smart cities

MDPI

Article

Multiple Linear Regression and Machine Learning for Predicting the Drinking Water Quality Index in Al-Seine Lake

Raed Jafar [1,2,*], Adel Awad [2], Iyad Hatem [1], Kamel Jafar [3], Edmond Awad [4] and Isam Shahrour [5]

[1] Engineering Faculty, Manara University, Lattakia HQ28 RFM, Syria; iyad.hatem@manara.edu.sy
[2] Environmental Engineering Department, Tishreen University, Lattakia P.O. Box 1385, Syria; a.awad@tishreen.edu.sy
[3] Information Technology Department, Syrian Virtual University, Damascus P.O. Box 35329, Syria; kamel_213663@svuonline.org
[4] Department of Economics, University of Exeter, Exeter EX4 4QJ, UK; e.awad@exeter.ac.uk
[5] Laboratory of Civil Engineering and Geo-Environment (LGCgE), University of Science and Technology of Lille, 59650 Villeneuve-d'Ascq, France; isam.shahrour@univ-lille.fr
* Correspondence: raed.jafar@manara.edu.sy or raedjafar@yahoo.fr

Abstract: Ensuring safe and clean drinking water for communities is crucial, and necessitates effective tools to monitor and predict water quality due to challenges from population growth, industrial activities, and environmental pollution. This paper evaluates the performance of multiple linear regression (MLR) and nineteen machine learning (ML) models, including algorithms based on regression, decision tree, and boosting. Models include linear regression (LR), least angle regression (LAR), Bayesian ridge chain (BR), ridge regression (Ridge), k-nearest neighbor regression (K-NN), extra tree regression (ET), and extreme gradient boosting (XGBoost). The research's objective is to estimate the surface water quality of Al-Seine Lake in Lattakia governorate using the MLR and ML models. We used water quality data from the drinking water lake of Lattakia City, Syria, during years 2021–2022 to determine the water quality index (WQI). The predictive performance of both the MLR and ML models was evaluated using statistical methods such as the coefficient of determination (R^2) and the root mean square error (RMSE) to estimate their efficiency. The results indicated that the MLR model and three of the ML models, namely linear regression (LR), least angle regression (LAR), and Bayesian ridge chain (BR), performed well in predicting the WQI. The MLR model had an R^2 of 0.999 and an RMSE of 0.149, while the three ML models had an R^2 of 1.0 and an RMSE of approximately 0.0. These results support using both MLR and ML models for predicting the WQI with very high accuracy, which will contribute to improving water quality management.

Keywords: Seine lake; machine learning; water quality; water quality index; evaluation; prediction

check for updates

Citation: Jafar, R.; Awad, A.; Hatem, I.; Jafar, K.; Awad, E.; Shahrour, I. Multiple Linear Regression and Machine Learning for Predicting the Drinking Water Quality Index in Al-Seine Lake. *Smart Cities* **2023**, *6*, 2807–2827. https://doi.org/10.3390/smartcities6050126

Academic Editor: Muhammad Shafique

Received: 18 September 2023
Revised: 5 October 2023
Accepted: 10 October 2023
Published: 12 October 2023

1. Introduction

Water is a fundamental natural resource for all life forms on planet Earth. Safe water should be free from harmful chemical substances or microorganisms at concentrations that cause health problems, according to the recommendations of the World Health Organization (WHO) [1]. Rivers and lakes are considered the main sources of freshwater and represent one of the most important water resources for various uses, such as drinking, agriculture, industry, and domestic needs. They resemble lifelines for communities and play a crucial role in social, economic, and environmental development [2].

However, these water bodies are severely depleted due to excessive human activities, such as manufacturing, urbanization, and population growth. Surface water sources including rivers and lakes have been subjected to widespread pollution from various sources, according to the United Nations Environment Programme [3]. In addition, poor water resource management and climate change have caused a decline in water quality in recent decades, leading to surface water pollution [4]. The surface water quality in a region largely depends on the nature and level of various human activities in the relevant watersheds.

The chemical, physical, and biological compositions of surface water are subject to numerous effects, including natural effects such as rainfall, watershed geography, atmosphere, and geology, as well as human effects such as industrial, agricultural, and household activities [5]. Increasing surface water pollution leads to the deterioration in water quality, threatens human health, affects the balance of the aquatic ecosystem, and hinders economic development and social progress [6]. According to a report by the WHO, polluted water causes about 80% of human diseases. When groundwater is polluted, its quality can be restored by stopping the flow of pollutants from the source [7]. Therefore, it is essential to continuously monitor the quality of surface and groundwater and improve the methods and means to protect them.

The water quality index (WQI) is used to assess and summarize the overall water quality of water [7]. It takes into account various physical, chemical, and biological parameters, including temperature, pH, dissolved oxygen, turbidity, and levels of pollutants such as nutrients and contaminants. The WQI provides a numerical value or rating that helps determine the health and suitability of water for different uses, such as drinking, recreation, or aquatic life. Higher WQI values generally indicate better water quality, while lower values suggest poorer water conditions. This index helps decision makers to take effective measures to manage water resources and maintain their quality [6,8]. The formulation and use of quality indices have been strongly supported by organizations responsible for water supply and pollution control. Nevertheless, the utilization of the WQI to evaluate groundwater and surface water quality was limited for a long time due to the lack of sufficient data and appropriate statistical and modeling methods.

In recent years, machine learning (ML) techniques have been widely used to evaluate water quality, including estimating the WQI [9]. These techniques have proven powerful tools for modeling complex linear and nonlinear relationships in environmental and water resource research [10]. The application of multivariate statistical methods, such as multiple linear regression (MLR), cluster analysis (CA), principal components analysis (PCA), factor analysis (FA), and discriminant analysis (DA), is useful in reducing the complexity of large water quality data sets (reducing the number of variables) without losing the original information [11]. Applying these statistical techniques helps interpret complex data to better understand the environmental water quality status and identify potential sources or factors that affect water systems, in addition to providing a quick solution to pollution problems for simple and cost-effective water quality assessment [12].

A literature review shows that each ML algorithm has its strengths and weaknesses, and its behavior depends on the water quality input variables in different study areas [13,14]. Gupta and Gupta investigated the health status of the Damodar River in India for drinking purposes using the WQI method. They analyzed eleven water quality parameters from ten monitoring sites along the river and applied an MLR model to predict WQI. The results showed that river health varied between good and unfit categories. In addition, it identified biochemical oxygen demand (BOD), total coliform (TC), and iron (Fe) as the primary factors affecting WQI values, and the MLR model was found to be effective for evaluating river health for efficient river management. The model exhibited a strong fit, indicating a robust relationship between the identified factors and the WQI values. The results underscore the potential of the MLR model as a valuable tool for evaluating river health [15].

The WQI method was used to investigate water quality in Taihu Basin of China. The results revealed generally moderate water quality, with notable variations among the six river systems studied and distinct seasonal patterns. Through a stepwise MLR analysis, the authors developed a simplified WQImin model consisting of five key parameters ($NH_4{}^{-}N$, COD_{Mn}, $NO_3{}^{-}N$, DO, and Tur). These parameters were found to account for a significant portion of the observed variance in water quality data within the basin [16].

Nair and Vijaya [17] developed ML models for predicting and classifying the WQI of the Bhavani River. Their models showed promising results, such as an MLP regressor in prediction and an MLP classifier that achieved very good classification accuracy. These findings have implications for effective water management strategies.

Malek et al. [18] investigated the use of MLR models to predict water quality classification in the Kelantan River. Among the seven models tested, gradient boosting with a learning rate of 0.1 demonstrated superior performance. The significant variables for predicting water quality classification were total suspended solid (TSS), ammoniacal nitrogen (NH3N), biochemical oxygen demand (BOD), and chemical oxygen demand (COD). The findings contribute to improving water quality and informing water resource management policies.

Duc et al. developed ML models to predict the WQI in irrigation systems of Vietnam's Red River Delta. They utilized parameters such as BOD_5, NH_4^+, PO_4^{3-}, turbidity, TSS, coliform, and DO for calculating the WQI. The gradient boosting model showed the best performance among the others. This research demonstrates the potential of ML for efficient WQI monitoring, particularly in developing countries, while limitations in model generalization and integration of external factors affecting water quality should be considered [19]. Rezaie-Balf et al. explored the prediction of the water quality index using physicochemical parameters. The study proposes a new approach combining an ensemble Kalman filter and artificial neural network models. A novel preprocessing technique is introduced to enhance the model's performance. The results demonstrate improved accuracy in predicting the water quality index. Overall, the study provides valuable insights for efficiently analyzing water quality and aiding water quality evaluation [20].

