E-ISSN 2586-9027
Homepage : https://tci-thaijo.org/index.php/SciTechAsia Science & Technology Asia
Vol.30 No.4 October - December 2025 Page: [201-216]
Original research article

Predictive Modeling of Brackish Surface Water Quality for Reverse Osmosis Desalination Plants Using Advanced Machine Learning Techniques

Cherki Lahlou[1,*], Belaid Bouikhalene[2], Jamaa Bengourram[2], Hassan Latrache[1], Radouane El Amri[3]

[1]*Laboratory of Industrial Engineering, Sciences and Technology Faculty, Sultan Moulay Slimane University, Beni Mellal 23000, Morocco*
[2]*Laboratory of Mathematics Innovation and Information Technology (LIMATI), Polydisciplinary Faculty Beni Mellal, Sultan Moulay Slimane University, Beni Mellal 23000, Morocco*
[3]*Environmental, Ecological and Agro-Industrial Engineering Laboratory, Department of Chemistry and Environment, Faculty of Sciences and Technology, University Sultan Moulay Slimane, Beni Mellal 23000, Morocco*

Received 4 April 2025; Received in revised form 27 July 2025
Accepted 5 September 2025; Available online 17 December 2025

ABSTRACT

Artificial intelligence (AI) has proven highly effective in optimizing water treatment processes, particularly for monitoring reverse osmosis (RO) desalination plants treating brackish surface water. These systems face complex, non-linear variations in feed water quality, making real-time monitoring crucial to avoid performance loss and membrane fouling. This study presents a machine learning (ML)–based framework to predict the water quality index (WQI) and enable rapid decision-making by classifying water quality (WQC) into four actionable categories: Excellent, Good, Poor, and Unsuitable. Using 11 key water quality parameters, the model provides an efficient and reliable approach for prediction and classification. Among tested algorithms, the multi-layer perceptron (MLP) achieved the best WQI prediction performance, with an R^2 of 98.19%, a mean absolute error (MAE) of 0.0182, and a mean squared error (MSE) of 0.0043. For WQC, the XGBoost algorithm outperformed others, reaching 99.84% accuracy. The results demonstrate the strong potential of ML techniques to enhance water quality monitoring and management in RO desalination plants, supporting efficient operation and timely intervention.

Keywords: Brackish water; Machine learning; Reverse osmosis; WQC; WQI

***Corresponding author:** cherki.lahlou@gmail.com **doi:** 10.14456/scitechasia.2025.79

1. Introduction

The strain on freshwater resources has never been more urgent, as climate change and population growth relentlessly push us toward a global water crisis. The World Health Organization (WHO) reports that over 2.1 billion people worldwide lack access to clean drinking water, a statistic that's poised to worsen [1].

Amid this looming scarcity, brackish water (BW) desalination has emerged as a viable solution, particularly in arid and semi-arid regions. BW, which accounts for approximately 1% of the Earth's water, offers an attractive middle ground between freshwater (0.8%) and seawater (96.5%) due to its lower energy consumption and higher efficiency in desalination processes [2, 3]. Despite efforts to improve the economic viability of reverse osmosis (RO) technology, its performance is significantly influenced by feed water quality and plant operating conditions. These factors can cause membrane fouling, which has a negative impact on RO performance such as attack flow rate and pressure, pressure difference across the membrane (ΔP), permeate quality, membrane life and permeate flow rate [4]. This results in increased energy consumption, more frequent cleaning, higher operating and maintenance costs, greater use of chemicals, and membrane replacements, ultimately leading to high water treatment costs. Consequently, an RO plant requires a highly efficient pretreatment process and an advanced system to precisely control feed quality parameters to maintain optimal operation [5].

Enter the age of Industry 4.0 and artificial intelligence (AI) to revolutionize how we monitor and optimize complex systems. Machine learning (ML), a subset of AI, has demonstrated remarkable success in modeling nonlinear systems across various domains, including health, sports, industry, environmental sciences, and water treatment [7-11]. A great deal of research has been dedicated to developing reliable indicators for monitoring and optimizing the performance of RO membranes with the help of ML techniques [12, 13].

Due to its intuitive, dimensionless nature and its ability to capture the combined influence of several parameters characterizing RO feedwater quality, the Water Quality Index (WQI) has become a widely used tool by researchers assessing surface and groundwater quality, particularly when categorizing individual water quality parameters proves difficult [14], [15]. Several methods have been developed to calculate the WQI, making it easier to assess water quality [16, 17].

This research aims to evaluate the quality of brackish surface water supplied to the RO demineralization plant in Morocco, with a focus on introducing a multistage prediction model leveraging ML techniques. The model predicts the WQI from critical input parameters and classifies these values into actionable categories: Excellent, Good, Poor, and Unsuitable, thereby establishing a novel WQC system. The ultimate goal is to enhance plant operation and maintenance by optimizing key system performance parameters, including feed rate, feed pressure transmembrane pressure (ΔP), permeate flow rate and energy consumption.

In this study, we gathered a dataset containing key physicochemical parameters, including silt density index (SDI), pH, turbidity, redox potential (ORP), chlorides, total iron, electrical conductivity (EC), alkalinity, calcium, and oxidizability. This dataset was used to train various machine learning models, such as k-Nearest Neighbors (KNN), random forests (RF), and Mul-

tilayer Perceptrons (MLP), to predict the WQI. In parallel, the same WQI values were used to place WQC into four categories: Excellent, Good, Poor, and Unsuitable. For this task, decision tree (DT), extreme gradient boosting (XGB-oost), and support vector machine (SVM) models were applied. The performance of the predictive models was rigorously evaluated using metrics such as mean absolute error (MAE), mean squared error (MSE), and the coefficient of determination (R^2). For classification tasks, additional evaluation criteria such as accuracy, precision, recall, and F1 score were employed. The novelty of this study lies in the monitoring of nonlinear variations in quality parameters in brackish surface waters feeding RO membranes through the development of a robust ML based framework capable of predicting WQI and classifying WQC into meaningful categories using 11 critical parameters.

