SCIENTIFIC Publishing Limited

Land Management and Utilization
https://ojs.ukscip.com/index.php/lmu

Article

Evaluation and Prediction of Key Parameters of Water Quality by Machine Learning in the Coastal Area, Southern Iran

Fariborz Mohammadi [1] and Maryam Heydarzadeh [2,*]

[1] Hormoz Studies and Research Center, University of Hormozgan, Bandar Abbas 7916193145, Iran

[2] Department of Water Engineering, Faculty of Agriculture, Minab Higher Education Center, University of Hormozgan, Bandar Abbas 7916193145, Iran

[*] Correspondence: m.heydarzadeh@hormozgan.ac.ir

Received: 12 February 2026; **Revised:** 22 May 2026; **Accepted:** 10 June 2026; **Published:** 25 June 2026

Abstract: Groundwater is the most reliable source of freshwater in arid and semi-arid regions, where surface water resources are scarce and highly variable. Its role in agriculture, drinking water supply, and ecosystem stability highlights the importance of reliable quality assessment. However, traditional monitoring approaches are often costly, labor-intensive, and spatially limited. Machine learning (ML) techniques offer promising alternatives for accurately predicting groundwater quality using a reduced set of hydrochemical parameters. This study evaluated the application of ML algorithms integrated with the Water Quality Index (WQI) framework to predict groundwater quality in the Minab Plain, Hormozgan Province, Iran. Groundwater samples were collected from 15 wells between 2010 and 2023 and analysed for six physicochemical indicators. Correlation analysis was performed to identify the most influential parameters, followed by the application of Decision Tree (DT), k-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Random Forest (RF) models to forecast WQI under variable input scenarios. WQI values ranged from 29.74 to 107.2, with 95.6% of samples classified as “Good” and 4.3% as “Excellent.” pH showed the strongest correlation with WQI (R^2 = 0.98), while Na and total dissolved solids (TDS) exhibited the weakest associations. Among the tested models, DT achieved the highest classification accuracy within this dataset; however, the results should be interpreted as case-study-specific rather than universally generalizable. The findings demonstrate that ML-based WQI modeling provides a robust, potentially cost-effective approach for groundwater quality prediction, supporting sustainable water management in arid and semi-arid regions.

Keywords: Groundwater Quality; Water Quality Index (WQI); Machine Learning (ML); Arid and Semi-Arid Regions; Coastal Area

1. Introduction

Groundwater is the principal source of irrigation and domestic supply in many regions, particularly in arid and semi-arid areas where surface water resources are scarce and unreliable [1]. The global importance of groundwater was underscored by the 2022 UN-Water campaign, Groundwater: Make the Invisible Visible, which highlighted that this resource constitutes nearly 99% of the Earth’s unfrozen freshwater reserves, yet remains inadequately recognized, monitored, and protected [2]. Despite its fundamental role in sustaining human and ecological systems, groundwater is increasingly threatened by overexploitation and declining quality [3].

Groundwater quality degradation is strongly linked to anthropogenic activities, including intensive agricultural practices, industrial discharges, and poor waste management, which collectively pose risks to human health and ecosystem stability [4]. For irrigation purposes, groundwater quality varies widely across regions, with salinity and chemical composition posing major challenges for sustainable agriculture [5,6]. Poor management practices exacerbate these problems, leading to reduced freshwater availability for human consumption and long-term ecological harm [7–9]. Strengthened research and monitoring are therefore essential to safeguard public health and ensure ecosystem integrity [3,10]. Traditional approaches to groundwater monitoring, such as field-based sampling and statistical regression models, face significant limitations. These methods are often costly and time-intensive, particularly when applied to large spatial domains [11]. Furthermore, statistical and empirical models generally require a fixed number of parameters and exhibit limited adaptability to dynamic hydrogeological conditions or to newly available datasets [12]. As groundwater systems are inherently complex and nonlinear, such approaches may fail to capture the full range of interactions among the influencing variables, thereby reducing their predictive power [13]. Machine learning (ML) techniques have emerged as powerful tools to address these limitations. Unlike traditional statistical approaches, ML models can handle nonlinear interdependencies and incorporate diverse datasets, thereby enhancing predictive accuracy and interpretability [12,13]. ML approaches also enable the use of reduced sets of input parameters while maintaining high accuracy, making them cost-effective and scalable solutions for groundwater quality prediction [14,15]. Recent advances in ML, including supervised, unsupervised, and reinforcement learning, have expanded the scope of groundwater research [16–18]. Supervised learning methods such as neural networks, support vector machines (SVM), and random forests (RF) are widely used for classification and regression tasks, including groundwater quality prediction [19,20]. Unsupervised methods, including clustering and dimensionality reduction techniques such as principal component analysis (PCA), are effective for identifying hidden structures in hydrochemical datasets [18]. Reinforcement learning, though less commonly applied in hydrogeology, offers potential for adaptive water resource management [17]. Applications of ML in groundwater studies have expanded over the past two decades, particularly in groundwater level forecasting, quality assessment, and aquifer vulnerability mapping [21,22]. Various models, including RF, SVM, Gradient Boosting, and XGBoost, have been successfully applied to predict the Water Quality Index (WQI), a widely accepted measure that integrates multiple water quality parameters into a single quantitative index [23,24]. WQI assessments commonly incorporate parameters such as pH, total dissolved solids (TDS), calcium, magnesium, and nitrates, with entropy-weighted water quality index (EWQI) approaches frequently employed to assign importance to different indicators [25,26]. Large datasets from multiple monitoring sites have been shown to enhance the robustness of ML-based predictions, enabling improved spatial mapping and regional understanding of groundwater quality [13]. For instance, Xu et al. [27] applied linear regression and seven ML models to water quality data from 17 rivers in Jiangsu Province, China, and reported that RF and XGB achieved the highest predictive reliability for WQI grades, with accuracies exceeding 85%. However, challenges remain, particularly regarding spatial uncertainty in predictions, data quality, and the integration of explainable AI approaches to support transparent decision-making [25]. In Iran, the Minab Plain aquifer is particularly vulnerable due to unsustainable withdrawals, desertification, and the proliferation of deep and semi-deep wells [28,29]. Overexploitation not only reduces water availability but also degrades water quality, especially in shallow aquifers [30]. Understanding groundwater quality dynamics in this region is therefore essential for ensuring agricultural sustainability and protecting drinking water resources.

Although WQI is conventionally calculated using all measured hydrochemical variables, routine groundwater monitoring is often constrained by analytical costs, laboratory capacity, and sampling frequency. Therefore, identifying the minimum number of hydrochemical parameters required for reliable WQI prediction is of considerable practical importance. Machine learning provides an opportunity not only to predict WQI but also to identify the smallest subset of variables that preserves prediction accuracy while reducing monitoring costs. Therefore, this study aims to (i) identify the minimum hydrochemical dataset required for reliable WQI prediction, (ii) compare the predictive performance of four machine learning algorithms using progressively reduced input variables, and (iii) evaluate the practical applicability of reduced-input models for groundwater quality monitoring in a coastal aquifer.

The novelty of this study lies in four aspects. First, six different variable-reduction scenarios were evaluated to identify the minimum hydrochemical dataset required for reliable WQI prediction. Second, both regression (continuous WQI prediction) and classification (WQI grades) performance were assessed simultaneously. Third, the study identifies the most influential groundwater quality parameters for efficient monitoring of this coastal aquifer. Finally, the proposed framework was evaluated using a long-term groundwater monitoring dataset (2010–2023), providing practical implications for cost-effective groundwater quality assessment.

2. Material and Method

2.1. Study Area

The Minab River Basin is located in Hormozgan Province, southern Iran, within the Makran folded zone south of the Jazmurian Depression. It extends eastward and northeastward toward the Minab and Zandan faults and southwestward to the Oman Sea, covering an area of approximately 10,421 km^2. Elevations range from 2,704 m in the upstream mountains to −16 m near the Persian Gulf, creating diverse topography that influences hydrological and land-use patterns. The basin experiences an arid-to-semi-arid climate according to the De Martonne and Thornthwaite classifications, with a mean annual temperature of 26.6 °C, average precipitation of 224 mm, and potential evapotranspiration of 2,668 mm. Rainfall is highly seasonal, primarily occurring from February to March (150–200 mm), contributing to a pronounced water deficit.

Geologically, the region comprises igneous intrusive rocks, metamorphic complexes, and sedimentary formations, reflecting the complex tectonic setting of the Makran zone, as shown in **Figure 1**. Field observations indicate widespread exposure of metamorphic, igneous, and sedimentary rocks across the basin. Alluvial sediments dominate the Minab Plain, especially at the river entrance, where coarse-grained sand and gravel deposits are prevalent. Particle size decreases westward near the coast, transitioning to finer silty and sandy sediments with higher salinity. Hydrologically, the Minab River originates from the Zagros Mountains and flows southeast to the Persian Gulf. Surface water is limited and variable, rendering groundwater the main source for irrigation, domestic supply, and industrial use. The most favorable groundwater occurs near the river entrance, where coarse alluvial deposits contain relatively fresh water. Westward, groundwater salinity increases due to finer sediments and lateral flow effects. Land use is dominated by irrigated agriculture, particularly in the middle and lower basin, while urban settlements, including the city of Minab, exert additional pressure on water resources. The combination of arid climate, high evapotranspiration, geological heterogeneity, and intensive groundwater use makes the Minab Basin a vulnerable socio-hydrological system, highlighting the importance of sustainable groundwater management and quality monitoring. **Figure 2** illustrates the workflow of the method applied in this study.