Saber et al. utilized artificial intelligence algorithms to predict the WQI in the Illizi region of southeast Algeria. The MLR model showed higher accuracy when all parameters were considered, while the RF model performed better in scenarios with limited data. Total dissolved solids (TDS) and total hardness (TH) were identified as influential factors in the WQI. The findings contribute to improving groundwater resource management for sustainable water planning [21]. Dani et al. provides a comprehensive review of existing methods for predicting water quality, examining parameters and artificial intelligence-based models. The study compares 83 publications and highlights the superiority of hybrid deep learning (DL) models. Potential solutions to data limitations are discussed, including the use of generative adversarial networks (GANs) for synthetic data generation and attention-based transformers for time series prediction. The article offers valuable insights for researchers interested in water quality forecasting [22].

Building on this rich literature, our aim in this work is to develop a robust model that can accurately predict the quality of drinking water based on various input variables. It is based on evaluating the quality of surface water sources and reducing their pollution by studying various physical and chemical pollution parameters over two consecutive years (2021–2022) using the arithmetic weighted WQI and the MLR models, as well as a set of various machine learning models based on different algorithms. These techniques help to interpret data sets, evaluate the quality of surface water, and reduce the number and frequency of different laboratory experiments by using these models.

As a case study, we focus on evaluating laboratory analysis results for pollution parameters in Al-Seine Lake (collected from the Lattakia drinking water intake as an approved monitoring point) by comparing them with the standards of WHO and then calculating the WQI to determine their classification and level of pollution. Located in the western part of Syria, Al-Seine Lake has a significant importance in providing drinking water to the cities of Lattakia and Tartous, as well as its other agricultural and industrial uses. The National Sanitation Foundation Water Quality Index (NSFWQI) was calculated for Al-Seine Lake for the years 1991–2004–2007–2011, and the evaluation results showed that the water quality ranged from good to fair according to the adopted monitoring point [23].

This paper presents a novel contribution by conducting a comprehensive performance comparison between the MLR and various ML models, including regression, decision tree, and boosting, in predicting the water quality index for the Al-Seine Lake intake. The objective is to assist lake managers and decision makers in selecting the most effective model, simplifying laboratory work and reducing costs, efforts, and time requirements. These advanced machine learning techniques support urban water management processes, promoting sustainable and resilient smart cities.

2. Materials and Methods

This research progressed through the following steps. First, we chose the Al-Seine Lake as a study area. Second, we conducted data collection. Third, we pre-processed the data sets, excluding outliers. Fourth, we chose the parameters to be included in the analysis. Fifth, we calculated the WQI using the formulas provided below. Sixth, we applied the MLR model and evaluated the quality of this model. Seventh, we applied a set of 19 ML models and compared them to the performance of the MLR model. Eighth, we constructed a time-series model to predict the WQI over time without the use of variables and we compared it to the previous models.

2.1. Study Area

Al-Seine Lake is formed from the Al-Seine spring, which is considered one of the important and main water sources in Syria. The lake is fed by 14 springs and is located between latitudes (35°15′13″ N)–(35°15′31″ N) and longitudes (35°58′09″ E)–(35°57′59″ E); Figure 1. According to Google Earth Pro, the lake has a perimeter of 1267 m and an area of 64,337 m^2. Its water capacity is 400,000 m^3 and its maximum depth is 9 m.

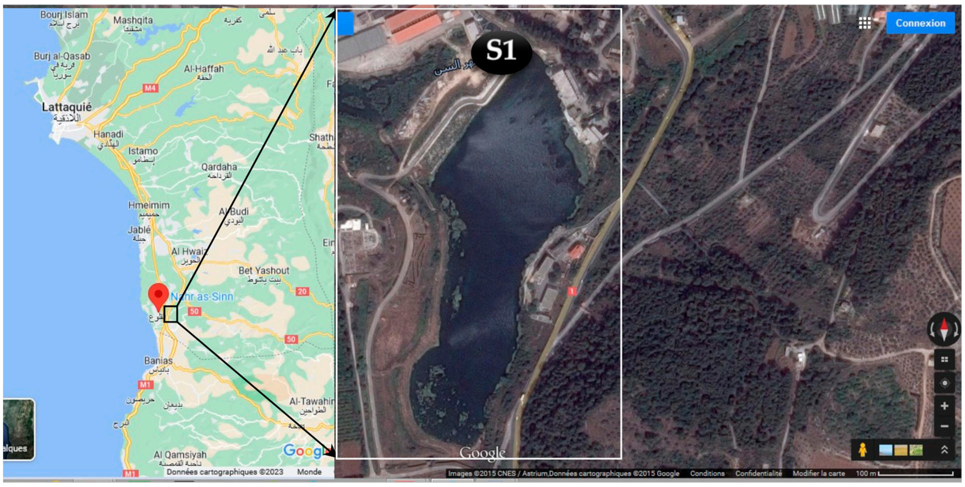


Figure 1. Location of the monitoring point for the drinking water intake in Lattakia city on Al-Seine Lake.

Al-Seine Lake supplies drinking water to the Syrian coast cities (Lattakia and Tartous) and surrounding villages, serving approximately 5% of Syria's population. The lake's water resources, including the Al-Seine spring, contribute about 60% of the drinking water in the coastal basin, with an average pumping rate of 13,500 m^3/h to consumption areas. Apart from drinking water, the lake is also utilized for irrigation and industrial purposes. The lake's significance is heightened by the growing water demand resulting from population growth, agricultural expansion, and industrial development in the Syrian coast cities and countryside. Efforts are currently underway to explore the possibility of utilizing the water of Al-Seine Lake to provide drinking water to additional cities. However, the presence of residential areas close to the lake and scattered throughout its basin, in addition to neighboring agricultural lands, contributes to the deterioration of its water quality, as well as the wells and springs in the nearby rural areas located in the Al-Seine basin.

2.2. Sample Collection and Analysis

This study is based on data collected from Lattakia's drinking water intake over 2021–2022 through the Al-Seine Lake monitoring and protection program. This included

530 measurements at a daily rate. Data preprocessing involved cleaning and removing any missing values and outliers. The water quality index (WQI) was calculated using the weighted arithmetic method. A multiple linear regression (MLR) model was developed to incorporate the significant variables in evaluating the intake water quality. Additionally, various machine learning (ML) models that utilized regression, decision tree, and boosting algorithms were applied to predict the WQI for Lattakia's water intake. The samples were specifically collected from the Lattakia water intake monitoring point S1, situated at latitude 35°15′31″ N, longitude 035°58′04″ E, and elevation of 19 m; Figure 1. This selection was based on the approved monitoring program, ensuring that the samples were representative of the intake water quality.

In the laboratory, comprehensive tests were conducted to determine various water quality parameters. These included pH, sulfates ($SO_4{}^{-2}$), nitrates ($NO_3{}^{-}$), nitrites ($NO_2{}^{-}$), ammonium ($NH_4{}^{+}$), phosphates ($PO_4{}^{-3}$), turbidity (Tur), and electrical conductivity (EC). The analysis of these parameters provided detailed insights into the chemical composition, clarity, and electrical characteristics of the water, contributing to a comprehensive understanding of the water quality at the Lattakia intake.

Table 1 shows the descriptive statistical analytical indicators for the measured and research-based data. Figure 2 illustrates the correlation matrix between all studied variables, including the calculated water quality index (WQI).

Table 1. Descriptive statistics for physicochemical parameters of water samples.

Descriptive Statistics					
	N	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
WQI	530	15.969	69.499	23.6427	4.24165
pH	530	7.00	08.90	7.7704	0.17769
$SO_4{}^{-2}$ (mg/L)	530	6.00	18.00	11.0811	2.19447
$NO_3{}^{-}$ (mg/L)	530	0.66	12.00	3.3711	0.89827
$NO_2{}^{-}$ (mg/L)	530	0.00	03.20	0.0212	0.13915
$NH_4{}^{+}$ (mg/L)	530	0.00	0.05	0.0023	0.00592
$PO_4{}^{-3}$ (mg/L)	530	0.01	0.90	0.2053	0.09166
Turbidity (NTU)	530	0.38	09.94	1.9497	1.00350
EC (µS/cm)	530	445.00	503.00	477.7830	7.92975

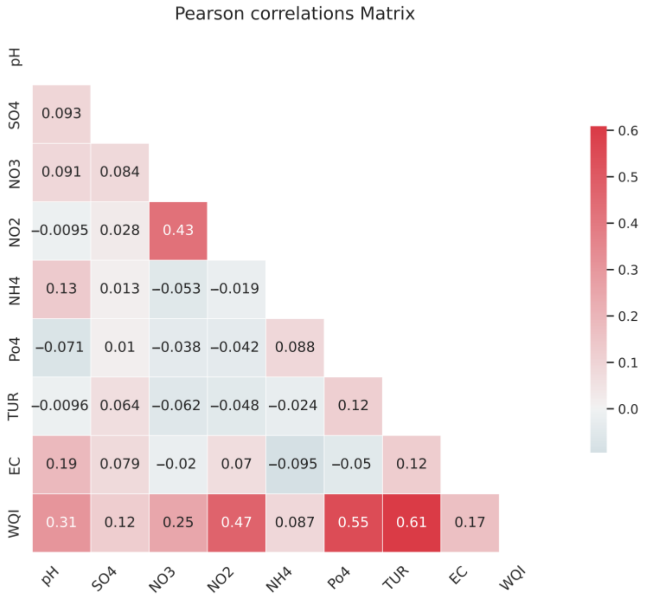


Figure 2. Correlation matrix for the studied variables.