2. Materials and Methods

This section provides a detailed description of the study area and the datasets used in this research. It also outlines the machine learning methods adopted for the analysis and explains the evaluation criteria applied to the developed models. Fig. 1 clearly illustrates the methodological process followed throughout this study.

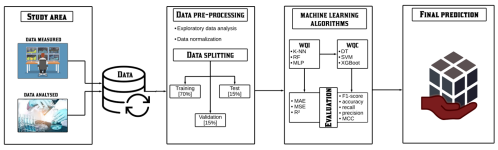

Fig. 1. Methodological framework for dataset analysis and machine learning model development.

2.1 Study area

The demineralization plant MAROC CENTRAL, the second of its kind in Morocco, was established in 2018. It is situated approximately 8.21 km southeast of Kasba Tadla in the Béni Mellal-Khénifra region, at coordinates 32.5874°N and -6.1847°W. It covers a total area of 129,321.99 m². The regional climate ranges from humid at higher altitudes to semi-arid on the plains, with temperatures reaching 46°C in August and dropping to -2°C in January [18]. The Ait Massoud dam supplies the plant, which has variable levels of suspended solids and high chloride concentrations that can affect the taste of treated water. The plant employs two main treatment processes (Fig. 2):

- **Conventional Pretreatment:** This includes coagulation flocculation (using ferric chloride and anionic polymers), settling, and filtration, aiming to prevent the deposition of suspended solids on the membranes and ensure optimal filtration.
- **Microfiltration and Reverse Osmosis (RO) membranes:** This process removes salts and impurities such as metals, radionuclides, bacteria, and organic substances from brackish, producing potable water.

A portion of the treated water is blended with demineralized water to adjust the chloride levels to the desired concentration. Since chloride levels in raw water fluctuate seasonally, the ratio of conventionally treated water to RO-treated water is adjusted accordingly to maintain optimal chloride levels in the final product.

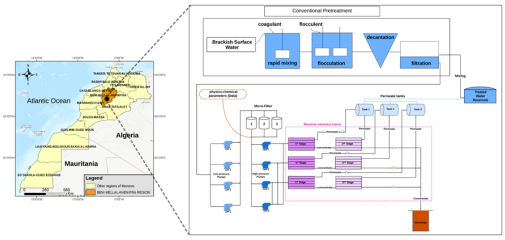

Fig. 2. Geographical location and treatment process of the studied maroc central demineralization plant.

2.2 Data Description

The data used in this study originates from 10,000 samples of 11 physicochemical parameters collected at the inlet of the MAROC CENTRAL demineralization plant. These samples, essential for assessing feed water quality, were acquired through a hybrid monitoring approach:

- **Real-time measurements:** were collected via an industrial SCADA system equipped with sensors connected through the Modbus-TCP/IP protocol. These sensors are installed at the membrane feed collector and continuously monitor key parameters.
- **Laboratory analyses:** were performed on parameters requiring higher analytical precision (such as total iron and SDI), using techniques such as spectrophotometry and manual titration, in the plant's on-site laboratory. Table 1 summarizes the key measured parameters and their operational relevance to the RO system. Each variable plays a critical role in predicting fouling potential, membrane scaling, system stability, and overall treatment performance. The table also indicates the measurement techniques used and the practical impact of each parameter on system efficiency and longevity.

Table 1. Key Parameters for Reverse Osmosis Inlet Water Quality.

Parameter	Average Measurements
Temperature (°C)	Real-Time Sensors
Silt Density Index (SDI)	Manually with SDI test kit
pH	Real-Time Sensors
Turbidity (NTU)	Real-Time Sensors
Redox Potential (ORP) (mV)	Real-Time Sensors
Chlorides ($mg{\cdot}L^{-1}$)	Titration Method
Total Iron ($mg{\cdot}L^{-1}$)	Spectrophotometry
Electrical Conductivity (EC) ($\mu{\cdot}Scm^{-1}$)	Real-Time Sensors
Alkalinity (0F)	Titration Method
Calcium ($mg{\cdot}\ L^{-1}$)	EDTA Titrimetric Method
Oxidizability ($mg{\cdot}L^{-1}$)	Permanganate Titration Method (COD Equivalent)

2.3 Machine learning models for WQI prediction

To predict WQI from physicochemical parameters, we used three proven regression models k-NN, RF, and MLP, which were selected for their ability to capture nonlinear patterns in environmental data. Table 2 summarizes the optimal hyperparameters used to fine-tune each model for maximum predictive performance.

2.3.1 K-Nearest Neighbour (k-NN)

k-NN is a simple yet effective nonparametric model that makes predictions based on the average output of the k most similar training samples, as determined by a distance metric such as Euclidean distance. This method works particularly well when local similarities in input data correlate strongly with the output values. While its performance is sensitive to the choice of k and distance function, its interpretability and ease of implementation make it a useful baseline for water quality prediction tasks [19].

2.3.2 Random Forest (RF)

RF is an ensemble learning method that constructs a large number of decision trees and outputs their average prediction. This approach mitigates the variance seen in single trees, enhancing robustness and generalization. RF handles missing data and

Table 2. Grid Search: Tested and Optimal Hyperparameters.