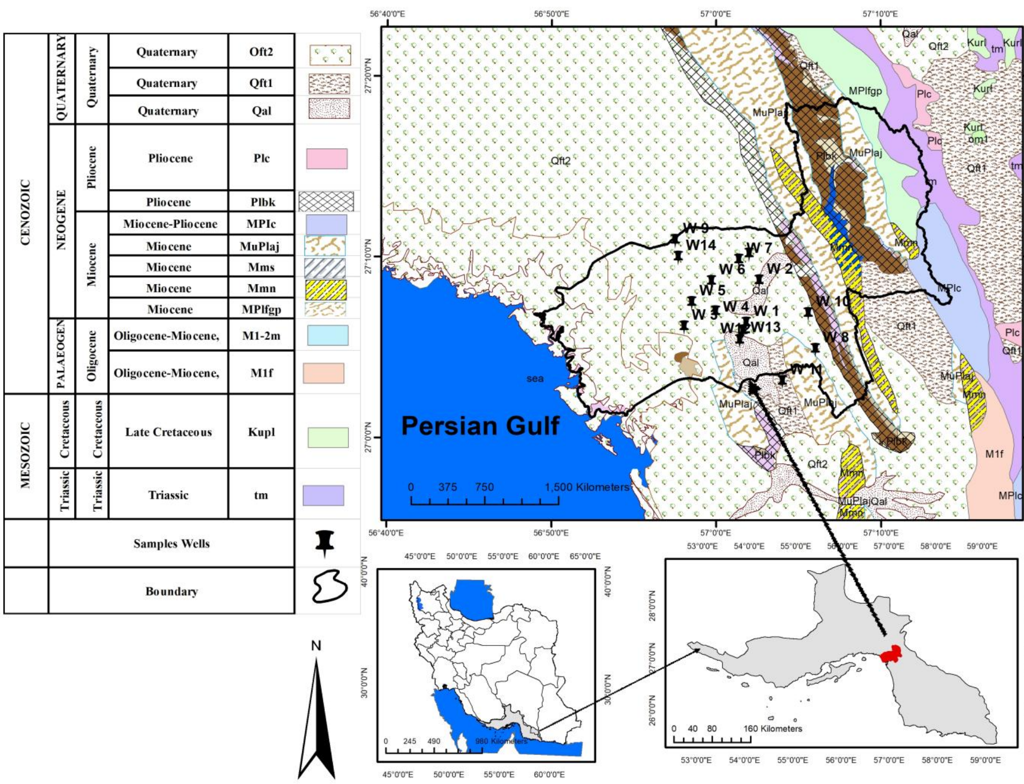


Figure 1. A View of the location of sample wells in Iran, Hormozgan, and the study area, and the geologic map.

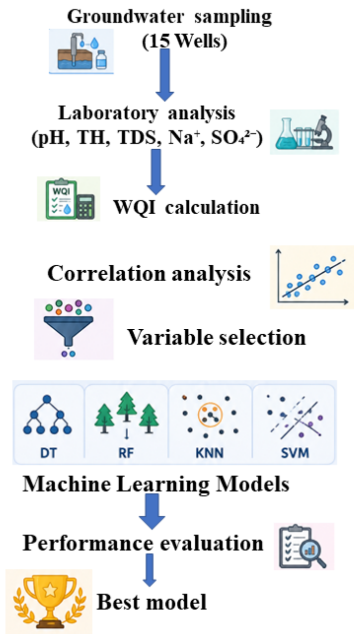


Figure 2. Workflow of the presented methodology.

2.2. Methods

2.2.1. Study Site and Sampling

The monitoring program was conducted within the agricultural areas of the Minab Plain aquifer. A total of 15 groundwater wells, distributed across different agricultural zones, were selected for the study. Groundwater samples were collected from these wells over 14 years (2010–2023), yielding 308 samples. The location and coordinates of each well were recorded using a Global Positioning System (GPS) to ensure accurate spatial representation. Sampling, transportation, and analysis were performed in accordance with standardized procedures for water and wastewater testing [31].

- **Water Quality Parameters**

Six hydrochemical parameters were analysed to determine the Groundwater Quality Index (GWQI): pH, total hardness (TH), total dissolved solids (TDS), sodium (Na^+), sulfate (SO_4^{2-}), and chloride (Cl^-). Total hardness represents the combined concentration of calcium and magnesium (mg/L) and is a key indicator of water suitability for irrigation, human consumption, and industrial use [32,33]. The pH of groundwater samples ranged from 7.9 to 8.3, indicating neutral to mildly alkaline conditions, consistent with typical natural waters [34].

Based on groundwater classification criteria [35], approximately 40% of the samples exhibited very high mineralisation (TDS > 1,500 mg/L), a condition that can be exacerbated by seawater intrusion and anthropogenic activities, such as domestic wastewater discharge [32]. Chloride and sodium are the predominant ions affecting irrigation water quality. Excessive chloride, originating from saltwater intrusion or atmospheric deposition, can act as a tracer for contamination [36,37]. High sodium concentrations reduce oil permeability, while elevated chloride levels can inhibit plant growth. Over time, the accumulation of salts from irrigation water and soil can increase soil salinity, adversely affecting crop productivity [38].

- **Geological Context**

All geological formations within the study area are classified as Quaternary deposits, including coastal sands, marine sediments, and alluvial deposits. The most favorable aquifer zones correspond to coarse-grained alluvial deposits at the mouth of the Minab River, which contain relatively fresh water, whereas finer sediments toward the west exhibit higher salinity. This comprehensive 14-year dataset enabled assessment of spatial and temporal variations in groundwater quality across the Minab Plain, forming the basis for WQI calculation and subsequent machine-learning-based predictive analyses. **Table 1** summarizes the threshold values for each water quality parameter according to international standards.

Table 1. Indicator Thresholds in Water Quality Standards.

	Units	WHO	Range of WQI	Groundwater Quality
pH	-	6.5–8.5	<50	Excellent
Na^{+}	mg/L	200	50–100	Good
SO_4^{2-}	mg/L	250	100–200	Poor
Cl^{-}	mg/L	300	200–300	Very Poor
TH	mg/L $CaCO_3$	500	>300	Unsuitable
TDS	mg/L	1,000		

2.2.2. WQI Calculation

The quality of groundwater for drinking purposes can be quantitatively assessed using the Water Quality Index (WQI), which integrates multiple hydrochemical parameters into a single, dimensionless value to facilitate [39,40]. In this study, WQI was calculated using six key parameters, each assigned a weight (W_i) reflecting its relative importance to overall water quality and potential health impacts. The calculation procedure involves three main steps. First, a sub-index (Q_i) is determined for each parameter, based on its measured concentration (V_i), the ideal or optimal value (V_0), and the standard permissible value (S_i) according to international guidelines:

$$Q_i = 100\left[\frac{V_i - V_0}{S_i - V_o}\right]$$

A proportionality constant (K) is used to calculate the weight for each parameter:

$$W_i = \frac{K}{S_i}$$

Finally, the overall WQI is computed as a weighted average of the sub-indices:

$$WQI = \frac{\sum W_i\, Q_i}{\sum W_i}$$

The resulting WQI values range from 0 to 100, providing a relative measure of water quality. Based on established thresholds, water quality can be categorised into five classes: "Excellent" (<50), "Good" (50–100), "Poor" (100–200), "Very Poor" (200–300), and "Unsuitable" (>300) [41]. This approach enables systematic ranking of wells and identification of areas where water quality may pose risks to human health, facilitating effective groundwater management and decision-making.

2.2.3. Machine Learning Algorithms for WQI Prediction

In this study, various combinations of groundwater quality parameters were employed to predict the Water Quality Index (WQI) using four widely applied machine learning (ML) models: Decision Tree (DT), k-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest (RF). Continuous variables were normalized using Min–Max scaling to improve model stability and comparability among variables with different units. The dataset was randomly divided within each cross-validation iteration into training (90%) and validation (10%) subsets. Model selection was based exclusively on the training folds, while final model performance was reported on the independent test set. Extreme values were examined using the interquartile range (IQR) method; however, no observations were removed because they represented realistic hydrochemical conditions of the aquifer. The original monitoring dataset contained no missing values because only complete laboratory records were retained for analysis. Therefore, no data imputation procedure was required.

The Decision Tree (DT) model emulates a sequential decision-making process, where water quality features are recursively partitioned to reach terminal nodes for classification. DT models are intuitive, interpretable, and allow visualisation of decision paths. However, they may be prone to overfitting on training data, potentially limiting generalisation [42,43]. The Classification and Regression Tree (CART) algorithm is widely used for its robustness and versatility [25,44], with parameters set to mtry = 5 and CP = 0.0001.

K-Nearest Neighbors (KNN) is a distance-based algorithm that classifies or predicts values by measuring similarity among neighboring data points using metrics such as Euclidean, Manhattan, or Minkowski distances. Despite its computational intensity, KNN has demonstrated superior performance compared to traditional methods [45]. The Euclidean distance is used in this study.

The Support Vector Machine (SVM) is a kernel-based model that optimizes global solutions while handling nonlinear relationships. Various kernel functions can be applied, including polynomial and radial basis function (RBF) kernels. In this study, SVM was implemented using the RBF kernel, with gamma = 0.5 and C = 1, ensuring high flexibility and regularisation [46,47].

Random Forest (RF) is an ensemble learning technique that combines multiple decision trees to improve predictive accuracy. By randomly selecting subsets of data and features for each tree, RF reduces overfitting while improving model stability. It is particularly effective in handling high-dimensional and complex datasets and allows evaluation of feature importance [48,49]. For this study, the RF model was configured with mtry = 5 and CP = 0.0001. Model hyperparameters were optimized using a grid-search procedure within the training folds of the cross-validation framework. The tested ranges included tree complexity parameters for DT and RF, the number of neighbors for KNN (k = 3–15), and the regularization (C = 0.1–10) and kernel parameters (γ = 0.01–1) for SVM.

To partially account for the temporal structure of the 2010–2023 dataset, model evaluation was repeated using a chronological hold-out test in which the most recent 20% of observations were reserved for independent testing, while the remaining data were used for model training and cross-validation.

2.2.4. Performance Evaluation

To assess the predictive capability of each model, several performance metrics were applied, including accuracy, coefficient of determination (R^2), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE) [50–52]. Accuracy, defined as the ratio of correctly predicted samples to the total number of samples, provides a general measure of classification performance [53]. RMSE and MSE quantify the deviation between predicted and observed values, while R^2 assesses the proportion of variance explained by the model. Collectively, these metrics provide a comprehensive evaluation of each ML algorithm's reliability in WQI prediction.

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{Y_i} - \mathrm{X_i})^2} \tag{1}$$

$$\mathrm{MSE} = \frac{1}{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{N}} (\mathrm{y_i} - \mathrm{x_i})^2 \tag{2}$$

$$r = \frac{n \sum_{i=1}^{n} x_i\, y_i - \sum_{i=1}^{n} y_i}{\sqrt{\left[n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right] \times \left[n \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}} \tag{3}$$

$$\mathrm{MAPE} = \frac{1}{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{N}} \left| \frac{\mathrm{y_i} - \mathrm{x_i}}{\mathrm{y_i}} \right| \tag{4}$$

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{5}$$

For a given point N in space and time, the mean of the observed (Y_i) and simulated (X_i) Values are calculated using all available data, including observations and outputs from multiple simulations. The R^2 value, as expressed in Equation (3), reflects how well the simulated values reproduce the variability of the observed data, with values closer to 1 indicating higher predictive accuracy and better model performance.