2.3. Water Quality Index (WQI)

The WQI was determined using eight physicochemical: pH, $SO_4{}^{-2}$, $NO_3{}^{-}$, $NO_2{}^{-}$, $NH_4{}^{+}$, $PO_4{}^{-3}$, turbidity, and EC. The WQI was calculated as follows:

1—Assigning each parameter of the eight previously mentioned parameters a weight ranging from 1 to 5 depending on its importance and its role in affecting water quality, based on expert opinions in previous reference studies.

The standard values, assigned weights, and relative weights of the eight parameters are shown in Table 2.

Table 2. Permissible standard values, assigned weights, and relative weights of the study parameters.

Parameter	Water Quality Stander (Si)	Assigned Weight (wi)	Relative Weight (Wi)
pH	6.5–8.5	4	0.1212
SO_4^{-2} (mg/L)	250	4	0.1212
NO_3^- (mg/L)	50	5	0.1515
NO_2^- (mg/L)	1	5	0.1515
NH_4^+ (mg/L)	0.50	3	0.0910
PO_4^{-3} (mg/L)	0.50	4	0.1212
Turbidity (NTU)	5	4	0.1212
EC (µS/cm)	1000	4	0.1212
Total		$\sum \text{wi} = 33$	$\sum \text{Wi} = 1$

The relative weight was calculated using the following equation:

$$\text{Wi} = \frac{\text{wi}}{\sum_{i=1}^{n} \text{wi}} \tag{1}$$

where:

Wi: The relative weight.

wi: The weight assigned to the parameter.

n: The number of parameters.

2—For each parameter, a quality assessment scale (qi) was calculated by dividing the laboratory measurement values of the pollutant concentrations by the standard values according to the World Health Organization (WHO). The result was multiplied by 100 using the following equation:

$$q_i = \frac{\text{Ci}}{\text{Si}} \times 100 \tag{2}$$

Note that for the pH and dissolved oxygen parameters, the quality assessment scale (qi) is calculated using the following equation:

$$q_i\ \text{Do, pH} = \frac{(\text{Ci} - \text{Vi})}{(\text{Si} - \text{Vi})} \times 100 \tag{3}$$

where:

Ci: The measured value of the water quality parameter.

Vi: The ideal values for dissolved oxygen (i.e., 14.7) and pH (i.e., 7).

Si: The standard value for the water quality parameter.

3—The assigned weight is multiplied by the relative weight to obtain the sub-indices Sli. The WQI is the sum of the sub-indices according to the following equation:

$$\text{Sli} = \text{Wi} \times q_i \tag{4}$$

$$\text{WQI} = \sum_{i=1}^{n} \text{Sli} \tag{5}$$

The WQI values are classified according to the scale proposed in previous studies, Table 3.

Table 3. Water quality index scale [24].

Water Quality Index	0–25	26–50	51–75	76–100	Above 100
Water Quality	Excellent	Good	Fair	Poor	Very Poor

2.4. Multiple Linear Regression (MLR)

Multiple linear regression is a statistical modeling technique used to model the relationship between a dependent variable (the variable being predicted or explained) and two or more independent variables (predictor variables). The goal of multiple linear regression is to find the best linear equation that can predict the value of the dependent variable based on the values of the independent variables. The relationship is expressed in terms of weights assigned to the independent variables, which allows for predictions to be made about the dependent variable.

The general equation for multiple linear regression (Equation (6)) can be written as:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn + \varepsilon \tag{6}$$

where Y is the dependent variable, X1, X2, . . ., Xn are the independent variables, β0 is the intercept, and β1, β2, . . ., βn are the regression coefficients (also known as the slope coefficients); these coefficient values signify both the strength and direction of the relationship. ε is the error term, which represents the variability in the dependent variable that is not explained by the independent variables.

The multiple linear regression model can be used for prediction, as well as for understanding the relationships between the dependent and independent variables. The model can also be used for hypothesis testing and model selection. Before building the MLR model, multi-collinearity diagnostics, cross-validation, or regularization procedures are implemented to prevent unstable results.

Using multiple linear regression (MLR) for WQI modeling involves analyzing several input parameters related to the water physicochemical characteristics and producing a single index score reflecting overall water quality. During the model building, we constructed and ran a series of MLR models based on chosen features (variables) covering data sets of different sizes. We split data into training and testing subsets to verify model performance before making the final decision of its utility.

2.5. Machine Learning Models (ML)

Machine learning (ML) is an aspect of artificial intelligence (AI) that enables machines to learn and adapt to data instead of relying on explicit programming. ML techniques include models such as artificial neural networks (ANNs), principal component analysis (PCA), and random forests (RFs). These models vary in their learning approaches and require expertise in mathematics, programming, and statistics, and domain-specific knowledge. Such techniques are used in many fields, including data analysis, speech and image recognition, prediction and classification, robot control, and improving game performance. They are also used in water quality management, to analyze real-time data, improve monitoring water quality, and analyze and predict the current and future water quality resulting from several influencing factors, such as acidity, turbidity, salts, nutrients, and pollutants.

3. Results and Discussion

3.1. Evaluation of Water Quality Parameters

The samples were analyzed for important pollution parameters, including pH, $SO_4{}^{-2}$, $NO_3{}^{-}$, $NO_2{}^{-}$, $NH_4{}^{+}$, $PO_4{}^{-3}$, turbidity, and EC, as follows:

3.1.1. pH

pH is an important indicator of drinking water quality as it helps to evaluate the suitability of water for drinking and other uses. The pH should range between 6.5 and 8.5 according to World Health Organization standards to ensure drinking water quality [25]. If the pH is below 6.5, the water is acidic and can cause corrosion of pipes and equipment used in water transport and distribution. If the pH is above 8.5, the water is alkaline and can cause calcium and magnesium deposition on surfaces and pipes.

The results shown in Figure 3 indicate the relative fluctuation of the concentration of H^+ ions in the water, ranging from 7 to 8.9, with all samples falling within the alkaline range. The alkaline values are due to the presence of bicarbonate ions, while the relative decrease in values may be due to an increase in salt concentration and dominance of the chloride and sulfate phase over the bicarbonate phase, leading to a slight decrease in pH towards acidity. Overall, the studied water samples are within the suitable alkaline limits for drinking, except for only two values that exceeded the permissible limit, as shown in Figure 3.

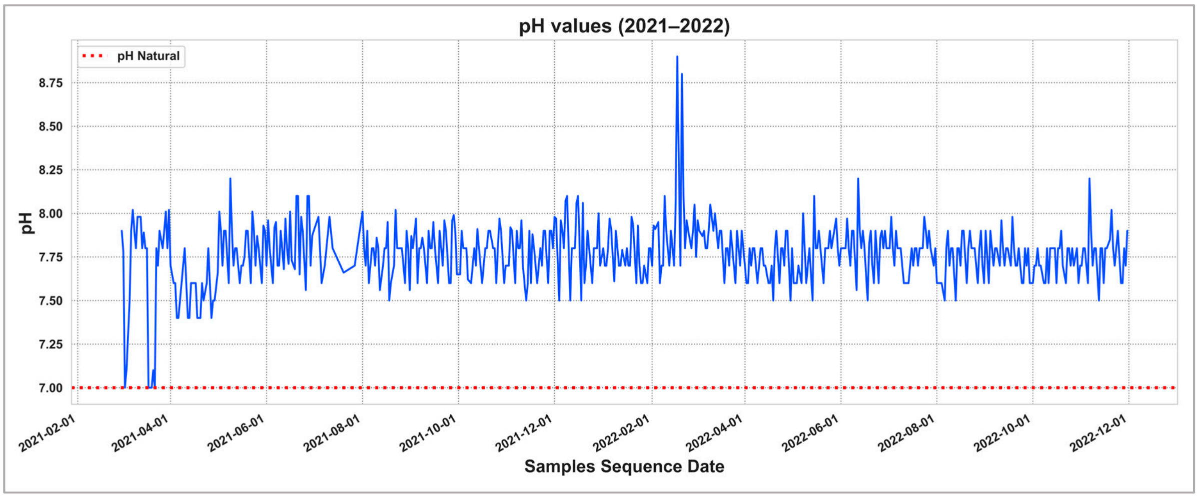


Figure 3. Variations in pH values in the Lattakia drinking water intake during the study period (2021–2022).

3.1.2. Sulfate ($SO_4{}^{-2}$)

Sulfates occur naturally in some sources of groundwater and surface water, and they do not pose a health risk to humans when present at low levels. However, excessive use of sulfite-containing chemicals in industry or agriculture can pollute water sources and increase sulfate concentrations to levels that pose a health risk. Sulfates are salts containing sulfur and are primarily formed through the biological oxidation of organic matter in surface and groundwater, as well as the oxidation of minerals in soil and rocks surrounding water sources [26]. Organic matter contributing to sulfate formation includes plant and animal matter, human and animal waste, and anaerobic bacteria.