Approach	Model	Hyperparameters		
		Parameter	**Values Tried**	**Best Value**
WQI	k-NN	n_neighbors	[3, 5, 7, 10]	5
		algorithm	[’auto’,’ball_tree’, ’kd_tree’, ’brute’]	auto
	RF	n_estimators	[50, 100, 150, 200]	100
		max_depth	[None, 10, 20]	None
		min_samples_split	[2, 5, 10]	2
	MLP	hidden_layer_sizes	[(50,50), (100,100), (50,100,50)]	(100,100)
		Activation	[’relu’, ’tanh’, ’Sigmoid’]	Relu
		Alpha	[0.0001, 0.001, 0.01]	0.0001
		Criterion	[’gini’, ’entropy’]	gini
		Splitter	[’best’, ’random’]	random
WQC	DT	max_depth	[None, 10, 20, 30, 40, 50]	None
		min_samples_split	[2, 5, 10]	2
		min_samples_leaf	[1, 2, 4]	1
		max_depth	[3, 6, 9]	3
	XGBoost	n_estimator	[100, 200, 300]	200
		gamma	[0, 0.1, 0.2]	0.1
		subsample	[0.8, 0.9, 1.0]	0.8
	SVM	C	[0.1, 1, 10, 100]	0.1
		Gamma	[’scale’, ’auto’, 0.001, 0.01, 0.1, 1]	auto
		kernel	[’linear’, ’poly’, ’rbf’, ’sigmoid’]	linear

multicollinearity well, and has been widely used in hydrology and water quality modeling [20]. In our study, RF provided strong baseline performance and helped assess the importance of individual input variables. Fig. 3 illustrates the RF prediction scheme.

2.3.3 Multi-layer perceptron regression model (MLP)

MLP is a class of feedforward artificial neural networks consisting of input, hidden, and output layers. It is capable of learning complex patterns and interactions between variables through backpropagation and nonlinear activation functions. Due to its flexibility and capacity to ap-

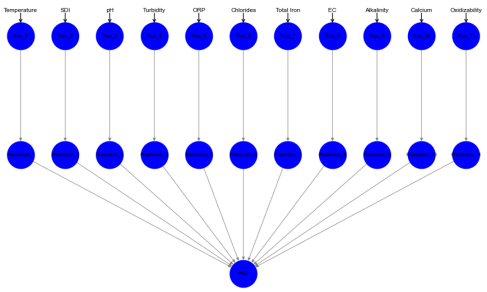

Fig. 3. The random forest (RF) prediction scheme.

proximate any continuous function, MLP has been successfully applied in various environmental and water quality forecasting contexts [21, 22]. We used a standard three-layer MLP model, trained with regularization to avoid overfitting.

By comparing these three approaches, we aim to evaluate both traditional and deep learning strategies for WQI prediction and determine which model best captures the nonlinear relationships inherent in brackish water quality data. Fig. 4 illustrates the MLP neural network architecture.

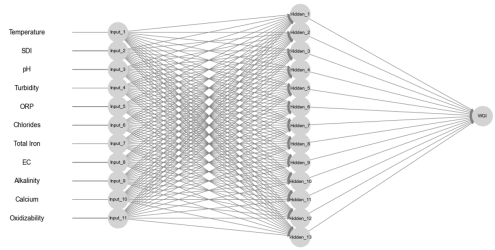

Fig. 4. The structure of a multi-layer perceptron (MLP).

2.4 Machine learning models for WQC

In classifying the quality of brackish surface water prior to reverse osmosis treatment, we used three robust ML models DT, XGBoost and SVM, which were chosen for their effectiveness in environmental classification and processing of nonlinear multivariate data. Table 2 shows the optimal hyperparameters selected for each model.

2.5 Decision Tree (DT)

Decision trees are simple, interpretable classification models that split data based on input features, forming a flowchart-like structure. They are well suited for short-term predictions and fast computations but can be prone to overfitting. While useful individually, decision trees typically perform better when combined in ensemble methods like random forests or boosting frameworks [23].

2.5.1 Extreme gradient boosting (XGBoost)

XGBoost is an advanced ensemble method that improves performance by iteratively combining multiple weak learners (typically decision trees). It includes regularization techniques that help prevent overfitting and is known for its scalability and accuracy in structured data classification tasks [24].

2.5.2 Support Vector Machines (SVM)

SVMs are powerful classifiers that identify the optimal hyperplane separating classes with the maximum margin. They are especially effective in high-dimensional spaces. For non-linearly separable problems, SVMs can employ kernel functions (e.g., radial basis function, polynomial) to map data into higher-dimensional feature spaces where linear separation becomes feasible [25].

2.6 Data preparation and exploration

Data preparation begins with a thorough exploration of the collected water quality parameters. This step is crucial for understand-ding the general characteristics of the datasets and identifying poten-

tial issues such as missing or outlier values. Several preprocessing techniques were employed in this study to ensure the reliability and quality of the data before applying ML models. These techniques include descriptive statistical analysis, data cleaning, normalization, and transformation.

2.6.1 Statistical analysis

Descriptive statistical analysis provides a detailed overview of the water quality parameters, identifying trends, detecting outliers, and assessing the overall distribution of the dataset. This analysis is crucial for understanding the data's characteristics and ensuring its adequacy for ML applications. Table 3 presents the statistical computations conducted on the dataset attributes, including key metrics such as mean, minimum, maxi-mum and standard deviation. These metrics illustrate the parameters' distribution and variability, establishing a robust foundation for further preprocessing and analytical steps.

Table 3. Statistical characteristics of data used.

Parameter	Max	Mean	Min	Std
EC	3350	3060	1270	345.12
Chlorides	1100	935.90	565.60	112.98
Turbidity	0.44	0.20	0.10	0.04
Temperature	29.00	20.07	9.00	5.18
pH	7.95	7.28	6.92	0.58
Alkalinity	23.00	18.90	13.3	3.03
Oxidability	2.84	1.87	0.78	0.40
Calcium	100.00	79.66	51.00	12.95
Total Iron	0.05	0.01	0.00	0.01
ORP	625.00	248	136.30	54.68
SDI	5.50	3.19	1.57	0.95
WQI	140.16	81.11	45.13	14.57

2.6.2 Data transformation and normalization

Data transformation was performed to ensure compatibility and comparability across parameters with different units of measurement. Normalization was particularly important to scale the values into a consistent range. The normalization technique used transforms data values into a range between 0.1 and 0.9 using the following equation:

$$x_{new} = 0.1 + (X - x_{\min}) \times \frac{0.9 - 0.1}{x_{\max} - x_{\min}}. \quad (2.1)$$

This approach ensures that all parameters, such as turbidity (measured in NTU) and conductivity (measured in µS/cm), are standardized, thereby improving the accuracy and performance of machine learning algorithms.