3. Result and Discussion

Groundwater quality is a key determinant of its suitability for drinking, irrigation, and industrial use. **Table 2** summarizes six hydrochemical parameters measured at 15 monitoring wells in the Minab aquifer from 2010 to 2023. These parameters were evaluated against World Health Organization (WHO) guidelines to assess water suitability [54]. Spatial distributions of groundwater quality were also analysed in relation to the basin's geological characteristics (**Figure 1**). The pH values ranged from 7.9 (Wells W12 and W15) to 8.3 (Wells W3 and W13), indicating neutral to mildly alkaline conditions. Chloride concentrations varied widely, from 170 to 2,393 mg/L, with a mean of 189 mg/L, while sulfate (SO_4^{2-}) ranged from 102 to 698.8 mg/L (mean 251 mg/L), reflecting contributions from both natural sources and anthropogenic activities [55]. Excessive chloride can impart a salty taste and indicate contamination from seawater intrusion or human activity. Total dissolved solids (TDS) in groundwater ranged from 720 to 4,889 mg/L (mean 1,614.5 mg/L), representing the sum of dissolved inorganic salts (WHO) [54]. Total hardness (TH) varied between 112 and 1,440 mg/L as $CaCO_3$ (mean 339 mg/L), while sodium concentrations ranged from 186 to 1,339 mg/L (mean 424 mg/L). Sodium levels above 200 mg/L may compromise taste and domestic usability [32,56,57]. Spatial variations in groundwater quality are closely linked to geological formations. TDS levels were generally lower in Quaternary sediments but higher in eastern wells, owing to high evaporation rates and anthropogenic activities such as intensive agriculture and water harvesting [41]. Elevated TH concentrations were observed in areas underlain by Miocene basement rocks, rich in calcite and dolomite, whereas Quaternary formations exhibited lower TH. Sodium concentrations followed a similar trend, with higher values in Miocene formations in the east. The Makran unit, composed of Miocene–Pliocene sediments including Goshi Marl, Khako Sandstone, Tiab Sandstone, and Minab Conglomerate, contributes to groundwater salinity due to high porosity, prolonged water contact, and elevated chloride content, predominantly as sodium chloride. The Minab aquifer generally receives recharge from the eastern highlands, suggesting an east-to-west flow. However, excessive groundwater extraction, particularly in central and eastern sections, has reversed the flow direction to west-to-east. The largest declines in the water table occur in the southeast, driven by intensive exploitation and overpumping, highlighting the critical need for sustainable groundwater management.

Table 2. Details of six water quality parameters at each monitoring well.

Areas Name	Parameter	Min	Max	Mean	S.D	Classification
W1	TH	320.0	381.0	351.8	17.8	Very Hard
	Na^+	443.7	540.3	475.9	27.6	Bad
	SO_4^{2-}	168.1	331.4	220.9	43.2	Good
	Cl^-	620.4	797.6	716	42.5	Bad
	pH	7.8	8.5	8.1	0.2	-
	TDS	1,657.0	1,974.4	1,780.5	101.5	Very High: Useful for irrigation
W2	TH	160.0	496	347.1	72.3	Very Hard
	Na^+	224.4	423.2	316.5	49.6	Not Bad
	SO_4^{2-}	150.8	380.4	250.2	62.1	Good
	Cl^-	202.1	443.1	330.5	56.5	Good
	pH	7.4	8.4	7.9	0.3	-
	TDS	873.6	1,652.0	1,326.3	201.2	High: Useful for irrigation
W3	TH	112	184	137	21.1	Hard
	Na^+	239.1	323.2	281	19.7	Not Bad
	SO_4^{2-}	131.1	273.3	183.4	30.5	Good
	Cl^-	226.9	265.9	246.6	10.0	Good
	pH	7.7	8.8	8.3	0.3	-
	TDS	896.6	1,043.2	958.0	38.1	Moderate: Permissible for drinking
W4	TH	144	392	220.4	67.9	Hard
	Na^+	284.2	450.1	326.6	39.4	Not Bad
	SO_4^{2-}	144.6	464	204.7	67.0	Good
	Cl^-	329.7	487.5	380.2	50.3	Not Bad
	pH	7.7	8.5	8.1	0.2	-
	TDS	1,031.0	1,751.7	1,199.9	184.7	Useful for irrigation
W5	TH	128	256	180.6	40	Hard
	Na^+	186.2	226.2	208.8	15.9	Good
	SO_4^{2-}	139.3	195	165.6	20.0	Good
	Cl^-	170.2	216.3	189.4	20.7	Good
	pH	7.8	8.5	8.1	0.2	-
	TDS	720.0	940.8	811.9	81.9	Moderate: Permissible for drinking
W6	TH	208.0	336	244.8	37.2	Hard
	Na^+	211.7	294.7	232.3	22.5	Not Bad
	SO_4^{2-}	102.8	293.9	201.2	38.9	Good
	Cl^-	191.4	322.6	216.3	30.8	Good
	pH	7.8	8.6	8.1	0.2	-
	TDS	92.0	1,228.2	926.2	206.6	Moderate: Permissible for drinking
W7	TH	176.0	584	425.3	88.5	Very Hard
	Na^+	239.3	480.3	420.9	50.7	Not Bad
	SO_4^{2-}	161.4	346.3	258.3	52.1	Good
	Cl^-	260.6	620.4	484.3	70.7	Not Bad
	pH	7.5	8.8	8.0	0.4	-
	TDS	1,012.5	2,074.2	1,718.6	221.5	Very High: Useful for irrigation

Table 2. *Cont.*

Areas Name	Parameter	Min	Max	Mean	S.D	Classification
W8	TH	144.0	352	251	61.4	Hard
	Na^+	349.7	478.2	430.5	46.6	Not Bad
	SO_4^{2-}	189.7	654.6	264.6	106.3	Good
	Cl^-	221.6	496.3	421.6	64.0	Not Bad
	pH	7.8	8.8	8.2	0.3	-
	TDS	1,340.2	1,647.4	1,517.4	91.5	Very High: Useful for irrigation
W9	TH	160.0	384	237.6	51.1	Hard
	Na^+	320.7	699.3	414.5	100.7	Not Bad
	SO_4^{2-}	184.4	372.7	246.2	42.3	Good
	Cl^-	336.8	531.8	398.6	63.5	Not Bad
	pH	7.5	8.7	8.2	0.3	-
	TDS	1,219.2	2,213.1	1,457.7	299.5	High: Useful for irrigation
W10	TH	248.0	1000	313.7	150.6	Very Hard
	Na^+	304.6	835.9	355.6	106.7	Not Bad
	SO_4^{2-}	160.9	597.0	241.0	87.3	Good
	Cl^-	280.1	1471.2	363.8	242.0	Not Bad
	pH	7.6	8.7	8.2	0.3	-
	TDS	1,194.2	3,601.3	1,395.1	485.5	High: Useful for irrigation
W11	TH	384.0	1,160	784.5	238.1	Very Hard
	Na^+	809.5	1,339.1	1,031.6	151.8	Bad
	SO_4^{2-}	286.3	658.5	482.1	102.7	Not Bad
	Cl^-	974.9	2,180.3	1,636.2	388.6	Bad
	pH	7.4	8.4	8.0	0.2	-
	TDS	2,826.2	5,052.8	3,886.0	658.5	Very High: Unfit for drinking and irrigation
W12	TH	288.0	1,440	630.0	329.9	Very Hard
	Na^+	425.3	1,263.3	787.8	262.5	Bad
	SO_4^{2-}	156.1	617.6	339.5	135.6	Not Bad
	Cl^-	567.2	2,393.0	1,217.5	550.4	Bad
	pH	7.6	8.4	7.9	0.3	-
	TDS	1,626.2	4,889.6	3,002.5	1,081.3	Very High: Unfit for drinking and irrigation
W13	TH	120.0	192	166.7	21.3	Hard
	Na^+	211.5	245.5	228.4	10.3	Good
	SO_4^{2-}	131.1	209.9	156.3	20.3	Good
	Cl^-	200.3	276.5	233.6	22.4	Good
	pH	7.7	8.7	8.3	0.2	-
	TDS	777.0	935	849.3	46.3	Moderate: Permissible for drinking
W14	TH	272.0	464	308.3	40.3	Very Hard
	Na^+	319.3	503	369.3	36.2	Not Bad
	SO_4^{2-}	147.4	454.8	244.7	66.1	Good
	Cl^-	354.5	620.4	383.9	54.2	Not Bad
	pH	7.6	8.5	8	0.3	-
	TDS	1,317.1	1,991.7	1,423.5	143.0	High: Useful for irrigation
W15	TH	288.0	640.0	488.8	77.6	Very Hard
	Na^+	367.6	591.1	480.3	65.9	Bad
	SO_4^{2-}	141.7	698.8	311.9	116.9	Not Bad
	Cl^-	354.5	584.9	477.4	73.2	Not Bad
	pH	7.3	8.8	7.9	0.4	-
	TDS	1,390.7	2,341.0	1,964.2	266.6	Very High: Useful for irrigation

The observed spatial variability reflects differences in hydrogeological conditions throughout the aquifer. Elevated salinity-related parameters (EC, TDS, Na^+, and Cl^-) are mainly associated with seawater intrusion, irrigation return flows, and intensive agricultural activities. In contrast, wells farther from the coastline exhibited lower salinity and, consequently, higher WQI values. These findings indicate that groundwater quality is primarily controlled by both natural geochemical processes and anthropogenic influences.

Figure 3 indicates a clear spatial pattern in groundwater quality across the Minab aquifer. Higher WQI values are mainly observed in inland recharge areas, whereas lower WQI values occur near the western coastal zone. This pattern is consistent with the regional hydrogeological setting, where seawater intrusion, groundwater overexploitation, and irrigation return flows contribute to increasing groundwater salinity. The spatial distribution therefore supports the hydrochemical interpretation presented in the subsequent machine learning analysis.

Table 3 presents the distribution of Water Quality Index (WQI) grades across the 15 monitoring wells in the Minab aquifer. The calculated WQI values ranged from 29.74 to 107.2, reflecting generally good groundwater quality throughout the study area. Well W15 exhibited the highest proportion of "Excellent" water quality, accounting for 40.9% of observations. In contrast, wells W1, W5, W6, and W8 did not record any instances of "Excellent" water, indicating 0% for this category. Overall, the most prevalent water quality class was "Good," with an average frequency exceeding 87% across all wells. Wells W1, W5, and W6 reported 100% of samples in the "Good" category, while six wells (W3, W4, W10, W11, W12, and W13) recorded more than 90% in this grade. Instances of "Poor" water quality were limited, observed in only six wells (W3, W8, W9, W10, W13, and W15), each exceeding 7% of total observations. Notably, none of the wells fell under the "Very Poor" category, indicating that extreme deterioration in water quality was absent in the study area. These findings underscore the overall suitability of Minab groundwater for domestic and agricultural use, while highlighting localized areas where water quality management may be required to maintain and improve standards.