Purification processes for surface and groundwater can remove some sulfates, but they cannot eliminate them, meaning most drinking water will contain certain levels of sulfates. To assess the presence of sulfates in drinking water, their concentration is reviewed and compared to allowable limits and approved health standards related to water quality. For example, according to the World Health Organization, the safe concentration of sulfates in drinking water should be less than 500 milligrams per liter.

The study results shown in Figure 4 indicate that the concentration value of sulfate ions ranged from 6 to 18 mg/L, which is within the allowable values according to the World Health Organization. This very low concentration may be due to the geological formation of the basin area, where sulfate ion concentrations in groundwater are influenced by the type of rocks and the period they have been exposed to. Gypsum and anhydrite (calcium sulfate) are the main sources of sulfates, while sandstone with very little limestone content is considered a minor source. In addition, bacterial activities in soil layers play an important role in oxidation-reduction reactions of sulfur species, and sulfate concentration in groundwater often increases with depth due to the increased likelihood of dissolution of adjacent rocks. Furthermore, anaerobic degradation of sulfur-containing compounds, gypsum, and anhydrite is the main process of geological formation, in which sulfur is reduced by *Thiobacillus desulfuricons* bacteria to form sulfur compounds that are oxidized to sulfuric acid under aerobic conditions.

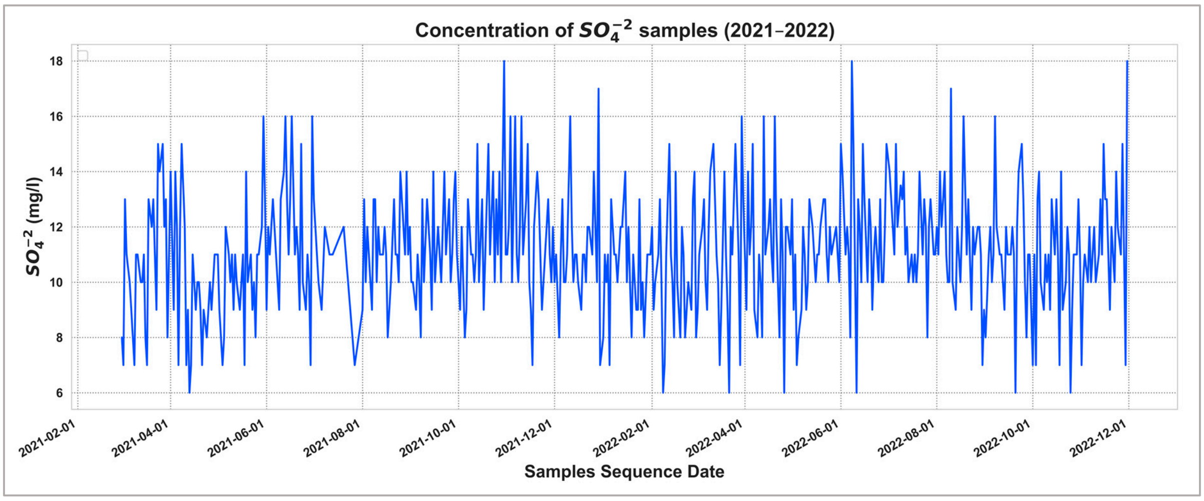


Figure 4. Variations in sulfate values $\mathbf{SO_4}^{-2}$ (mg/L) in the Lattakia drinking water intake during the study period (2021–2022).

3.1.3. Nitrate (NO_3^-)

Nitrates are one of the most widespread pollutants in groundwater, and a major problem in some shallow groundwater layers. The primary sources of nitrates in groundwater are sewage and industrial wastewater, plant residues, and animal waste, as well as chemical fertilizers and herbicides used in agricultural activities.

Controlling the concentration of nitrates in drinking water is essential due to the potential health problems associated with increased levels. Nitrates are known carcinogens, and when consumed by humans, they can be converted into nitrite (NO_2^-) within the intestines. This conversion affects the ability of red blood cells to carry oxygen, leading to a condition called methemoglobinemia. Infants and pregnant women are particularly vulnerable, and the condition is commonly known as blue baby syndrome [27]. According to WHO standards, the maximum permissible level of nitrates in drinking water is 45 mg/L.

We represent the laboratory analysis results for daily measurements of nitrate concentrations in water samples taken from the Lattakia city intake over the two years (2021–2022) in Figure 5.

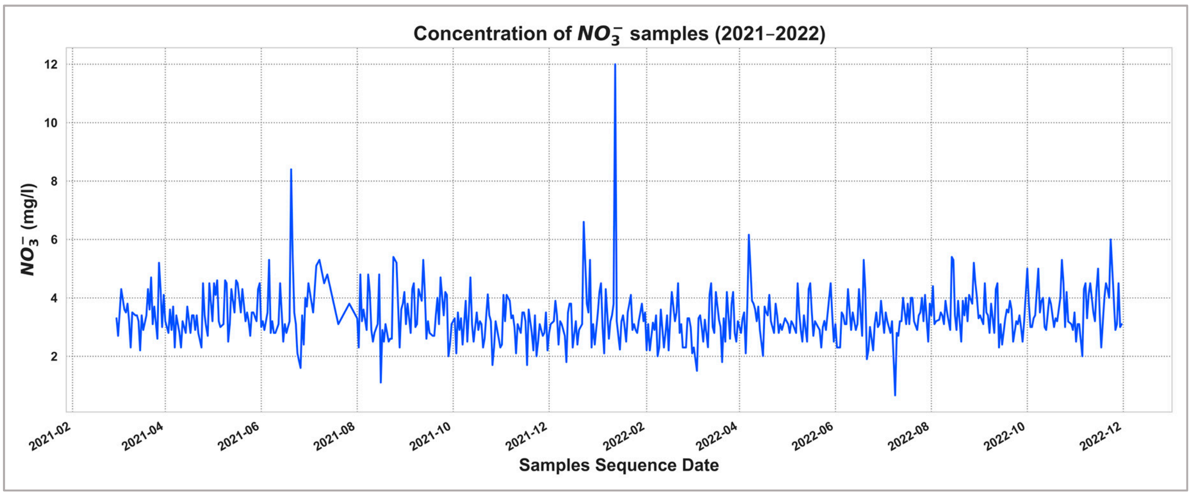


Figure 5. Changes in nitrate (NO_3^-) concentrations (mg/L) in Lattakia drinking water intake during the study period (2021–2022).

Based on the laboratory analysis results and graphical representations, several observations can be made. The range of nitrate concentration in the studied water samples fell within the allowable limit for drinking water according to the WHO standards, which is 45 mg/L. Additionally, the highest concentration of nitrate (12 mg/L) was recorded during the winter season, as indicated in Figure 5. This is thought to be due to the role of rainfall in transporting fertilizer residues from soil and agricultural lands into the groundwater reservoir. The lowest recorded value for nitrate concentration (0.66 mg/L) was observed in July 2022.

3.1.4. Nitrite (NO_2^-)

Nitrite is usually formed in drinking water when the water is contaminated from various sources such as agricultural fertilizers, animal waste, wastewater, and chemical industries. Nitrite can also naturally form in groundwater and surface water when water reacts with nitrates in the soil.

Several mechanisms can lead to the formation of nitrite in drinking water, including:

- Agricultural pollution: Nitrates, ammonium, and urea are added to the soil as nitrogen fertilizers for crops, and water can seep into the groundwater containing nitrates and nitrites. Nitrates in agricultural fertilizers can also decompose and turn into nitrites through bacterial processes.
- Industrial pollution: Chemicals in raw materials or used in industry can leak into the soil and groundwater, and then into surface water sources. Some industries using nitrates, such as fertilizers, insecticides, and other chemicals, can contribute to nitrite pollution in surface and groundwater.
- Sewage: Sewage and animal waste from animal and poultry facilities can seep into groundwater and surface water sources, leading to increased concentrations of nitrates and nitrites.
- Air pollution: Rain, snow, and airborne spray can carry industrial and agricultural pollutants into water sources, leading to increased concentrations of nitrates and nitrites in surface water.
- Drug pollution: Some drugs can seep into the soil and groundwater from various sources and may be found in surface and groundwater sources.

The level of nitrite in water is measured in specialized laboratories using various methods, including the colorimetric method (which depends on the reaction of nitrite with chemicals to produce a measurable color). The nitrite concentration in drinking water should not exceed the allowable limit according to the World Health Organization standards, which is 1 mg/L [25].

The results of the study shown in Figure A1 indicate that the nitrite ion concentration ranged from 0 to 0.13 mg/L, which is within the allowable values according to the World Health Organization, except for a single value in January 2022, which reached 3.2 mg/L. This low concentration may be due to strict measures regarding human activities allowed in the buffer area of the Al-Seine basin.

3.1.5. Ammonium NH_4^+

Ammonia is typically formed in drinking water due to the natural decomposition of organic, plant, and animal matter in the aquatic environment. Ammonia can also form because of the leakage of animal waste, fertilizer, and industrial waste into groundwater and surface water.