2.6.3 WQI calculation and class classification

The primary goal of this analysis is to compute the WQI, a composite indicator that integrates several physicochemical parameters. Each parameter contributes uniquely to the overall evaluation for rapid decision-making on the need for intervention by classifying feed water into four distinct categories: Excellent, Good, Poor and Unsuitable, thus forming the WQC. This classification provides actionable insights into water quality entering the reverse osmosis systems.

- **Calculating the WQI** The Water Quality Index (WQI) in this study is calculated using a weighted aggregation method, where each water quality parameter is assigned a specific weight that reflects both its relative impact on RO system performance and its operational relevance. These weights were not chosen arbitrarily. Instead, they result from a hybrid approach based on:
  - Technical guidelines and threshold values recommended by membrane manufacturers for optimal RO feedwater

quality (e.g., DuPont FilmTec, Hydranautics, Toray) [26–29];

- Industrial feedback from operational monitoring at the MAROC CENTRAL Demineralization Plant, where certain parameters (such as SDI, iron content, and turbidity) were consistently correlated with membrane fouling and degradation;
- Established literature where similar weighting strategies were applied in WQI computation for RO and other water treatment systems [30, 31].

This weighting approach allows for a composite and dimensionless indicator that captures the combined influence of individual water quality parameters, thus facilitating better operational decisions and risk management in RO desalination systems.

General WQI Formula:

$$WQI = \sum_{i=1}^{n} (P_i \times W_i), \qquad (2.2)$$

where n is the number of parameters included in the calculation. W_i is the relative weight assigned to the ith parameter based on its importance in RO feed water quality. P_i is the normalized value of the ith parameter, defined as:

$$P_i = \frac{X_i - X_{i,\min}}{X_{i,\max} - X_{i,\min}}, \qquad (2.3)$$

with X_i being the actual value of the parameter, and, $X_{i,\max}, X_{i,\min}$ the permissible minimum and maximum values, respectively.

Table 4 summarizes the final weights used for each parameter, as well as the permissible values recommended by RO system manufacturers.

Table 4. Weights and permissible limits for parameters used to calculate WQI.

Parameter	Weight (%)	Permissible Values for Feed Water RO (per Manufacturer)
Temperature (dimensioning value)	12.2	9–30 °C
Turbidity	16.2	<1 NTU
EC	4.0	3700 μ·Scm$^{-1}) (dimensioning value)
pH	4.0	6.5–7.5
Alkalinity	12.2	<100 mg/L (as $CaCO_3$)
Oxidizability	16.2	<2 mg/L O_2
Calcium	12.2	<70 mg/L (as Ca^{2+})
Chlorides (dimensioning value)	4.0	<1200 mg/L
Total Iron	20.3	<0.05 mg/L
ORP	12.2	<260 mV
SDI	24.3	<5

- Water Quality Classification (WQC)

The WQC is an essential post processing step derived from the WQI to categorize water quality into actionable classes. These categories facilitate rapid decision-making in reverse osmosis (RO) operations, guiding interventions such as membrane cleaning, pressure adjustments, or flow regulation.

In this study, the WQC was divided into four predefined categories based on WQI values, in alignment with established water quality standards, according to function (2.12).

$$WQC = \begin{cases} \text{Excellent} & \text{if } WQI \leq 60, \\ \text{Good} & \text{if } 60 < WQI \leq 80, \\ \text{Poor} & \text{if } 80 < WQI \leq 100, \\ \text{Unsuitable} & \text{if } WQI > 100. \end{cases} \qquad (2.4)$$

Initially, the class distribution was imbalanced, which posed challenges for classifying performance. To overcome this, we applied data augmentation and resampling strategies to balance the number of samples across all WQC categories. This adjustment aimed to enhance model reliability, minimize classification bias, and improve the accuracy for underrepresented

classes. Fig. 5 presents the updated and balanced distribution of water samples across the four WQC categories following augmentation.

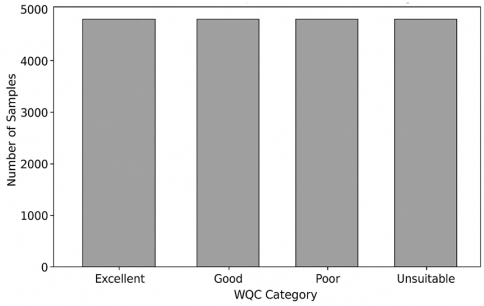

Fig. 5. Updated distribution of water samples across WQC categories.

This classification framework is essential for operational monitoring of RO system performance indicators, including feed rate, feed pressure, transmembrane pressure (ΔP), permeate flow rate, and energy consumption. As shown in Fig. 6, there is a clear relationship between WQC levels and key RO performance metrics.

2.7 Model training and evaluation

The process of training and evaluating of the ML models implies the use of a certain methodology to make the models as accurate and effective as possible. The data set is split into three parts: The first part is the training set, which contains 70% of the data, the validation set contains 15% of the data and the test set also contains 15% of the data. The training subset is used to train the ML models from which the models are able to learn the patterns and relationships that are present in the data. The validation subset is used in the hyperparameter tuning and to check the performance of the model in the training process thus helping to prevent overfitting. The testing subset is data that was not used during training at all and is used for the testing of the models' ability to generalize. The evaluation phase is very important in order to determine how well the trained models perform. For WQI prediction tasks, the performance of the trained models is evaluated using root mean square error (RMSE), mean absolute error (MAE) and coefficient of determination (R^2). With respect to WQC classification, the performance of the models is evaluated using metrics like accuracy, precision, recall and F1-score thus providing a rich characterization of the classification performance. In this way, the most effective and reliable of the algorithms are selected for the subsequent studies and applications. Table 5 presents the equations for the evaluation metrics employed in this study.