(a)

(b)

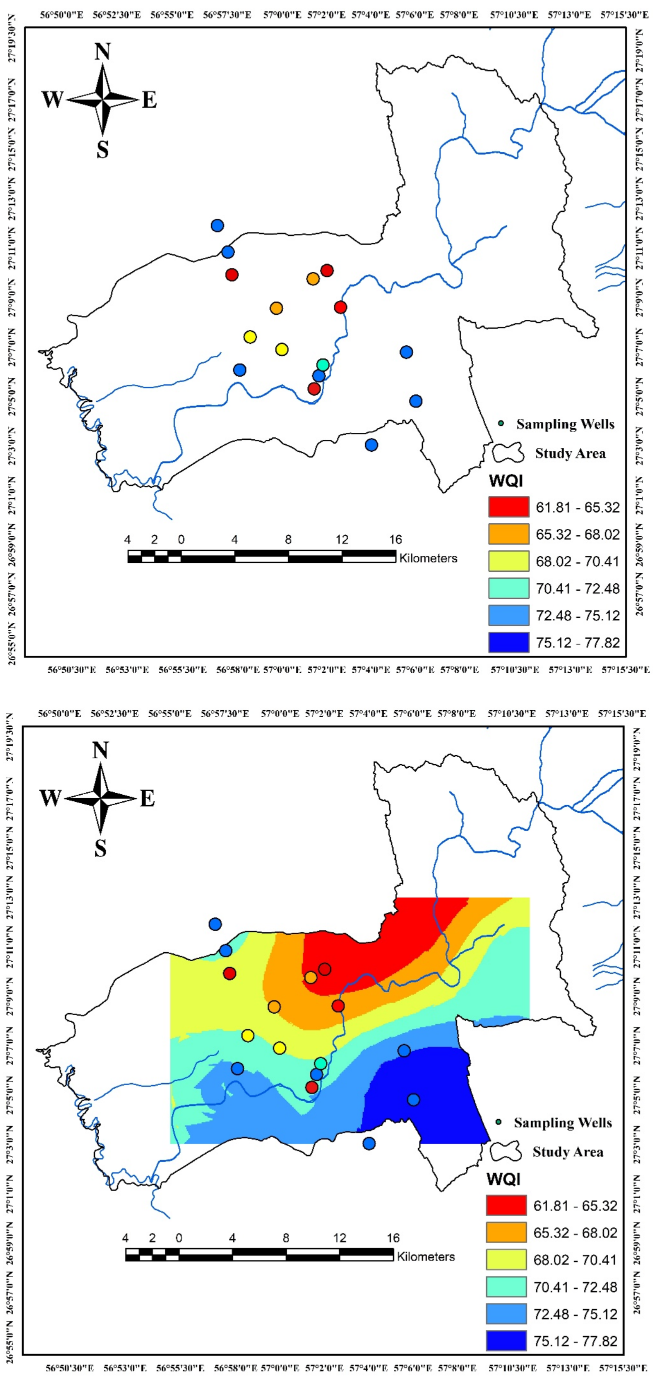


Figure 3. Spatial distribution of the Water Quality Index (WQI) across the Minab aquifer. (**a**) WQI class at the 15 monitoring wells according to the WHO classification. (**b**) Spatial interpolation of WQI values shows the gradual decline in groundwater quality toward the western coastal zone due to salinity intrusion and agricultural activities.

Table 3. Summary of the WQI classification at all monitoring wells.

No.Well	Excellent (%)	Good (%)	Poor (%)	Very Poor (%)	Count
W1	-	100	-	-	15
W2	31	68	-	-	22
W3	4.5	91	4.5	-	22
W4	4.5	95.5	-	-	22
W5	-	100	-	-	7
W6	-	100	-	-	24
W7	30	70	-	-	20
W8	-	88.8	11.2	-	18
W9	6	82	12	-	17
W10	4.4	91.2	4.4	-	23
W11	6.7	93.3	-	-	15
W12	6.3	93.7	-	-	16
W13	4.3	91.4	4.3	-	23
W14	13.6	86.4	-	-	22
W15	40.9	54.6	4.5	-	22

Because the dataset is highly imbalanced (95.6% 'Good' class), classification accuracy was interpreted together with class-wise prediction behavior rather than as the sole indicator of model quality.

3.1. Correlation Analysis of Water Quality Parameters

Based on the collected groundwater quality data, six input parameters were selected to evaluate their influence on the Water Quality Index (WQI). **Figure 4a** illustrates the correlation coefficients between each parameter and WQI across all monitoring wells. Among the variables, pH exhibited the strongest positive influence on WQI, with an average correlation coefficient of 0.98, indicating its critical role in determining groundwater quality. Total hardness (TH) and chloride (Cl^-) displayed negative correlations with WQI, whereas sulfate (SO_4^{2-}) showed a positive relationship. Sodium (Na^+) and total dissolved solids (TDS) exhibited the lowest correlations, suggesting a minimal impact on WQI predictions. These findings imply that Na^+ and TDS could be monitored less frequently without significantly affecting the accuracy of water quality assessment, thereby reducing sampling costs. The correlation patterns also varied spatially across wells. For instance, in Well W3, all parameters except TH were positively associated with WQI, whereas in Well W15, all parameters except pH were negatively associated with WQI. These spatial variations highlight the importance of site-specific assessment when evaluating groundwater quality. Consistent with previous studies, pH, along with trace elements such as manganese (Mn) and nickel (Ni), has been identified as the most influential parameter in predicting WQI [20]. Overall, the correlation analysis underscores the dominant role of pH in governing groundwater quality and provides guidance for optimizing sampling strategies and parameter selection in water quality monitoring programs.

(a)

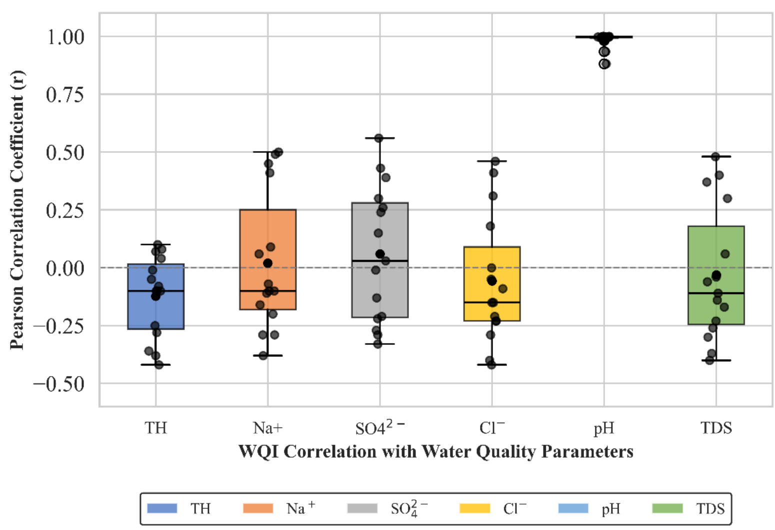


Figure 4. *Cont.*

(b)

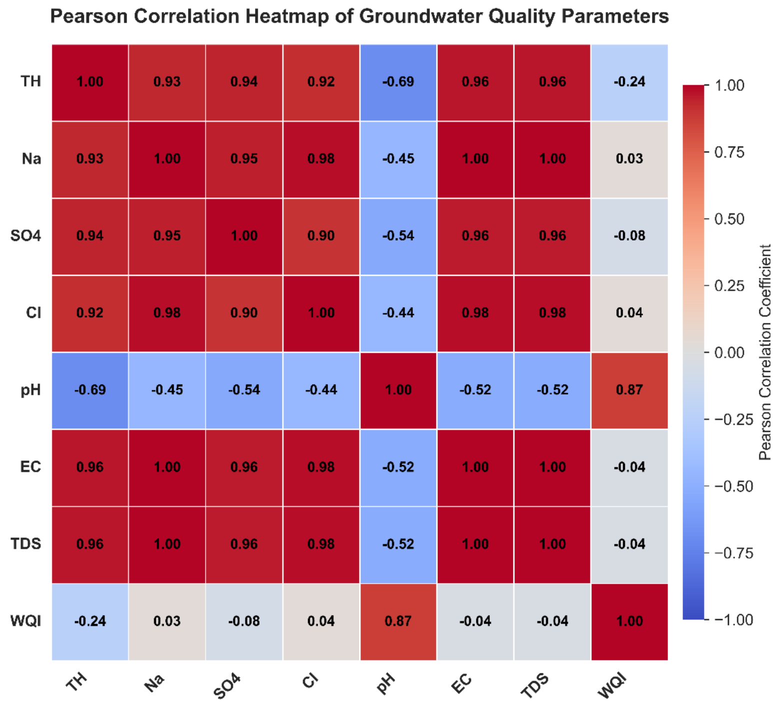


Figure 4. Pearson correlation heatmaps (XLSTAT) showing (**a**) multicollinearity among hydrochemical parameters (TH, Na^+, SO_4^{2-}, Cl^-, pH, EC, TDS) and (**b**) their correlations with WQI for feature selection. Color scale: −1.0 to 1.0.

Figure 4b illustrates the Pearson correlation matrix among the hydrochemical parameters and WQI. Strong positive correlations were observed between pH and WQI, whereas TH, chloride, and sodium exhibited relatively strong negative correlations. These results indicate that only a limited number of hydrochemical parameters explain most of the variability in groundwater quality. Therefore, correlation analysis was used as the initial criterion for selecting variables for subsequent machine learning scenarios. Although pH exhibited the highest individual correlation with WQI, the progressive feature-addition scenarios were designed according to the study's predefined analytical framework rather than solely on pairwise correlation coefficients. Therefore, the order of variable inclusion should not be interpreted as a ranking of variable importance but as a practical evaluation of cumulative predictive performance under different monitoring scenarios.

Accordingly, six groups of input water quality parameters were determined and defined for each of the six scenarios, based on existing data from the study area, as shown in **Table 4**.

Table 4. Various combinations of input variables to predict WQI.