In general, the presence of ammonia in drinking water can be evaluated by analyzing the water and measuring the level of ammonia in it. It is preferable to maintain low levels of ammonia in drinking water, as high levels of ammonia can indicate the presence of contamination in the water with organic matter that can be decomposed. This represents a health hazard to water users.

In general, the concentration of NH_4^+ in drinking water can be measured by conducting chemical tests. This is usually undertaken by adding a solution of chlorine to the water sample and then measuring the level of ammonia in the water after a specified period. The

NH_4^+ level in the water should be less than 0.5 parts per million (ppm) according to the World Health Organization's standards for drinking water.

The study results during the years 2021–2022 shown in Figure A2 indicate that the concentration of the ammonium ion ranged from 0 to 0.05 mg/l and all values were within the allowable limits according to the World Health Organization [25].

3.1.6. Phosphate PO_4^{-3}

Phosphate compounds are produced by the combination of phosphorous and oxygen compounds. Organic phosphates are produced by the decomposition of plant and animal residues, as well as the breakdown of waste and food remnants. Inorganic phosphates, on the other hand, are mainly sourced from fertilizers used for agricultural purposes, as well as soap manufacturing and domestic wastewater. Therefore, measuring the concentration of phosphate in drinking water and natural water sources is an important indicator when studying water source quality [28]. According to World Health Organization (WHO) standards, the maximum allowed limit for phosphate ion concentration in drinking water is 0.5 mg/L. Figure 6 shows the laboratory analysis results of measured changes in phosphate ion concentration values in water samples collected from the Lattakia drinking water intake over the two years (2021–2022).

Based on the laboratory analysis results and graph, several observations can be made regarding the phosphate ion concentration in the studied water samples. The measured values ranged from 0.01 mg/L to 0.9 mg/L, with the lowest concentration observed in July 2022 and the highest in December. Although most of the measured values fell within the allowed limit for drinking water of 0.5 mg/L, five points exceeded this limit. These points were recorded in October 2021, December 2021, and April 2021 and 2022. It can also be seen from Figure 6 that all the points exceeding the limit were recorded during the winter season. This can be attributed to rainfall events during this season, which wash away fertilizer and pesticide residues from soil and agricultural lands and transport runoff rich in pollutants, including phosphate ions, from nearby catchment basins to the groundwater reservoir.

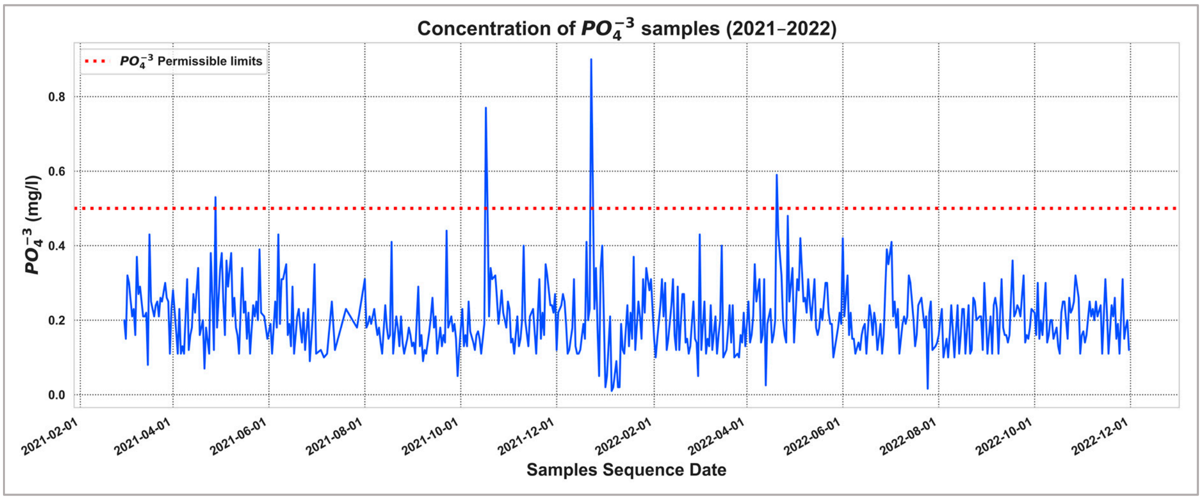


Figure 6. Variations in the concentrations of phosphate (PO_4^{-3}) (mg/L) in the Lattakia drinking water intake during the study period (2021–2022).

3.1.7. Turbidity

Turbidity in drinking water is usually caused by the presence of very small, suspended particles, microorganisms, and dissolved organic and inorganic matter that are difficult to see with the naked eye. These particles and organic matter can originate from various sources such as soil, leaves, and industrial and animal waste [23]. Turbidity is affected by several factors, including the water source, water flow, temperature, storage duration, nature of suspended matter in the water, and level of biological and chemical pollution.

Turbidity can affect the taste, odor, and overall appearance of the water, and in some cases, it can cause health problems. Therefore, the allowable limits for turbidity concentration in drinking water are determined according to international and local quality standards. The maximum allowable limit according to the World Health Organization (WHO) standards is 5 NTU, where turbidity level depends on the size and number of suspended particles and solid matter in the water.

Several methods can be used to reduce turbidity in water, such as filtration, reverse osmosis (RO), and removal of suspended particles by adding materials that help coagulate and remove them.

The study results during the two years (2021–2022), as shown in Figure A3, indicate that the values of turbidity ranged from 0.38 to 9.94 NTU, all of which were within the allowable values according to the WHO, except for eleven values that were recorded during the rainy months.

3.1.8. Electrical Conductivity (EC)

Electrical conductivity measurement in drinking water provides an idea about the quantity of dissolved substances in the water, especially mineral salts that dissolve in water to form ions. Electrical conductivity is affected by the concentration of these ions.

When chemicals dissolve in water, they are separated into negative and positive ion particles, which move in the water to achieve electrical balance. Therefore, electrical conductivity in water increases as the concentration of dissolved salts increases. Electrical conductivity is used to evaluate the quality of drinking water, where high values of electrical conductivity indicate the presence of large amounts of mineral salts in the water, and vice versa, indicating that the water is not suitable for drinking.

Electrical conductivity is measured using a conductivity meter, and the resulting value represents the number of positive and negative ions in the water. The value is measured in Siemens units, which indicate the amount of electrical current flowing through the water between poles at a certain temperature and pressure.

The study results during the two years (2021–2022), as shown in Figure A4, indicate that the concentration values of electrical conductivity ranged from 445 to 503 µS/cm, all of which were below the allowable value according to the WHO, which is 1000 µS/cm [25].

Figure 7 shows the results of applying the water quality index (WQI) to the Lattakia drinking water intake for the period 2021–2022. These results indicate that the quality of the intake water fell within the excellent classification for 70.4% of the observations and within the good classification for 29.2% of the observations, while we had only two values (0.004%) that fell within the poor classification.

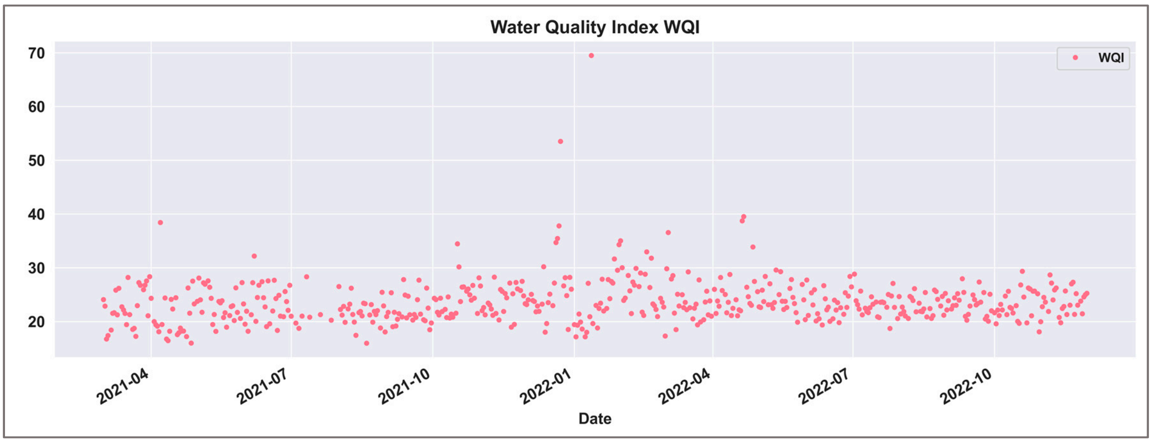


Figure 7. Changes in water quality index (WQI) values in the Lattakia drinking water intake during the study period (2021–2022).

3.2. Multiple Linear Regression (MLR)

To determine the relationship between the water quality index (WQI) in the Lattakia drinking water intake and the explanatory variables used in the study (pH, sulfates, nitrates, nitrites, ammonium, phosphates, turbidity, and electrical conductivity), a multiple linear regression model was developed with the variables listed in Table 4, where the previous variables were considered as independent explanatory variables and the WQI variable was considered as the dependent variable.