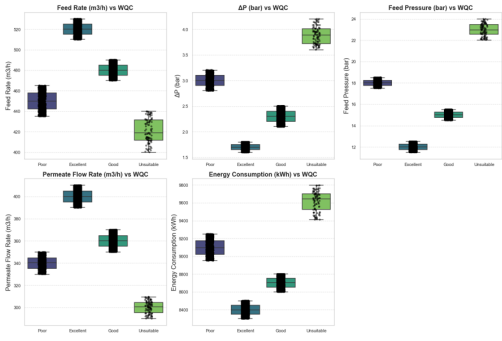

Fig. 6. Relationship between RO System Performance and Water Quality Classification (WQC).

Table 5. Mathematical formulations of evaluation metrics for model performance assessment.

APPROACH	METRIC	EQUATION
WQI	MAE	$\frac{1}{N}\sum_{i=1}^{N}\lvert y_{rèel}i - y_{pred}i\rvert$ (2.5)
	MSE	$\frac{1}{N}\sum_{i=1}^{N}(y_{rèel}i - y_{pred}i)^2$ (2.6)
	R^2	$1 - \frac{\sum_{i=1}^{N}(y_{rèel}{}^{i} - y_{pred}{}^{i})^2}{\sum_{i=1}^{N}(y_{rèel}{}^{i} - \bar{y})^2}$ (2.7)
WQC	Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$ (2.8)
	Recall	$\frac{TP}{TP+FN}$ (2.9)
	Precision	$\frac{TP}{TP+FP}$ (2.10)
	F1 score	$\frac{2\times Recall\times Precision}{Recall+Precision}$ (2.11)

2.8 Data distribution and correlation analysis

Following the preprocessing phase, the dataset was analyzed to assess its distribution and ensure readiness for model training and testing. Understanding the distribution of features is critical for detecting potential imbalances or skewness that could affect the performance of machine learning models. Histograms were created for each water quality parameter, as illustrated in Fig 8. The analysis revealed a diverse range of distributions across the parameters. For instance, Conductivity and Chlorides demonstrated a right-skewed distribution, indicating that higher values are more frequent in the dataset. On the other hand, pH exhibited a near-normal distribution, which aligns with expected natural variability. Certain parameters, such as SDI and Turbidity, displayed uniform distributions, suggesting a relatively even spread of values across their respective ranges

These insights provided a deeper understanding of the data and guided subsequent steps in the modeling process, such as feature scaling or normalization, to address potential biases. The identification of skewed features also informs the selection of machine learning models and loss functions that can handle non-normal distributions effectively.

The correlation matrix depicted in Fig. 8, provides a comprehensive view of the relation-ships among various physicochemical parameters. Using Pearson's correlation coefficient, we quantified the linear dependencies between pairs of variables. The coefficients, ranging from -1 to 1, reflect the strength and direction of these relationships, where values closer to 1 or -1 indicate strong positive or negative correlations, respectively, and values near zero denote weak or no linear correlation. The

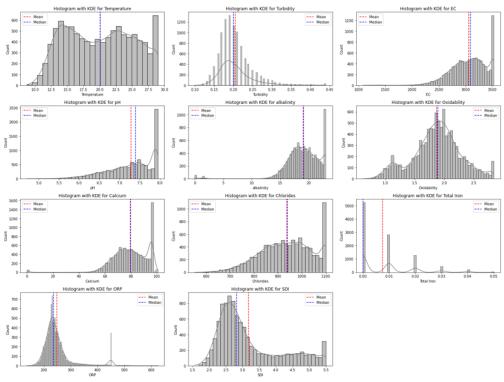

Fig. 7. Distribution histograms of water quality parameters.

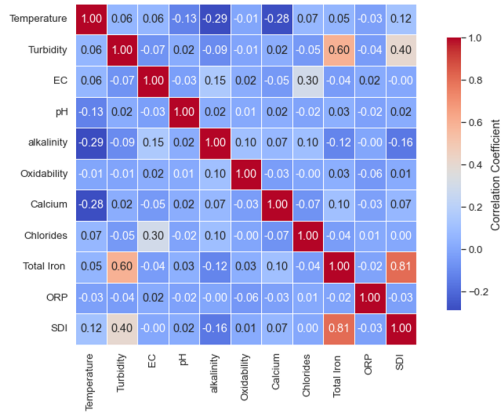

Fig. 8. Heatmap representation of the correlation matrix for water quality parameters.

heatmap's color gradient enhances the interpretability of these relationships, with shades of red highlighting positive correlations and blue representing negative ones. Such visualizations are invaluable for identifying key interdependencies and guiding feature selection in predictive modeling efforts.

3. Results and Discussion

3.1 Results for regression algorithms

Table 6 summarizes the predictive performance of the RF, KNN, and MLP models for WQI estimation during the training, testing, and validation phases. The

results demonstrate that the MLP model outperformed the RF and KNN models across all metrics, achieving superior accuracy with the lowest MAE and MSE, as well as the highest R^2 values. Specifically, the MLP model recorded MAE values of 0.0182, 0.0278, and 0.0265 for the training, testing, and validation datasets, respectively, with corresponding R^2 values of 0.9819, 0.9796, and 0.9806.

These results highlight the MLP's exceptional ability to capture the complex relation-ships inherent in the data.