Scenario	Input Variables	Objective
S1	TH	Evaluate the predictive ability of the single most influential variable
S2	TH, Na^+	Assess the effect of adding the second-ranked variable
S3	TH, Na^+, SO_4^{2-}	Progressive feature addition
S4	TH, Na^+, SO_4^{2-}, Cl^-	Evaluate the cumulative contribution of major hydrochemical variables
S5	TH, Na^+, SO_4^{2-}, Cl^-, pH	Identify the minimum optimal predictor set
S6	TH, Na^+, SO_4^{2-}, Cl^-, pH, TDS	Reference model using the complete reduced dataset

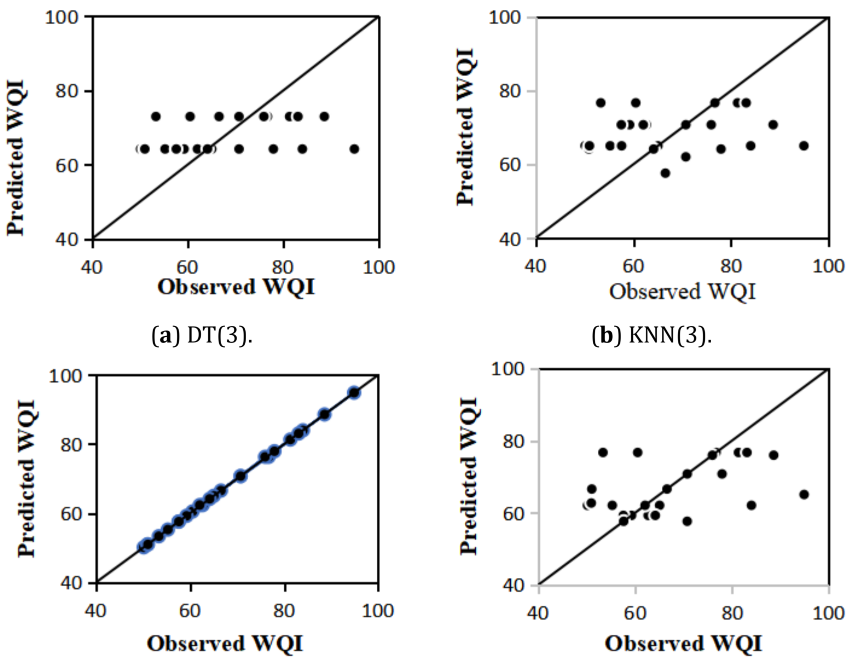
The results indicate that the predictive performance improved as additional hydrochemical variables were incorporated. However, the improvement became marginal after the inclusion of five variables (TH, Na^+, SO_4^{2-}, Cl^-, and pH), suggesting that these parameters contain most of the information required for reliable WQI estimation. Therefore, increasing the number of input variables beyond this subset provides limited additional predictive benefit.

Compared with the full six-parameter dataset, the five-variable model eliminates one laboratory analysis while maintaining nearly identical predictive performance. Although the reduction appears modest, TDS measurement represents one of the routine laboratory analyses in groundwater monitoring. Therefore, eliminating this parameter can reduce analytical workload and monitoring cost, particularly in long-term monitoring programs involving large numbers of sampling locations.

All machine learning (ML) models were validated using 10-fold cross-validation to prevent data overlap between the training and test sets and to ensure reliable, representative results [20]. Performance metrics calculated for both phases included Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and coefficient of determination (R^2). In the Decision Tree (DT) model, increasing the number of input water quality parameters improved model performance. MSE, RMSE, and MAPE decreased, while R^2 increased from 0.34 to 0.96 during training and from 0.41 to 0.96 during testing, indicating enhanced predictive capability. Notably, scenarios 5 and 6, covering 93% of wells, achieved the lowest error metrics and the highest R^2. For K-Nearest Neighbors (KNN), increasing the number of input features did not significantly improve performance. R^2 values ranged from 0.27 to 0.29 in training and from 0.28 to 0.32 in testing, with the best performance observed in scenarios 1–3. Scenarios 5 and 6 showed limited improvement, suggesting that KNN is less effective in capturing complex relationships among WQI predictors in this study. The Support Vector Machine (SVM) model exhibited marked improvement with additional parameters. When six water quality variables were employed, MSE decreased to below 2.8 (training) and 1.4 (testing), while RMSE decreased to 3.8–1.55 and 2.8–0.89, respectively. R^2 exceeded 0.98 for both training and testing phases, demonstrating high accuracy in WQI prediction.

Random Forest (RF) performed moderately well, with predictions less accurate than those of DT and SVM, particularly at lower WQI values. Although RF captured general trends, scatter around the reference lines was larger, reflecting moderate predictive precision.

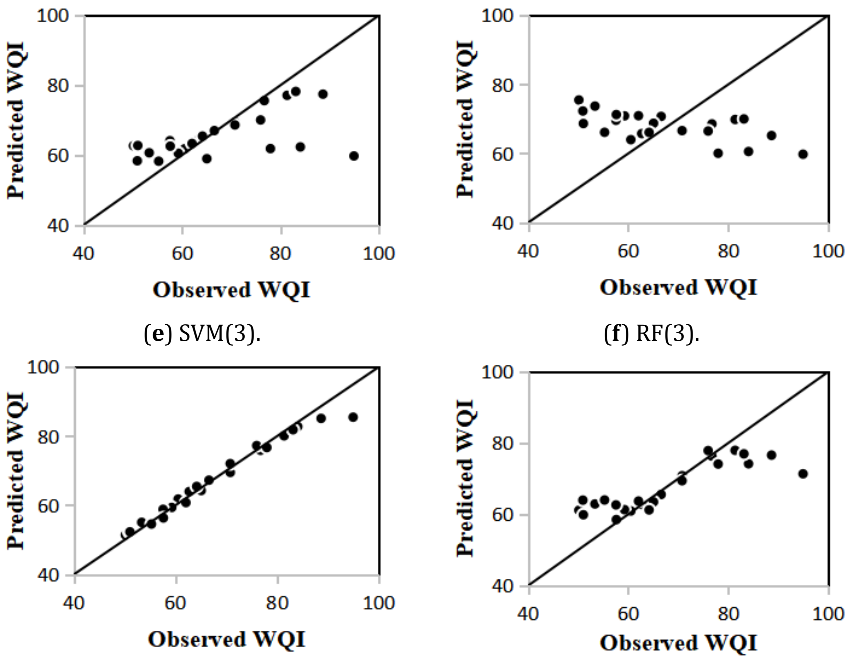
Overall, DT and SVM models, especially when using five to six input parameters (TH, Na^+, SO_4^{2-}, Cl^-, and pH), outperformed KNN and RF. According to Grassi et al. [58] models with $R^2 > 0.9$ for both training and testing phases are considered satisfactory; this criterion was met only by DT and SVM. **Figure 5** illustrates the comparison of actual versus predicted WQI values, showing that DT(6) and SVM(6) closely align with the ideal 1:1 reference line, confirming their superior predictive capability. Increasing the number of input variables consistently improved model performance, highlighting the importance of multivariate data integration for accurate WQI prediction [59,60].



(**a**) DT(3). (**b**) KNN(3).

(**c**) DT(6). (**d**) KNN(6).

Figure 5. *Cont.*



(**e**) SVM(3). (**f**) RF(3).

(**g**) SVM(6). (**h**) RF(6).

Figure 5. Comparisons between actual and predicted WQI from three and six water quality parameters in various Machine Learning models.

The superior performance of DT and SVM can be attributed to their ability to capture nonlinear relationships between hydrochemical variables and WQI. DT effectively partitions the feature space using thresholds on influential parameters such as pH and chloride, whereas SVM with the RBF kernel models complex nonlinear boundaries in the predictor space. In contrast, KNN is sensitive to local data density and distance metrics, and RF may be affected by the relatively small sample size available in this study.

One of the most important findings of this study is that reliable WQI prediction does not necessarily require the complete set of measured hydrochemical parameters. The reduced-input scenarios demonstrated that only five variables were sufficient to achieve prediction accuracy comparable to that obtained using the larger dataset. This finding has direct implications for groundwater monitoring programs where laboratory analyses represent a major component of operational costs. The reduced-variable framework may substantially decrease the analytical burden associated with routine groundwater monitoring. Because fewer laboratory measurements are required, the proposed methodology can reduce monitoring costs, shorten laboratory processing time, and facilitate more frequent groundwater quality assessments without significant loss of prediction accuracy. In addition to continuous WQI values, categorical WQI grades "Excellent," "Good," "Poor," "Very Poor," and "Unsuitable" provide a widely used framework for assessing groundwater quality. In the Minab dataset, the majority of records (95.6%) were classified as "Good," while "Excellent" and "Poor" grades represented 4.3% and 0.14%, respectively. No samples were categorized as "Very Poor" or "Unsuitable." **Table 5** summarizes the prediction accuracy of each machine learning model across various scenarios. When using a single input variable (TH), model accuracy ranged from 78% to 84%. In general, iterative inclusion of additional water quality parameters improved model performance. For Decision Tree (DT) and Support Vector Machine (SVM) models, accuracy increased from an average of 85% with two variables to approximately 98% with five to six variables. Specifically, DT achieved a maximum accuracy of 98.2%, while SVM reached 96.5% with five parameters, demonstrating that multivariate inputs significantly enhance predictive performance. The k-Nearest Neighbors (KNN) model achieved its highest accuracy with four and five variables (84% and 84.8%, respectively), after which additional parameters did not improve predictions, indicating a plateau effect. Similarly, Random Forest (RF) exhibited a moderate improvement as the number of parameters increased, achieving above 86% accuracy with five to six input variables, highlighting the importance of including key parameters to optimize model performance. Following the recommendations by Freeman et al. [61], prediction accuracies above 0.85 are considered satisfactory for WQI grade assessment. Overall, DT and SVM models consistently outperformed KNN and RF in predicting WQI grades, corroborating previous findings that decision-tree-based approaches are highly effective for water-quality prediction [62].

Table 5. Accuracy of water quality grades based on predicted WQI values from various statistical and machine-learning models across nine groups of variables.

Actual WQI		Excellent		Good		Poor		Accuracy %
		NO	Value	NO	Value	NO	Value	
Model	**Scen**	**30**	**43.9**	**250**	**72.98**	**10**	**106.2**	
DT	1	8	43.93	281	71.85	0	0	84.48
	2	7	42.95	282	72.10	0	0	85.90
	3	19	49.55	270	72.37	0	0	85.94
	4	10	39.70	279	71.88	0	0	86.85
	5	24	42.03	260	72.90	4	103.47	98.23
	6	26	42.80	258	72.98	4	106.77	96.94
KNN	1	11	45.79	277	71.05	0	0	83.74
	2	11	45.05	276	70.94	0	0	83.65
	3	11	44.85	276	71.15	0	0	83.76
	4	13	45.89	274	71.16	0	0	84.07
	5	13	45.89	274	71.12	1	105.1	84.88
	6	16	47.32	273	70.71	1	102.4	83.39
SVM	1	3	47.45	286	69.97	0	0	83.35
	2	3	46.85	286	70.32	0	0	86.02
	3	10	48.03	279	70.87	0	0	87.77
	4	11	47.71	278	70.82	0	0	88.94
	5	24	44.40	265	72.08	0	0	96.49
	6	25	44.68	264	72.05	0	0	96.37
RF	1	11	43.03	278	72.14	0	0	78.19
	2	8	46.53	281	70.97	0	0	78.37
	3	4	45.15	285	70.74	0	0	78.71
	4	6	46.05	283	70.98	0	0	79.66
	5	11	47.40	278	71.17	0	0	86.12
	6	10	46.90	279	71.31	0	0	87.68
Sum		295		6,622		10		

Given the strong dominance of the "Good" class, classification accuracy should be interpreted with caution and alongside regression metrics (R^2, RMSE, and MAPE), which provide a more comprehensive evaluation of model performance.