Table 4. Results of the multiple linear regression analysis model.

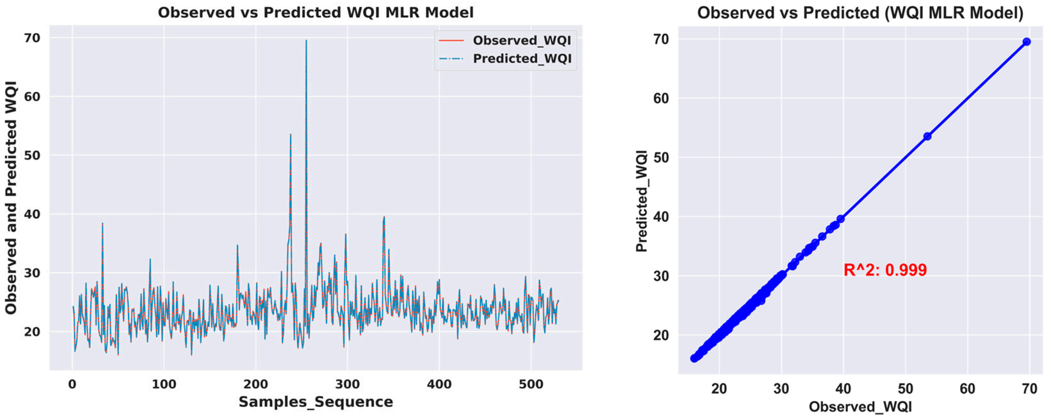
MLR Model	B	T	Sig.	Correlation Coefficient R	Determination Coefficient R^2	VIF Variance Inflation Factor	F	Sig.
(Constant)	−56.751	−124.833	0.00	0.999	0.999		69,855.695	0.00
pH	8.227	217.177	0.00			1.057		
NO_3^-	0.304	37.335	0.00			1.244		
NO_2^-	15.167	289.537	0.00			1.239		
PO_4^-	24.370	337.218	0.00			1.024		
Turbidity	2.428	365.470	0.00			1.037		
EC	0.011	13.183	0.00			1.070		

The results of the regression model in Table 4 showed that the regression model was significant, with an F-test value of 69,855.695 and a statistical significance (0.00) smaller than the significance level (0.01). This indicates the quality of the relationship model and the reliability of relying on the model results without errors. The results also indicate that the independent explanatory variables explain 99% of the variance in the water quality index (WQI), based on the coefficient of determination (R^2). The beta values (B), which indicate the relationship between WQI and the explanatory variables, were all statistically significant, as shown by the t-values and associated functions. For example, the beta value that indicates the relationship between WQI and turbidity was 2.43, and it was statistically significant (0.00). This means that if the turbidity value improved by one unit, the WQI level would improve by 2.43 units; the same interpretation applies to the other variables. Table 4 also shows the results of the multicollinearity test, which revealed that all the variance inflation factors (VIFs) were less than three (VIF < 3), indicating no problem of multicollinearity between the model variables. Therefore, we can formulate the regression equation (Equation (7)), which consists of the beta values (B) for each of the previous independent variables and the constant value, as follows:

$$\begin{aligned}\text{Water Quality Index (WQI)} = {} & -56.751 + 8.227\,(\text{pH}) + 0.304\,(NO_3) + 15.167\,(NO_2) + 24.370\,(PO_4) \\ & + 2.428\,(\text{Turbidity}) + 0.011\,(\text{EC})\end{aligned} \tag{7}$$

Figure 8a shows the mathematical relationship between the measured values and the estimated values by applying the MLR model to the data matrix. Figure 8b shows the model's performance through the coefficient of determination (R^2).

Based on the highly performing multiple linear regression model (MLR), we have designed and programmed a graphical user interface (GUI) tool to calculate the water quality index for practical and simple use by decision makers and stakeholders responsible for monitoring and controlling water quality. The tool allows entering the values of the model variables resulting from laboratory analysis and immediately determining the drinking water quality index for the studied intake in Al-Seine Lake, as shown in Figure A5.



(**a**) (**b**)

Figure 8. (**a**) Observed vs. predicted values of WQI samples sequence—MLR model. (**b**) MLR model efficiency.

3.3. Machine Learning Models (ML)

As an important first step in building an ML model, enough input variables must be chosen that contain sufficient basic information to predict the WQI. In addition, this selection can improve the model's accuracy by avoiding unwanted influence on predictive performance. In the current study, eight water quality variables were identified as potential inputs. There are several current methods for evaluating input sets, including self-correlation, partial self-correlation, mutual correlation, and the correlation coefficient. From these methods, the correlation coefficient was used in this study because of its accuracy and efficiency [4]. The correlation matrix in Figure 2 shows that the WQI variable is correlated with turbidity, followed by $PO_4{}^{-3}$, $NO_2{}^{-}$, pH, $NO_3{}^{-}$, and EC, respectively. We can neglect the $NH_4{}^{+}$ and $SO_4{}^{-2}$ variables due to their weak correlation with WQI.

After selecting the input WQI variables, 19 ML models were built and compared to predict water quality index, as shown in Table 5, where the database was divided into 70% for training and 30% for testing, and the performance of these models was evaluated during testing [29]. In this study, two packages, Scikit-Learn and PyCaret, were used, built in a programming environment using Python, to develop nineteen ML models for predicting WQI. The root mean square error (*RMSE*) and coefficient of determination (R^2) were used as typical model efficiency statistics to evaluate the performance of ML models and measure the quality of compatibility between predicted and measured values. *RMSE* (Equation (8)) measures the deviation between observed and predicted values, while R^2 (Equation (9)) measures the degree of correlation between observed and predicted data.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{n}} \tag{8}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2} \tag{9}$$

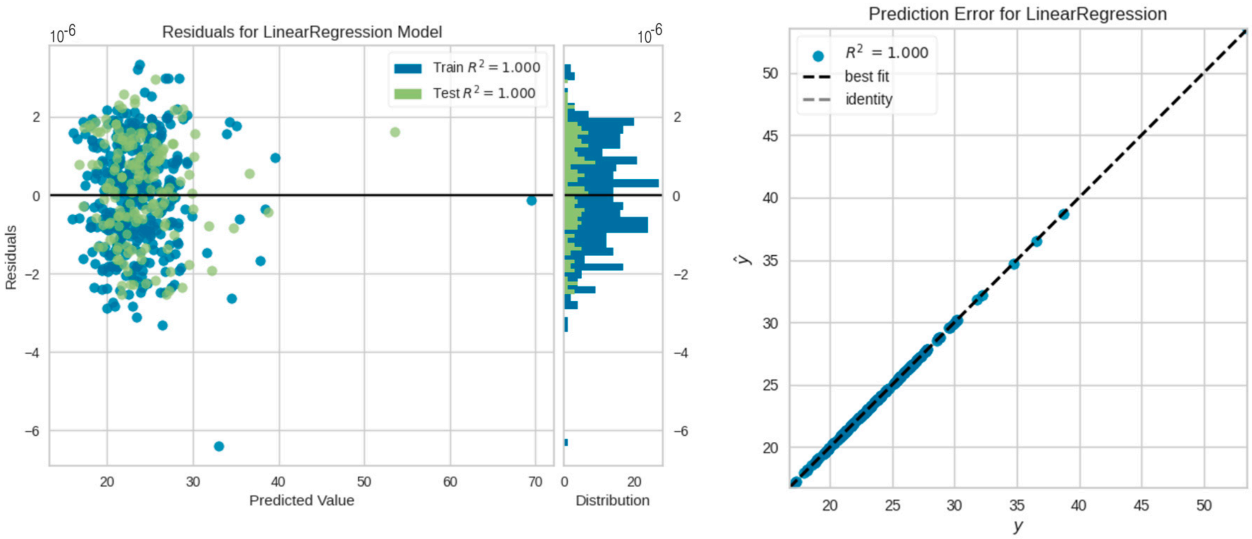
where n is the total number of predicted values, O_i is the observed value, $\overline{O}$ is the average of observed values, and P_i is the predicted value.

The results of machine learning (ML) models for predicting the quality of drinking water at the Lattakia intake shown in Table 5, based on the studied pollution variables during the period 2021–2022, indicate that the three ML models (linear regression (LR), least angle regression (LAR), and Bayesian ridge (BR)) achieved the best performance, with a correlation value between measured and predicted values of 100%.

Table 5. Comparison of machine learning (ML) model performance in calculating the drinking water quality index for the Lattakia intake (ranked by performance).