Fig. 9 presents the time series plots of predicted versus actual WQI values for the three models during the training, testing, and validation phases. The red dashed line represents the actual values, while the black line represents the predicted values. For the MLP model, the predicted and actual values align closely, indicating a highly precise fit. The RF model also shows good alignment, although with minor deviations in the testing and validation phases. The

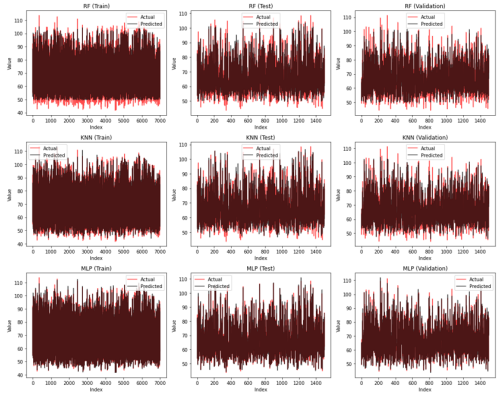

Fig. 9. Time series plot of WQI: training, testing, and validation phases.

KNN model, however, displays more pronounced discrepancies, particularly during the testing phase, consistent with its lower accuracy metrics.

The findings underscore the potential of advanced machine learning models, particularly MLP, for accurate and reliable water quality assessment in complex datasets.

Fig. 10 illustrates the regression plots for the RF, KNN, and MLP models, showcasing their performance in predicting WQI during the training, testing, and validation phases. The regression plots visualize the relationship between each model's observed WQI values (target, x-axis) and the predicted WQI values (output, y-axis). The diagonal dashed line (Y = T) represents the ideal scenario where predictions perfectly match the observed values. The solid red line illustrates the best-fit linear regression between the predicted and observed values. For the MLP model, the points are closely clustered around the dashed line, indicating a strong correlation and highly accurate predictions. The regression equations for the MLP model also show slopes close to 1, further demonstrating its efficiency in capturing the underlying relationships in the data.

3.2 Results for classification algorithms

This section details the performance of three ML models DT, XGBoost, and SVM applied to the classification of WQC. The dataset was split into three distinct subsets: 70% for training, 15% for testing, and 15% for validation. The evaluation metrics include accuracy, recall, precision, and F1-score, which comprehensively assess each model's ability to classify water quality accurately.

Table 7 shows the results obtained for each model across the three phases. The XGBoost model consistently outperformed the other two algorithms, achieving the highest accuracy during the training (99.84%), testing (100%), and validation (99.73%) phases. The DT model demonstrated similar levels of performance, albeit

Table 6. Model performance for predicting the WQI.

Metric	Training			Testing			Validation		
	RF	KNN	MLP	RF	KNN	MLP	RF	KNN	MLP
MAE	0.0524	0.0325	0.0182	0.0572	0.0346	0.0278	0.0564	0.0341	0.0265
MSE	0.0097	0.0058	0.0043	0.0115	0.0084	0.0049	0.0109	0.0081	0.0046
R^2	0.9130	0.9617	0.9819	0.9010	0.9247	0.9796	0.9029	0.9213	0.9806

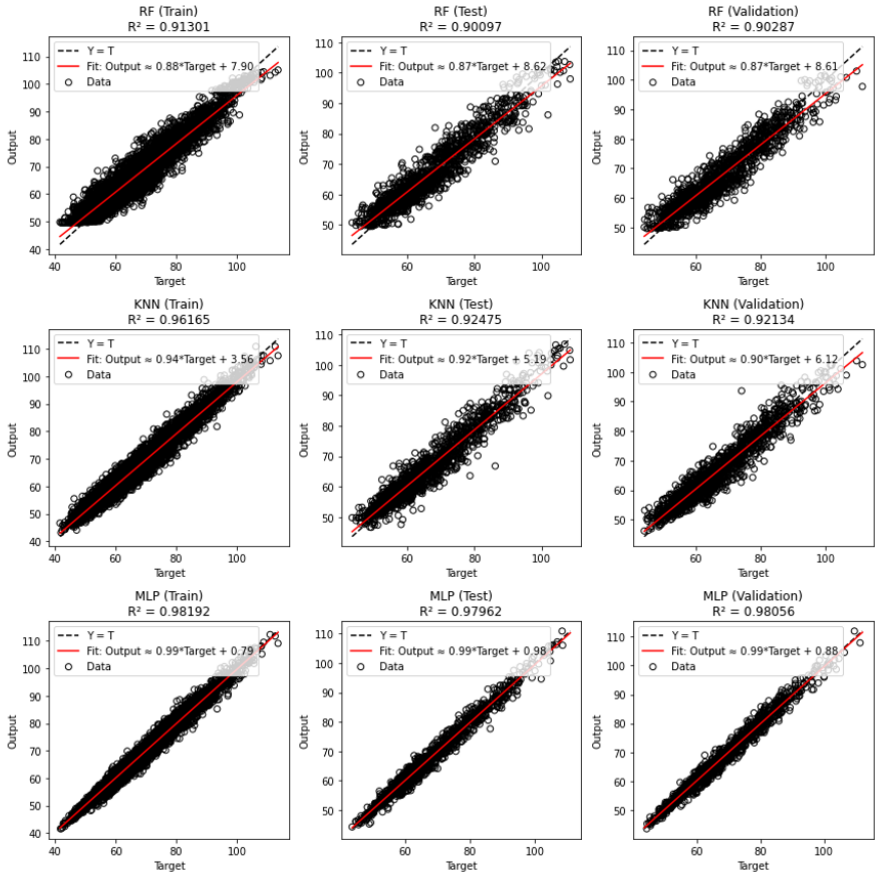

Fig. 10. Actual Values vs. Predicted Values for Regressor Models.

slightly below XGBoost. The SVM model, while slightly less accurate, still delivered strong results, with a validation accuracy of 99.13%.

Fig. 11 presents the confusion matrices for the DT, XGBoost, and SVM models, offering a comprehensive assessment of their classification performance across the training, testing, and validation phases. The x-axis represents the actual class labels, while the y-axis shows the predicted class labels. The WQC are defined as 0 (Excellent), 1 (Good), 2 (Poor), and 3 (Unsuitable). Among the models, XGBoost stood out with exceptional performance, demonstrating minimal false positives and false negatives throughout all phases. Its high accuracy, precision, and

Table 7. Performance Metrics for DT, XGBoost, and SVM.