The graphical summaries presented in **Figure 6** corroborate the quantitative findings in **Supplementary Materials Table S1**. The heatmaps (**Figure 6a,b**) clearly reveal that although DT and SVM exhibited comparable performance in the first four scenarios (SE1–SE4), a marked divergence occurred in SE5 and SE6, with SVM achieving R^2 values exceeding 0.96 and RMSE values below 1.0—particularly for the test datasets. In contrast, KNN and RF exhibited persistently poor generalization, with R^2 values remaining below 0.3 across all scenarios, regardless of the input feature combinations. The grouped bar charts (**Figure 6c**) further emphasize the superior predictive accuracy of SVM, which significantly outperformed all other models in both error reduction and explained variance. The slope chart (**Figure 6d**) illustrates a clear upward trend in predictive performance for DT and SVM from SE1 to SE6, whereas the performance trajectories for KNN and RF remained relatively flat, indicating their insensitivity to scenario-specific variations. Finally, the residual plot for the SVM model (**Figure 6e**) demonstrates a random, homoscedastic distribution of errors around zero, with no discernible pattern, thereby confirming the absence of systematic bias and the robustness of the SVM predictions under the most complex scenarios. The present findings are consistent with those reported by Haghiabi et al. [63], who also demonstrated the superior predictive capability of SVM for groundwater quality variables. Similarly, Chen et al. [64] observed that tree-based algorithms generally outperform Chen and other machine learning approaches in nonlinear environmental datasets. Similarly, Yongo et al. [65] selected three key parameters from 8 water quality indicators using stepwise multiple linear regression to predict WQI, achieving an R^2 of 0.783.

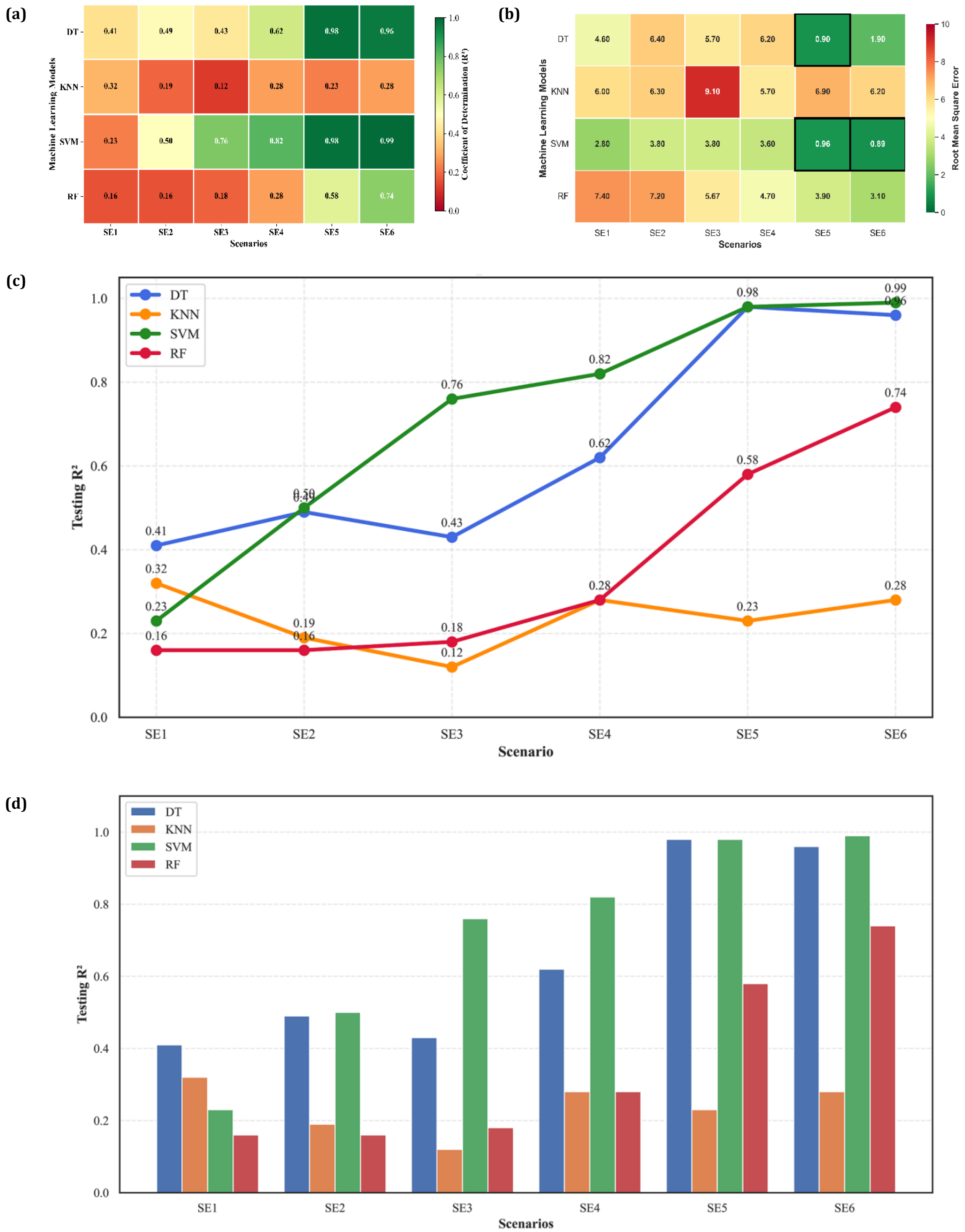


Figure 6. *Cont.*

(e)

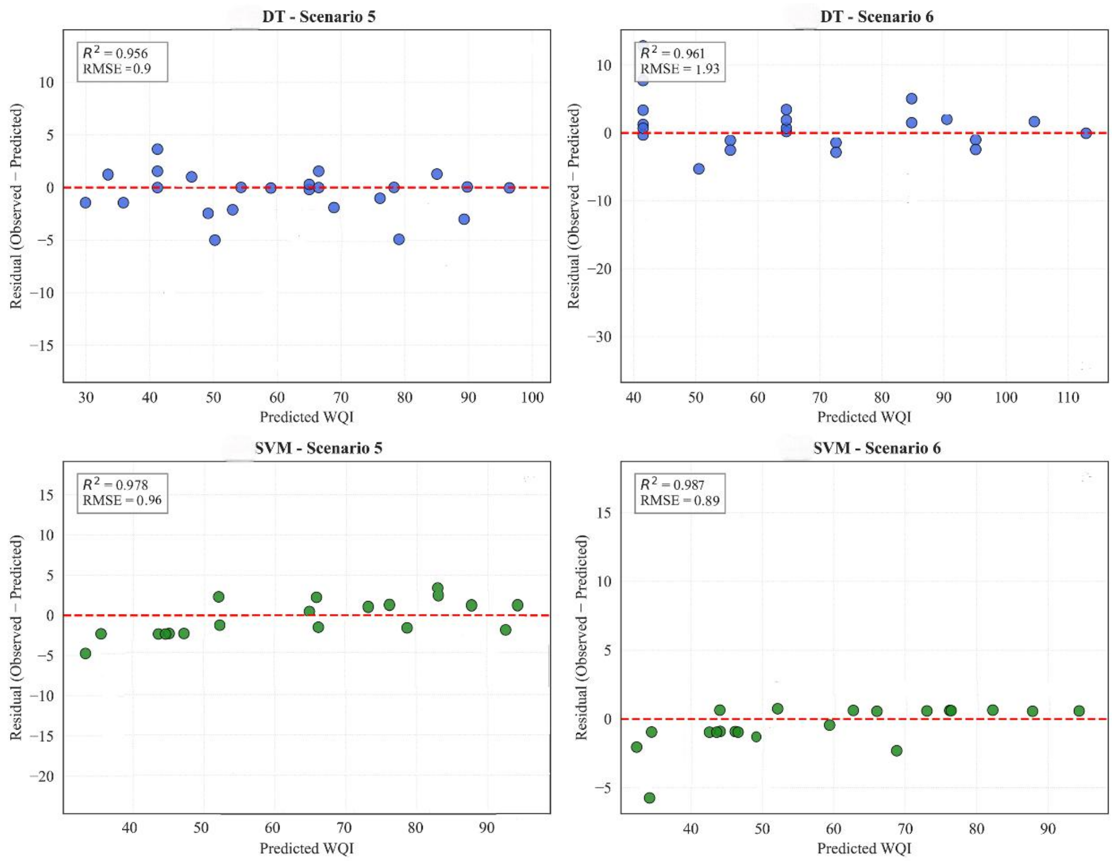


Figure 6. Graphical synthesis of the statistical performance of the evaluated models across six observation-well scenarios. (**a**) Heatmap of the coefficient of determination (R^2). (**b**) Heatmap of the root-mean-squared error (RMSE). (**c**) Grouped bar chart comparing R^2 and RMSE values for the testing dataset across all models and scenarios. (**d**) Slope chart illustrating the relative changes in model performance from SE1 to SE6. (**e**) Residual plot for the best-performing models, indicating error distribution and model bias.

However, unlike the study of Xue et al. [66], which required six variables to achieve satisfactory prediction accuracy, the present study suggests that comparable performance can be achieved using only five hydrochemical parameters. This difference is likely related to the hydrogeochemical characteristics of the Minab coastal aquifer and the variable selection strategy adopted in this study.

The observed superiority of DT and SVM also reflects the relatively nonlinear relationships among groundwater quality variables in coastal aquifers. A Decision Tree partitions the predictor space using threshold values, whereas an SVM optimizes nonlinear decision boundaries via kernel transformations. In contrast, instance-based KNN relies heavily on local neighborhood similarity, which becomes less effective when hydrochemical observations overlap substantially.

Overall, although DT and SVM achieved high predictive performance, the similarity between training and testing statistics should be interpreted cautiously because no external dataset was available for validation. Therefore, the reported performance reflects internal model validation only and should not be considered evidence of universal model generalisation. Consequently, the reported performance should be interpreted as evidence of internal predictive consistency rather than proof of external model transferability.

The superior performance of the DT and SVM models can be attributed to their ability to capture nonlinear relationships between hydrochemical variables and WQI. Decision Tree recursively partitions the predictor space based on threshold values of influential parameters, enabling effective representation of complex interactions among groundwater quality variables. Likewise, SVM employs nonlinear kernel functions that can model complex decision boundaries even when the relationships between predictors and WQI are nonlinear. These characteristics explain why both algorithms consistently outperformed KNN and RF across the evaluated scenarios.

In contrast, the relatively lower performance of KNN may be attributed to its reliance on local neighborhood structure and distance metrics, which make it sensitive to sample distribution and noise. Similarly, although RF generally performs well on large, heterogeneous datasets, its performance in this study was likely constrained by the relatively small sample size and limited variability among the predictor variables. Consequently, the additional ensemble complexity of RF did not provide a substantial improvement over the simpler DT model.