Model		MAE	MSE	RMSE	R2	RMSLE	MAPE
Ir	Linear Regression	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
lar	Least Angle Regression	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
br	Bayesian Ridge	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
ridge	Ridge Regression	0.6266	5.7816	1.3336	0.8822	0.0406	0.0235
et	Extra Trees Regressor	0.6691	7.7039	1.6521	0.8227	0.0514	0.0240
gbr	Gradient Boosting Regressor	0.6713	7.7261	1.6444	0.8216	0.0505	0.0240
xgboost	Extreme Gradient Boosting	0.6906	7.8954	1.7136	0.8092	0.0531	0.0249
rf	Random Forest Regressor	0.8350	8.2094	1.8274	0.7900	0.0589	0.0312
huber	Huber Regressor	1.2459	10.5125	2.1805	0.6997	0.0772	0.0492
dt	Decision Tree Regressor	1.2773	9.9978	2.3195	0.6650	0.0782	0.0501
lightgbm	Light Gradient Boosting Machine	1.0778	9.8756	2.3649	0.6560	0.0762	0.0403
ada	AdaBoost Regressor	1.4226	10.1218	2.4847	0.6166	0.0871	0.0579
lasso	Lasso Regression	2.2606	13.7407	3.3067	0.2884	0.1212	0.0941
llar	Lasso Least Angle Regression	2.2606	13.7407	3.3067	0.2884	0.1212	0.0941
en	Elastic Net	2.3091	13.9869	3.3621	0.2580	0.1231	0.0961
knn	K Neighbors Regressor	2.4627	14.7339	3.5339	0.1645	0.1295	0.1011
omp	Orthogonal Matching Pursuit	2.7662	17.7386	3.9547	−0.0604	0.1456	0.1154
dummy	Dummy Regressor	2.7546	17.9070	3.9646	−0.0620	0.1462	0.1152
par	Passive Aggressive Regressor	3.6783	28.3909	5.0279	−0.9922	0.1892	0.1473

Figure 9a illustrates the optimal performance of the machine learning model in the case study of linear regression (LR) through residual error and its distribution between measured and predicted values by applying this model to the data matrix of the training and testing phase. Figure 9b shows the model's performance strength through the value of the coefficient of determination (R^2).



(**a**) (**b**)

Figure 9. (**a**) The best performance of the machine learning model (linear regression (LR)) (residual error). (**b**) The best performance of the machine learning model (Linear regression (LR)) (coefficient of determination).

Machine learning (ML) models can effectively handle time series data analysis; therefore, they are a good option for analyzing such data. This is due to their ability to deal with data related to time order and analyze it accurately and efficiently. They are usually used in time series data analysis to predict future values or analyze temporal behavior of data. They can be used to analyze temporal patterns in productivity, sales, the environment, or any other type of time series data.

For example, the XGBoost model can be trained using available time series data saved in chronological order. After training, the model can be used to predict future values based on available historical data.

The K-Neighbors regressor can also be used successfully for time series data analysis, but it requires a good understanding of time series data, statistical analysis methods, and machine learning techniques. It is also important to verify the quality of the data and ensure that there are sufficient data for training and analysis. This algorithm falls within supervised learning techniques and is used to solve regression and classification problems. It is a non-parametric algorithm, which means that it does not make any assumptions about the underlying data and seeks to better fit the training data in building the function, while maintaining some ability to generalize based on the unseen data. Hence, it can fit a large number of functions.

Therefore, we applied nineteen ML models to the available time series during the study period (2021–2022) and we verified the ability of these models to predict the quality of drinking water at the Lattakia intake. Regardless of the pollution variable values (which did not enter into building the models), the K-Neighbors regressor (K-NN) model achieved the best performance with a determination coefficient value of 34% between measured and expected values of the water quality index, and an RMSE value of 4.62; Figure 10a. Figure 10b shows the model's performance strength through the value of the coefficient of determination (R^2).

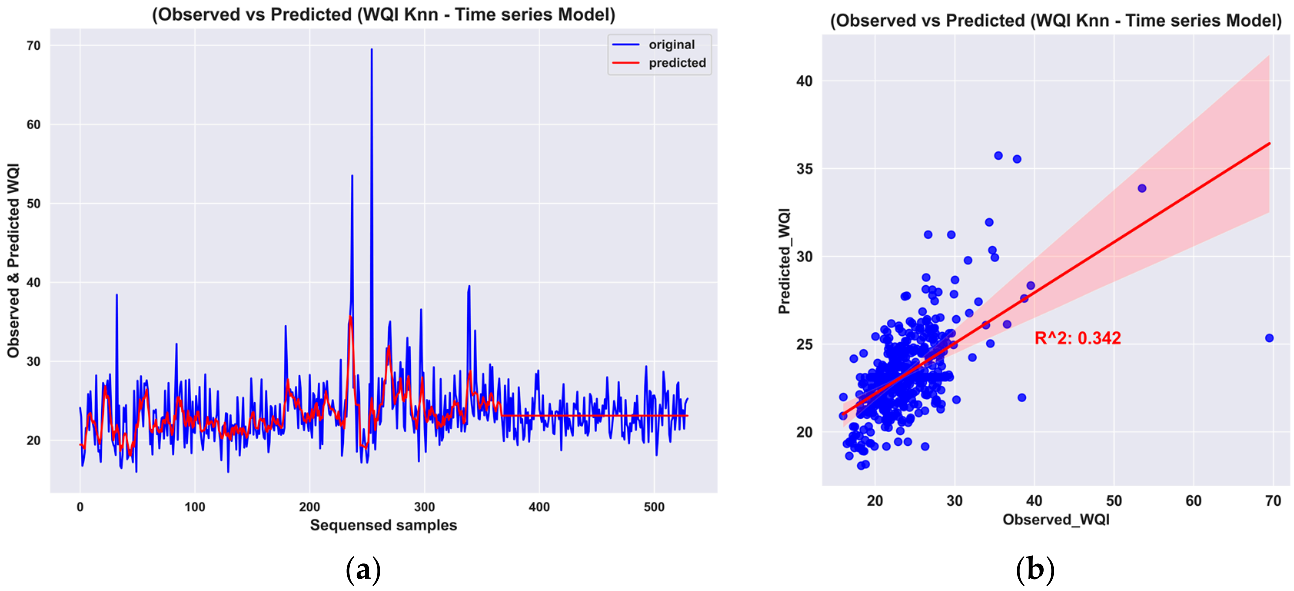


Figure 10. (**a**) Observed vs. predicted values of WQI samples sequence—time series model (K-NN). (**b**) K-NN time series model efficiency.

4. Conclusions

This research explored the capacity of multiple linear regression (MLR) and machine learning (ML) techniques to predict surface water quality. Analyses were conducted on data collected from the intake of Lattakia city in Al-Seine Lake over 2021–2022. Data included the following water quality parameters: pH, sulfates (SO_4^{-2}), nitrates (NO_3^{-}), nitrites (NO_2), ammonium (NH_4^{+}), phosphates (PO_4^{-3}), turbidity (Tur), and electrical conductivity (EC). The research findings demonstrated the excellent performance of MLR in accurately predicting the WQI and identifying influential factors, achieving excellent results by explaining 99% of the WQI variations. Furthermore, a graphical user interface (GUI) was developed to facilitate the utilization of the MLR model by decision makers and water quality monitoring personnel. Tests conducted with 19 ML models showed that the best performances were provided by the linear regression (LR), least angle regression (LAR), and Bayesian ridge (BR) models, with a correlation value of 100% between measured and predicted values. The K-Neighbors regressor (K-NN) model performed the best for time series data, with a coefficient of determination of 34% and a root mean square error of 4.62. This research demonstrates the effectiveness of using ML techniques to manage surface water in Al-Seine Lake and shows their high capacity to predict the water quality index.

Author Contributions: Conceptualization, R.J., A.A., K.J. and I.S.; methodology, R.J., A.A., I.H., K.J., E.A. and I.S.; software, R.J., K.J., I.H. and E.A.; validation, R.J., A.A., K.J. and I.S.; formal analysis, R.J., A.A., K.J., I.H., E.A. and I.S.; data curation, R.J. and K.J.; writing—original draft preparation, R.J., A.A., K.J., I.H., E.A. and I.S.; writing—review and editing, R.J., A.A., K.J., E.A. and I.S.; visualization, R.J., K.J. and I.H.; supervision, A.A. and I.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability Statement: The data sets are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

