et al. Water quality index using modified random forest technique: assessing novel input features. *Comput Model Eng Sci.* 2022;132(3):1011-38. DOI: 10.32604/cmescs.2022.019244.
- [39] Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Control.* 1974;19(6):716-23. DOI: 10.1109/TAC.1974.1100705.
- [40] Schwarz G. Estimating the dimension of a model. *Ann Stat.* 1978;6(2):461-4. DOI: 10.1214/aos/1176344136.
- [41] Mallows CL. Some comments on Cp. *Technometrics.* 2000;42(1):87-94. DOI: 10.1080/00401706.2000.10485984.
- [42] Gilmour SG. The interpretation of Mallows's Cp-statistic. *J R Stat Soc Ser D (The Statistician).* 1996;45(1):49-56. DOI: 10.2307/2348411.
- [43] Kuhn M, Johnson K. *Applied Predictive Modeling*. New York: Springer; 2013. DOI: 10.1007/978-1-4614-6849-3.
- [44] Modaresi F, Araghinejad S. A comparative assessment of support vector machines, probabilistic neural networks, and K-nearest neighbor algorithms for water quality classification. *Water Resour Manage.* 2014;28(12):4095-111. DOI: 10.1007/s11269-014-0730-z.
- [45] Azlaoui M, Karef S, Foufou A, Haied N, Azlaoui N, Rabehi A, et al. Machine learning-based prediction of irrigation water quality index with SHAP interpretability: application to groundwater resources in the semi-arid region, Algeria. *Water.* 2026;18(8):959. DOI: 10.3390/w18080959.
- [46] Nefla L, Bergal A, Boumaraf W, Gheid S, Bouksiba C, Khammar H, et al. Integrated GIS-machine learning approach to irrigation water quality assessment in coastal aquifers. *Sci Rep.* 2026. DOI: 10.1038/s41598-025-25461-y.
- [47] Farhi N, Al-Mutiry M, Bennia A, Kreri S, Djerida A, Kebir LW, et al. WQI-machine learning integration with spatial data augmentation for robust groundwater quality assessment in data-limited arid regions. *Sustainability.* 2026;18(7):3493. DOI: 10.3390/su18073493.
- [48] O'Brien RM. A caution regarding rules of thumb for variance inflation factors. *Qual Quant.* 2007;41:673-90. DOI: 10.1007/s11135-006-9018-6.