The identification of TH, Na^+, SO_4^{2-}, Cl^-, and pH as the optimal predictor set is also hydrogeochemically meaningful. Chloride and sodium are widely recognized as indicators of salinity intrusion and groundwater mineralization, particularly in coastal aquifers. Total hardness reflects carbonate dissolution and water-rock interactions, while sulfate is associated with mineral weathering and agricultural inputs. The inclusion of pH indicates the importance of groundwater buffering processes in determining the overall WQI. Therefore, the selected variables are not only statistically significant but also consistent with the dominant hydrogeochemical processes controlling groundwater quality in the Minab aquifer.

The results further demonstrate that reducing the number of input variables from six to five had only a minor influence on predictive performance. This finding suggests that a limited number of carefully selected hydrochemical indicators can adequately represent overall groundwater quality. Such variable reduction is particularly valuable for long-term monitoring programs where analytical costs and laboratory resources are limited.

From a practical perspective, the proposed framework offers an efficient approach to groundwater quality monitoring in coastal aquifers. By identifying the minimum number of laboratory measurements required, the methodology can reduce monitoring costs while maintaining reliable WQI estimation. This may help water authorities optimize sampling strategies and prioritize monitoring resources in regions experiencing increasing groundwater stress.

Despite the encouraging results, several limitations should be acknowledged. The models were developed using data from a single coastal aquifer, and their applicability to other hydrogeological settings remains to be verified. In addition, the relatively limited number of monitoring wells may have influenced model generalization. Future research should evaluate the proposed methodology using larger independent datasets and investigate explainable artificial intelligence techniques to improve model interpretability.

3.2. Practical Implications

The results demonstrate that reliable WQI prediction can be achieved using only five hydrochemical variables (TH, Na^+, SO_4^{2-}, Cl^-, and pH). This finding has practical significance because routine groundwater monitoring programs often operate under budget and laboratory constraints. Reducing the number of required analyses can substantially decrease monitoring costs while maintaining acceptable prediction accuracy, thereby improving the efficiency of long-term groundwater quality assessment.

3.3. Limitations

This study has several limitations. First, the machine learning models were developed using data from only one coastal aquifer, which may limit their generalizability to other hydrogeological settings. Second, although the monitoring period covered 2010–2023, the number of monitoring wells was relatively limited. Third, external validation on an independent dataset was unavailable. Future studies should evaluate the proposed framework on larger datasets from diverse climatic and geological regions and investigate explainable artificial intelligence techniques to improve model interpretability.

4. Conclusions

This study evaluated the applicability of four machine learning algorithms (DT, KNN, SVM, and RF) for predicting the Water Quality Index (WQI) using long-term groundwater quality data collected from the Minab coastal aquifer between 2010 and 2023. Unlike conventional studies that primarily compare machine learning algorithms, the principal objective of this research was to determine whether reliable WQI prediction could be achieved with fewer hydrochemical variables.

The hydrochemical assessment showed that groundwater quality was generally classified as Good, although considerable spatial variability was observed across the aquifer. Lower groundwater quality was mainly associated with coastal areas affected by seawater intrusion, groundwater salinization, and intensive agricultural activities, whereas inland recharge areas generally exhibited better groundwater quality.

Among the evaluated machine learning models, Decision Tree and Support Vector Machine consistently produced the highest predictive performance, with R^2 values exceeding 0.95 and substantially lower prediction errors than KNN and RF. Their superior performance is attributed to their ability to represent nonlinear relationships between groundwater quality variables and WQI.

An important outcome of this study is that reliable WQI prediction was achieved using only five hydrochemical variables (TH, Na^+, $SO_4{}^{2-}$, Cl^-, and pH). The results indicate that excluding one routinely measured parameter resulted in only a negligible reduction in predictive performance while potentially reducing laboratory analyses and monitoring effort. Consequently, the proposed framework may support more cost-effective groundwater quality monitoring programs, particularly where analytical resources are limited.

Nevertheless, several limitations should be acknowledged. The models were developed using observations from a single coastal aquifer, and no independent external dataset was available for validation. Therefore, the reported predictive performance reflects internal model validation and should not be interpreted as evidence of universal applicability.

Future studies should validate the proposed framework using independent datasets from diverse hydrogeological settings, investigate explainable artificial intelligence techniques to improve model interpretability, and evaluate the transferability of reduced-input machine learning models across diverse climatic and geological conditions.

Finally, it should be emphasised that the proposed machine learning framework is not intended to replace the deterministic WQI calculation procedure. Rather, its primary contribution is identifying the minimum hydrochemical dataset needed to estimate WQI with high reliability, thereby improving the efficiency of routine groundwater quality monitoring and reducing analytical requirements.

Supplementary Materials

The supporting information can be downloaded at https://ojs.ukscip.com/uploads/2/asset-file/2026/09/asset-file_6f00d5f01e98d9a32f6c10c6489d45a2.docx.

Author Contributions

Conceptualization, F.M. and M.H.; methodology, F.M.; software, M.H.; validation, F.M. and M.H.; formal analysis, F.M.; investigation, F.M.; resources, M.H.; data curation, M.H.; writing—original draft preparation, F.M.; writing—review and editing, M.H.; visualization, M.H.; supervision, F.M.; project administration, F.M. Both authors have read and agreed to the published version of the manuscript.

Funding

This work received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data that support the findings of this study will be available from the corresponding author upon reasonable request.

Acknowledgments

The authors gratefully acknowledge the Hormozgan Regional Water Company for providing the data and technical information used in this study.

Conflicts of Interest

The authors declare no conflict of interest.

AI Use Statement

During the preparation of this work, the authors used [AI-assisted language editing software] to improve the English expression. The authors did not use any AI tools for the research methodology, model implementation, or data interpretation. The authors subsequently reviewed and edited the content as necessary and took full responsibility for the final content of the published article.

References

1. Yang, Y.; Yuan, Y.; Xiong, G.; et al. Patterns of Nitrate Load Variability under Surface Water-Groundwater Interactions in Agriculturally Intensive Valley Watersheds. *Water Res.* **2024**, *267*, 122474. [CrossRef]
2. Foster, S.S.D.; Chilton, P.J. Downstream of Downtown: Urban Wastewater as Groundwater Recharge. *Hydrogeol. J.* **2004**, *12*, 115–120. [CrossRef]
3. Jasechko, S.; Seybold, H.; Perrone, D.; et al. Rapid Groundwater Decline and Some Cases of Recovery in Aquifers Globally. *Nature* **2024**, *625*, 715–721. [CrossRef]
4. Ravindiran, G.; Rajamanickam, S.; Sivarethinamohan, S.; et al. A Review of the Status, Effects, Prevention, and Remediation of Groundwater Contamination for Sustainable Environment. *Water* **2023**, *15*, 3662. [CrossRef]
5. Nadjai, S.; Bouderbala, A.; Khammar, H.; et al. Assessment of Groundwater Suitability for Drinking and Irrigation Purposes in the Middle Cheliff Aquifer, Algeria. *Desalination Water Treat.* **2024**, *319*, 100528. [CrossRef]
6. Raheja, H.; Goel, A.; Pal, M. A Holistic Assessment of Groundwater Quality for Drinking and Irrigation Purposes. *Water Pract. Technol.* **2023**, *19*, 134–153. [CrossRef]
7. Biswas, A.; Tortajada, C. Water Quality Management: A Globally Neglected Issue. *Int. J. Water Resour. Dev.* **2019**, *35*, 913–916. [CrossRef]
8. Gleeson, T.; Wada, Y.; Bierkens, M.F.P.; et al. Water Balance of Global Aquifers Revealed by Groundwater Footprint. *Nature* **2012**, *488*, 197–200. [CrossRef]
9. Lall, U.; Josset, L.; Russo, T. A Snapshot of the World's Groundwater Challenges. *Annu. Rev. Environ. Resour.* **2020**, *45*, 171–194. [CrossRef]
10. Basuki, T.M.; Pramono, I.B. Peak Flood Volume and Its Suspended Sediment at Various Rainfall in Kedungbulus Catchment in Gombong, Central Java, Indonesia. *J. Degrad. Min. Lands Manag.* **2021**, *9*, 3211–3217.
11. Sahour, S.; Khanbeyki, M.; Gholami, V.; et al. Evaluation of Machine Learning Algorithms for Groundwater Quality Modeling. *Environ. Sci. Pollut. Res.* **2023**, *30*, 46004–46021. [CrossRef]
12. Torres-Martínez, J.A.; Mahlknecht, J.; Kumar, M.; et al. Advancing Groundwater Quality Predictions: Machine Learning Challenges and Solutions. *Sci. Total Environ.* **2024**, *949*, 174973. [CrossRef]
13. Jibrin, A.; Al-Suwaiyan, M.; Aldrees, A.; et al. Machine Learning Predictive Insight of Water Pollution and Groundwater Quality in the Eastern Province of Saudi Arabia. *Sci. Rep.* **2024**, *14*, 20031. [CrossRef]
14. Ghahramani, Z. Probabilistic Machine Learning and Artificial Intelligence. *Nature* **2015**, *521*, 452–459. [CrossRef]
15. Chen, K.; Chen, H.; Zhou, C.; et al. Comparative Analysis of Surface Water Quality Prediction Performance and Identification of Key Water Parameters Using Different Machine Learning Models Based on Big Data. *Water Res.* **2020**, *171*, 115454. [CrossRef]
16. Knipe, D.; Padmanathan, P.; Newton-Howes, G.; et al. Suicide and Self-Harm. *Lancet* **2022**, *399*, 1903–1916. [CrossRef]
17. Campos Zabala, F.J. Supervised and Unsupervised Learning. In *Grow Your Business with AI*; Apress: Berkeley, CA, USA, 2023; pp. 217–244. [CrossRef]
18. Landau-Donnelly, F.; Roskamm, N.; Pohl, L. *[Un]Grounding-Post-Foundational Geographies*; Transcript Verlag: Bielefeld, Germany, 2022.
19. Choubin, B.; Rahmati, O. Groundwater Potential Mapping Using Hybridization of Simulated Annealing and Random Forest. In *Water Eng. Model. Math. Tools*; pp. 391–403. [CrossRef]
20. Zheng, H.; Hou, S.; Liu, J.; et al. Advanced Machine Learning and Water Quality Index (WQI) Assessment: Evaluating Groundwater Quality at the Yopurga Landfill. *Water* **2024**, *16*, 1666. [CrossRef]
21. Torkashvand, M.; Neshat, A.; Javadi, S.; et al. New Hybrid Evolutionary Algorithm for Optimizing Index-Based Groundwater Vulnerability Assessment Method. *J. Hydrol.* **2021**, *598*, 126446. [CrossRef]