Model	Phase	Accuracy	Recall	Precision	F1-Score
DT	Training	99.84%	99.84%	99.84%	99.84%
	Testing	99.93%	99.93%	99.94%	99.93%
	Validation	99.67%	99.67%	99.67%	99.67%
XGBoost	Training	99.84%	99.84%	99.84%	99.84%
	Testing	100%	100%	100%	100%
	Validation	99.73%	99.73%	99.74%	99.73%
SVM	Training	99.31%	99.31%	99.31%	99.31%
	Testing	99.33%	99.33%	99.33%	99.33%
	Validation	99.13%	99.13%	99.14%	99.13%

F1-scores confirm its consistency and establish it as the top-performing model. The DT model performed similarly, delivering competitive metrics that make it a strong and reliable alternative. Conversely, the SVM model, while overall robust, showed a slightly higher frequency of false negatives during the validation phase, leading to a marginal reduction in its F1-score.

3.3 Discussion and comparative evaluation

The assessment and prediction of water quality using ML techniques are critical for the effective monitoring and management of reverse osmosis (RO) systems. This study focused on evaluating the quality of brackish surface water feeding into a RO demineralization plant in Morocco while optimizing key system performance indicators such as feed rate, transmembrane pressure (ΔP), permeate flow rate, energy consumption, and membrane longevity.

The proposed ML framework demonstrated a high capacity to predict the WQI and to classify Water Quality Categories (WQC) with remarkable accuracy. In particular, the MLP model achieved strong predictive performance with an MAE of 0.0182, an MSE of 0.0043, and an R^2 of 98.19%, while the XGBoost model yielded a classification accuracy of 99.84%, supported by equally high precision, recall, and F1-score.

To evaluate the added value of our approach, we compared its performance with both conventional techniques (manual WQI computation and fixed-threshold classify-cation) and results from recent studies in the literature. This comparison, summarized in Table 8, highlights the superior performance and practical benefits of our proposed models. While conventional methods are static and sensitive to human error, our AI-based approach is adaptive, handles nonlinearities effectively, and integrates both prediction and classification in a unified system.

Table 8. Comparative analysis between the proposed AI-based approach and conventional manual and literature methods for water quality prediction and classification.

Ref	Method	Predi-ction	Accuracy R^2	MSE
This study (Proposed)	MLP XGBoost	WQI WQC	98.19% 99.84%	0.0043
Ahmed et al, [32]	Poly. Regr MLP	WQI	85.07%	7.9467
Sakizadeh et al, [33]	ANN	WQI	77%	9.25
Gad et al, [34]	RNA ANN	WQI	99%	-

Compared to existing works, such as those by Ahmed et al., Sakizadeh et al., and Gad et al., our methodology demonstrates improved accuracy, a broader set of input parameters, and the unique advantage of targeting brackish surface water used in reverse osmosis, which remains underexplored in the current literature. These results confirm the robustness and relevance

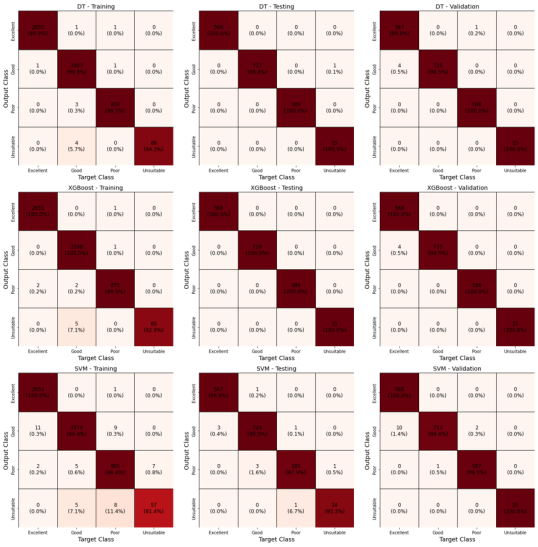

Fig. 11. Confusion matrix for DT, XGBoost and SVM algorithms.

of the proposed approach for real-time monitoring, operational decision-making, and preventive maintenance in large-scale desalination contexts. The proposed system is distinguished by its focus on brackish surface water feeds to RO membranes, which present unique challenges due to their complex quality variations.

4. Conclusion

This research set out to evaluate the quality of surface water feeding the RO demineralization plant in Morocco, with a focus on optimizing system performance metrics such as feed rate, transmembrane pressure (ΔP), permeate flow rate, energy consumption, and membrane longevity. By leveraging ML models, the study aimed to enhance the predictive accuracy of the WQI and the WQC. The findings highlight the exceptional performance of the MLP model for WQI prediction, achieving an MAE of 0.0182, an MSE of 0.0043, and an R^2 of 98.19% during the training phase.

For WQC, the XGBoost model excelled, achieving a testing accuracy of 100% and a validation accuracy of 99.84%. The DT model followed closely with comparable metrics, while the SVM model, despite decent performance (validation accuracy of 99.13%), fell short of XGBoost's reliability and robustness. These results align with global research trends that underscore the utility of advanced AI models for water quality monitoring. The findings of this study extend these insights by tailoring them to Morocco's unique environmental context and water quality challenges.

Future work should focus on integrating Internet of Things (IoT) technologies and real-time monitoring systems for continuous data collection and immediate response to water quality fluctuations, advanced hyperparameter tuning and the exploration of deep learning architectures could further enhance predictive accuracy and classification reliability. Additionally, enriching the dataset with new explanatory variables and applying feature selection techniques could help identify the most critical factors influencing WQI and WQC.

Acknowledgements

The authors would like to thank the laboratory teams at the MAROC CENTRAL demineralization plant for their invaluable help and support.

References

[1] Chen X, Xia X, Liang P, Cao X, Sun H, Huang X. Stacked microbial desalination cells to enhance water desalination efficiency. Environ Sci Technol. 2011;45(6):2465–70.