22. Sarkar, S.; Adalder, A.; Paul, S.; et al. Iron Phthalocyanine Hollow Architecture Enabled Ammonia Production via Nitrate Reduction to Achieve 100% Faradaic Efficiency. *Appl. Catal. B* **2024**, *343*, 123580. [CrossRef]
23. Mohseni, U.; Pande, C.B.; Chandra Pal, S.; et al. Prediction of Weighted Arithmetic Water Quality Index for Urban Water Quality Using Ensemble Machine Learning Model. *Chemosphere* **2024**, *352*, 141393. [CrossRef]
24. Sangwan, V.; Bhardwaj, R. Machine Learning Framework for Predicting Water Quality Classification. *Water Pract. Technol.* **2024**, *19*, 4499–4521. [CrossRef]
25. Wu, Z.; Wu, Y.; Yu, Y.; et al. Assessment of Groundwater Quality Variation Characteristics and Influencing Factors in an Intensified Agricultural Area: An Integrated Hydrochemical and Machine Learning Approach. *J. Environ. Manage.* **2024**, *371*, 123233. [CrossRef]
26. Yadav, A.; Raj, A.; Yadav, B. Enhancing Local-Scale Groundwater Quality Predictions Using Advanced Machine Learning Approaches. *J. Environ. Manage.* **2024**, *370*, 122903. [CrossRef]
27. Xu, J.; Mo, Y.; Zhu, S.; et al. Assessing and Predicting Water Quality Index with Key Water Parameters by Machine Learning Models in Coastal Cities, China. *Heliyon* **2024**, *10*, e33695. [CrossRef]
28. Amiraslani, F.; Dragovich, D. Combating Desertification in Iran over the Last 50 Years: An Overview of Changing Approaches. *J. Environ. Manag.* **2011**, *92*, 1–13. [CrossRef]
29. Hosseinzadeh, M.M.; Nohegar, A. Studying the Effect of Drought on Water Resources over Two Decades and Occurrences of the Sinking Phenomenon in Minab Plain. *Environ. Sci.* **2011**, *9*.
30. Bouselsal, B.; Saibi, H. Evaluation of Groundwater Quality and Hydrochemical Characteristics in the Shallow Aquifer of El-Oued Region (Algerian Sahara). *Groundw. Sustain. Dev.* **2022**, *17*, 100747. [CrossRef]
31. APHA. *Standard Methods for the Examination of Water and Wastewater*; American Public Health Association: Washington, DC, USA, 2012.
32. World Health Organization (WHO). *Guidelines for Drinking-Water Quality: Fourth Edition Incorporating the First and Second Addenda*; World Health Organization: Geneva, Switzerland, 2022.
33. Rosborg, I.; Soni, V.; Kozisek, F. Potentially Toxic Elements in Drinking Water in Alphabetic Order. In *Drinking Water Minerals and Mineral Balance*; Springer: Cham, Switzerland, 2015; pp. 79–101. [CrossRef]
34. Sellamuthu, K.M.; Malathi, P.; Kumaraperumal, R. Assessment of Ground Water Quality for Irrigation in the Farms of Horticultural College and Research Institute, Periyakulam, Theni District. *Madras Agric. J.* **2021**, *108*, 213–219. [CrossRef]
35. van der Aa, M. Classification of Mineral Water Types and Comparison with Drinking Water Standards. *Environ. Geol.* **2003**, *44*, 554–563. [CrossRef]
36. Boateng, T.K.; Opoku, F.; Acquaah, S.O.; et al. Groundwater Quality Assessment Using Statistical Approach and Water Quality Index in Ejisu-Juaben Municipality, Ghana. *Environ. Earth Sci.* **2016**, *75*, 489. [CrossRef]
37. Li, C.; Gao, X.; Li, S.; et al. A Review of the Distribution, Sources, Genesis, and Environmental Concerns of Salinity in Groundwater. *Environ. Sci. Pollut. Res.* **2020**, *27*, 41157–41174. [CrossRef]
38. Şener, E.; Şener, Ş.; Varol, S. Evaluation of Irrigation Water Quality Using GIS-Based Analytic Hierarchy Process (AHP) in Kızılırmak Delta (Turkey). *Arab. J. Geosci.* **2022**, *15*, 678. [CrossRef]
39. Adimalla, N.; Li, P. Occurrence, Health Risks, and Geochemical Mechanisms of Fluoride and Nitrate in Groundwater of the Rock-Dominant Semi-Arid Region, Telangana State, India. *Hum. Ecol. Risk Assess.* **2019**, *25*, 81–103. [CrossRef]
40. Bouaissa, M.; El Khadir, G.; Ghalit, M.; et al. Identifying the Origin of Groundwater Salinization in the Bokoya Massif (Central Rif, Northern Morocco) Using Hydrogeochemical and Isotopic Tools. *Groundw. Sustain. Dev.* **2021**, *14*, 100646. [CrossRef]
41. Shuaibu, A.; Kalin, R.M.; Phoenix, V.; et al. Hydrogeochemistry and Water Quality Index for Groundwater Sustainability in the Komadugu-Yobe Basin, Sahel Region. *Water* **2024**, *16*, 601. [CrossRef]
42. Pradhan, B. A Comparative Study on the Predictive Ability of the Decision Tree, Support Vector Machine and Neuro-Fuzzy Models in Landslide Susceptibility Mapping Using GIS. *Comput. Geosci.* **2013**, *51*, 350–365. [CrossRef]
43. Radhakrishnan, N.; Pillai, A. Comparison of Water Quality Classification Models Using Machine Learning. In Proceedings of the 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 10–12 June 2020; pp. 1183–1188. [CrossRef]
44. Jena, M.; Dehuri, S. DecisionTree for Classification and Regression: A State-of-the Art Review. *Informatica* **2020**, *44*. [CrossRef]
45. Zhang, M.L.; Zhou, Z.H. ML-KNN: A Lazy Learning Approach to Multi-Label Learning. *Pattern Recognit.* **2007**, *40*, 2038–2048. [CrossRef]
46. Shen, J.; Qin, Q.; Wang, Y.; et al. A Data-Driven Modeling Approach for Simulating Algal Blooms in the Tidal

Freshwater of James River in Response to Riverine Nutrient Loading. *Ecol. Modell.* **2019**, *398*, 44–54. [CrossRef]

47. Yaseen, Z.M.; Sulaiman, S.O.; Deo, R.C.; et al. An Enhanced Extreme Learning Machine Model for River Flow Forecasting: State-of-the-Art, Practical Applications in Water Resource Engineering Area and Future Research Direction. *J. Hydrol.* **2019**, *569*, 387–408. [CrossRef]
48. El Bilali, A.; Taleb, A.; Brouziyne, Y. Groundwater Quality Forecasting Using Machine Learning Algorithms for Irrigation Purposes. *Agric. Water Manag.* **2021**, *245*, 106625. [CrossRef]
49. Naloufi, M.; Lucas, F.S.; Souihi, S.; et al. Evaluating the Performance of Machine Learning Approaches to Predict the Microbial Quality of Surface Waters and to Optimize the Sampling Effort. *Water* **2021**, *13*, 2457. [CrossRef]
50. Uddin, M.G.; Nash, S.; Rahman, A.; et al. A Comprehensive Method for Improvement of Water Quality Index (WQI) Models for Coastal Water Quality Assessment. *Water Res.* **2022**, *219*, 118532. [CrossRef]
51. Uddin, M.G.; Nash, S.; Rahman, A.; et al. Assessing Optimization Techniques for Improving Water Quality Model. *J. Clean. Prod.* **2023**, *385*, 135671. [CrossRef]
52. Xiong, Y.; Zhang, T.; Sun, X.; et al. Groundwater Quality Assessment Based on the Random Forest Water Quality Index—Taking Karamay City as an Example. *Sustainability* **2023**, *15*, 14477. [CrossRef]
53. Powers, D.M.W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *J. Mach. Learn. Technol.* **2020**, *2*, 37–63.
54. World Health Organization (WHO). *Guidelines for Drinking-Water Quality*, 4th ed.; World Health Organization: Geneva, Switzerland, 2017.
55. Adimalla, N.; Taloor, A.K. Hydrogeochemical Investigation of Groundwater Quality in the Hard Rock Terrain of South India Using Geographic Information System (GIS) and Groundwater Quality Index (GWQI) Techniques. *Groundw. Sustain. Dev.* **2020**, *10*, 100288. [CrossRef]
56. Kawo, N.S.; Karuppannan, S. Groundwater Quality Assessment Using Water Quality Index and GIS Technique in Modjo River Basin, Central Ethiopia. *J. Afr. Earth Sci.* **2018**, *147*, 300–311. [CrossRef]
57. Vaiphei, S.P.; Kurakalva, R.M.; Sahadevan, D.K. Water Quality Index and GIS-Based Technique for Assessment of Groundwater Quality in Wanaparthy Watershed, Telangana, India. *Environ. Sci. Pollut. Res.* **2020**, *27*, 45041–45062. [CrossRef]
58. Grassi, L.; Schileo, E.; Taddei, F.; et al. Accuracy of Finite Element Predictions in Sideways Load Configurations for the Proximal Human Femur. *J. Biomech.* **2011**, *45*, 394–399. [CrossRef]
59. Pany, R.; Rath, A.; Swain, P.C. Water Quality Assessment for River Mahanadi of Odisha, India Using Statistical Techniques and Artificial Neural Networks. *J. Clean. Prod.* **2023**, *417*, 137713. [CrossRef]
60. Wang, S.; Peng, H.; Liang, S. Prediction of Estuarine Water Quality Using Interpretable Machine Learning Approach. *J. Hydrol.* **2022**, *605*, 127320. [CrossRef]
61. Freeman, E.A.; Moisen, G.G. A Comparison of the Performance of Threshold Criteria for Binary Classification in Terms of Predicted Prevalence and Kappa. *Ecol. Modell.* **2008**, *217*, 48–58. [CrossRef]
62. Uddin, M.G.; Rahman, A.; Nash, S.; et al. Marine Waters Assessment Using Improved Water Quality Model Incorporating Machine Learning Approaches. *J. Environ. Manage.* **2023**, *344*, 118368. [CrossRef]
63. Haghiabi, A.H.; Nasrolahi, A.H.; Parsaie, A. Water Quality Prediction Using Machine Learning Methods. *Water Qual. Res. J.* **2018**, *53*, 3–13. [CrossRef]
64. Chen, X.; Zhang, M.; Zhang, W. Landscape Pattern Changes and Its Drivers Inferred from Salt Marsh Plant Variations in the Coastal Wetlands of the Liao River Estuary, China. *Ecol. Indic.* **2022**, *145*, 109719. [CrossRef]
65. Yongo, E.; Mutethya, E.; Zhang, P.; et al. Comparing the Performance of the Water Quality Index and Phytoplankton Index of Biotic Integrity in Assessing the Ecological Status of Three Urban Rivers in Haikou City, China. *Ecol. Indic.* **2023**, *157*, 111286. [CrossRef]
66. Xue, D.; Zhou, J.; Zhao, X.; et al. Impacts of Climate Change and Human Activities on Runoff Change in a Typical Arid Watershed, NW China. *Ecol. Indic.* **2021**, *121*, 107013. [CrossRef]