[2] Greenlee LF, Lawler DF, Freeman BD, Marrot B, Moulin P, Ce P. Reverse osmosis desalination: Water sources, technology, and today's challenges. Water Res. 2009;43(9):2317–48.

[3] Mun I, Rodríguez A. Reducing the environmental impacts of reverse osmosis desalination by using brackish groundwater resources. Water Res. 2008;42:801–11.

[4] Badruzzaman M, Voutchkov N, Weinrich L, Jacangelo JG. Selection of pretreatment technologies for seawater reverse osmosis plants: A review. Desalination. 2019;449:78–91.

[5] Abbas A. Model predictive control of a reverse osmosis desalination unit. Desalination. 2006;194:268–80.

[6] Sircar A, Yadav K, Rayavarapu K, Bist N, Oza H. Application of machine learning and artificial intelligence in oil and gas industry. Pet Res. 2021;6(4):379–91.

[7] Bertolini M, Mezzogori D, Neroni M, Zammori F. Machine learning for industrial applications: A comprehensive literature review. Expert Syst Appl. 2021;175:114820.

[8] Stiglic G, Kocbek P. Interpretability of machine learning-based prediction models in healthcare. Wiley Interdiscip Rev Data Min Knowl Discov. 2020;10(5):e1379.

[9] Ferdous M, Debnath J, Chakraborty NR. Machine learning algorithms in healthcare: A literature survey. In: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT). 2020. p. 1–6.

[10] Bilali AE, Taleb A, Mazigh N, Mokhliss M. Prediction of chemical water quality used for drinking purposes based on artificial neural networks. Moroccan J Chem. 2020;8(3):665–72.

[11] Zaveri J, Dhanushkodi SR, Bansal L. Towards machine learning in water treatment: A diagnostic tool for assessing water quality. Desalin Water Treat. 2023;286:64–72.

[12] Ridwan MG, Altmann T, Yousry A, Das R. Intelligent framework for coagulant dosing optimization in an industrial-scale seawater reverse osmosis desalination plant. Mach Learn Appl. 2023;12:100475.

[13] Akhtar N, Izzuddin M, Ishak S, Ahmad MI, Umar K, Yusuff MS, et al. Modification of the Water Quality Index (WQI) process for simple calculation using the multi-criteria decision-making. Water. 2021;13(7):905.

[14] Zhang Z, Zhang W, Hu X, Li K, Luo P, Li X, et al. Evaluating the efficacy of point-of-use water treatment systems using the water quality index in rural southwest China. Water. 2020;12(3):867.

[15] Shams MY, Elshewey AM, Sayed E, El M. Water quality prediction using machine learning models based on grid search method. Multimed Tools Appl. 2024;83(12):35307–34.

[16] Krushna BV, Sasikala D. Comparative analysis of machine learning models for water quality prediction. In: 2024 4th International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT 2024). 2024.

[17] Lebrini Y, Boudhar A, Lionboui RHH, Arrach LER. Identifying agricultural systems using SVM classification approach based on phenological metrics in a semi-arid region of Morocco. Earth Syst Environ. 2019;3(2):277–88.

[18] Salinas-Rodríguez SG, Schippers JC, Amy GL, Kim IS, Kennedy MD. Seawater reverse osmosis desalination: Assessment and pre-treatment of fouling and scaling. 1st ed. Elsevier; 2021.

[19] Kubat M, Kubat JA. An introduction to machine learning. 3rd ed. Cham (Switzerland): Springer International Publishing; 2017. p. 321–9.

[20] Breiman L. Random forests. Mach Learn. 2001;45:5–32.

[21] Chen Y, Song L, Liu Y, Yang L, Li D. A review of the artificial neural network models for water quality prediction. Water. 2020;10(17):5776.

[22] Ahmed AN, Othman FB, Afan HA, Ibrahim RK, Fai CM, Hossain MS, Elshafie A. Machine learning methods for better water quality prediction. J Hydrol. 2019;578:124084.

[23] Lu H, Ma X. Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere. 2020;249:126169.

[24] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p. 785–94.

[25] Guo Y, Zhan W, Li W. Application of support vector machine algorithm incorporating slime mould algorithm strategy in ancient glass classification. Appl Sci. 2023;13(6):3718.

[26] DuPont Water Solutions. FilmTec™ reverse osmosis membranes technical manual. DuPont; 2022.

[27] Hydranautics Technical Service Bulletin. Recommended feed water quality guidelines. Hydranautics; 2021.

[28] Toray Membrane USA. RO membrane design and operating limits. Toray; 2020.

[29] Brown RM, McClelland NI, Deininger RA, Tozer RG. A water quality index—do we dare. Water Sewage Works. 1970;117(10).

[30] Pesce SF, Wunderlin DA. Use of water quality indices to verify the impact of Córdoba City (Argentina) on Suquía River. Water Res. 2000;34(11):2915–26.

[31] Tyagi S, Sharma B, Singh P, Dobhal R. Water quality assessment in terms of water quality index. Am J Water Resour. 2013;1(3):34–8.

[32] Ahmed U, Mumtaz R, Anwar H, Shah AA, Irfan R, García-Nieto J. Efficient water quality prediction using supervised machine learning. Water (Switzerland). 2019;11(11):1–14.

[33] Sakizadeh M. Artificial intelligence for the prediction of water quality index in groundwater systems. Model Earth Syst Environ. 2016;2(1):1–9.

[34] Gad M, Saleh AH, Hussein H, Elsayed S. Water quality evaluation and prediction using irrigation indices, artificial neural networks, and partial least square regression models for the Nile River, Egypt. Water. 2023;15(12):2244.

[35] Mura I, Franco JF. Enriched spatial analysis of air pollution: Application to the city of Bogotá, Colombia. Front Environ Sci. 2022;10:966560. doi:10.3389/fenvs.2022.966560.