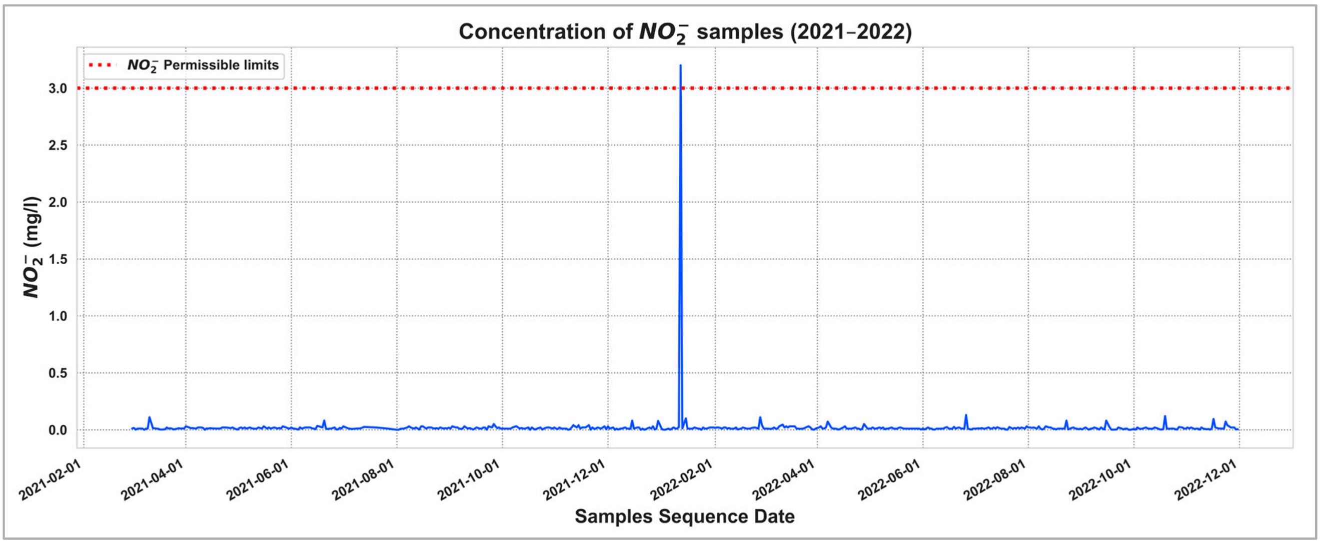


Figure A1. Variations in nitrite ion ($NO_2{}^-$) concentration values (mg/L) in the Lattakia drinking water intake during the study period (2021–2022).

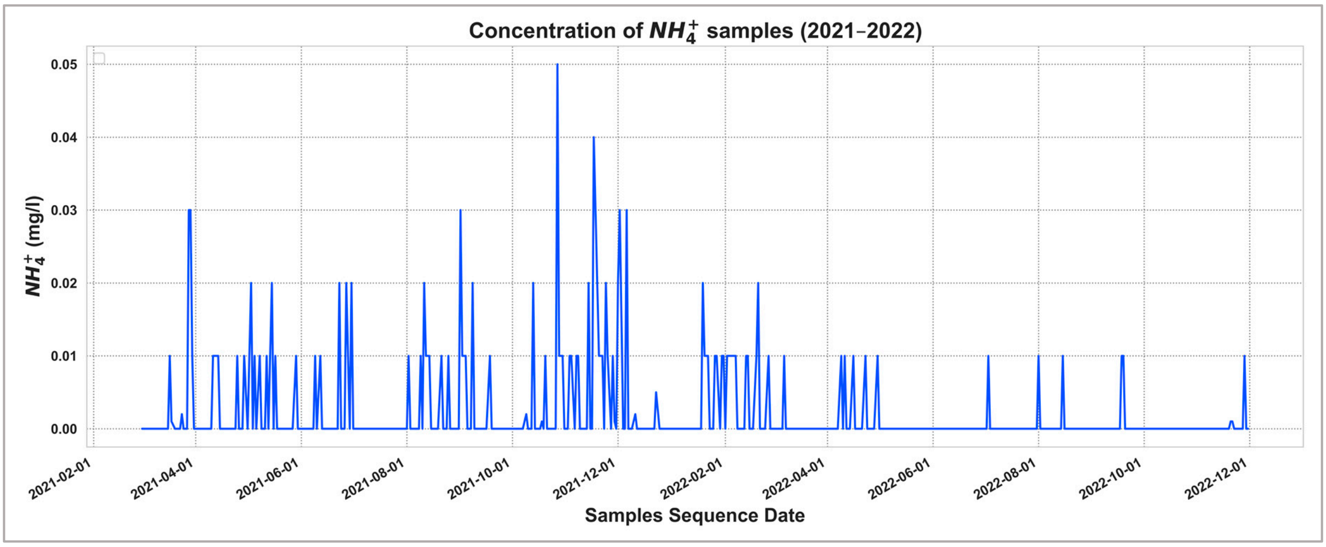


Figure A2. Variations in the concentrations of ammonium ion (NH_4^+) (mg/L) in the Lattakia drinking water intake during the study period (2021–2022).

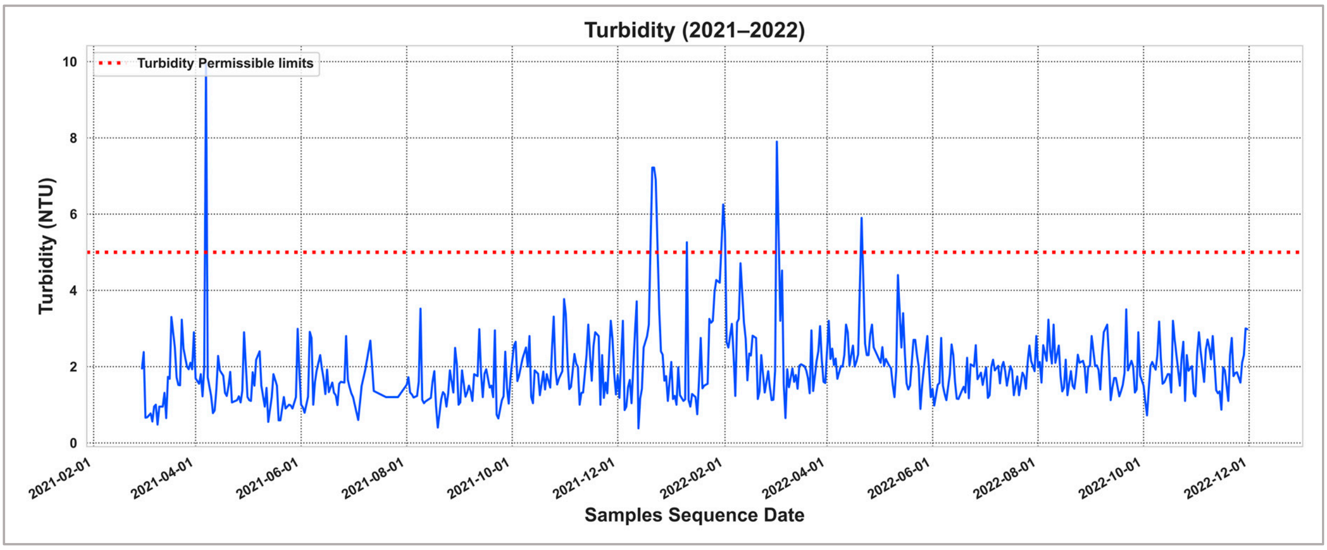


Figure A3. Changes in turbidity values (NTU) in the Lattakia drinking water intake during the study period (2021–2022).

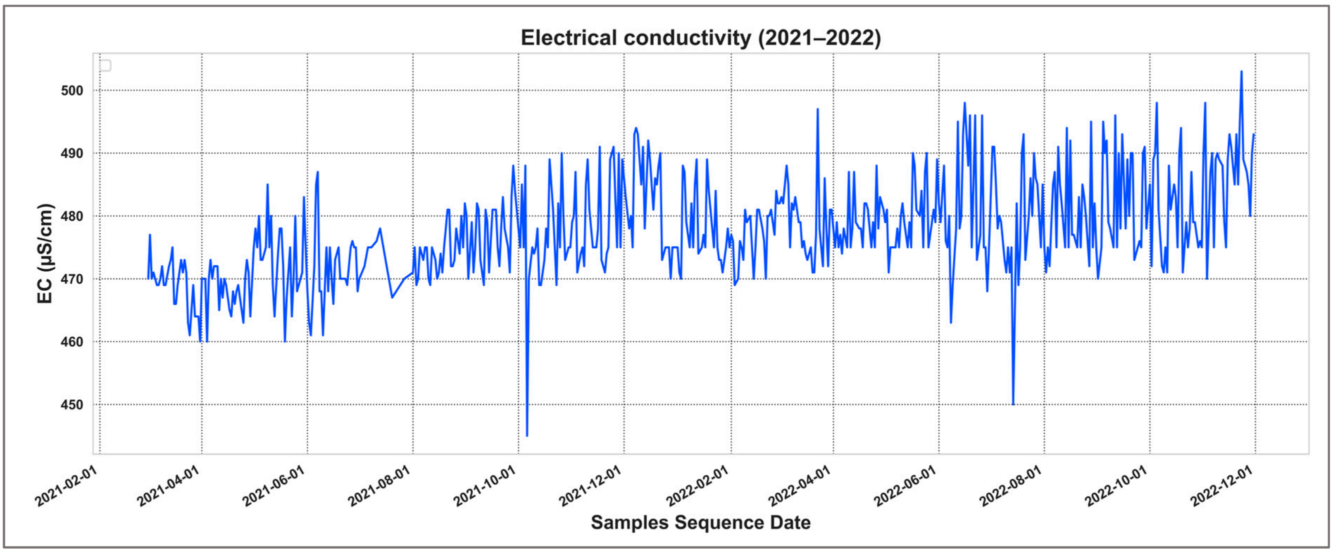
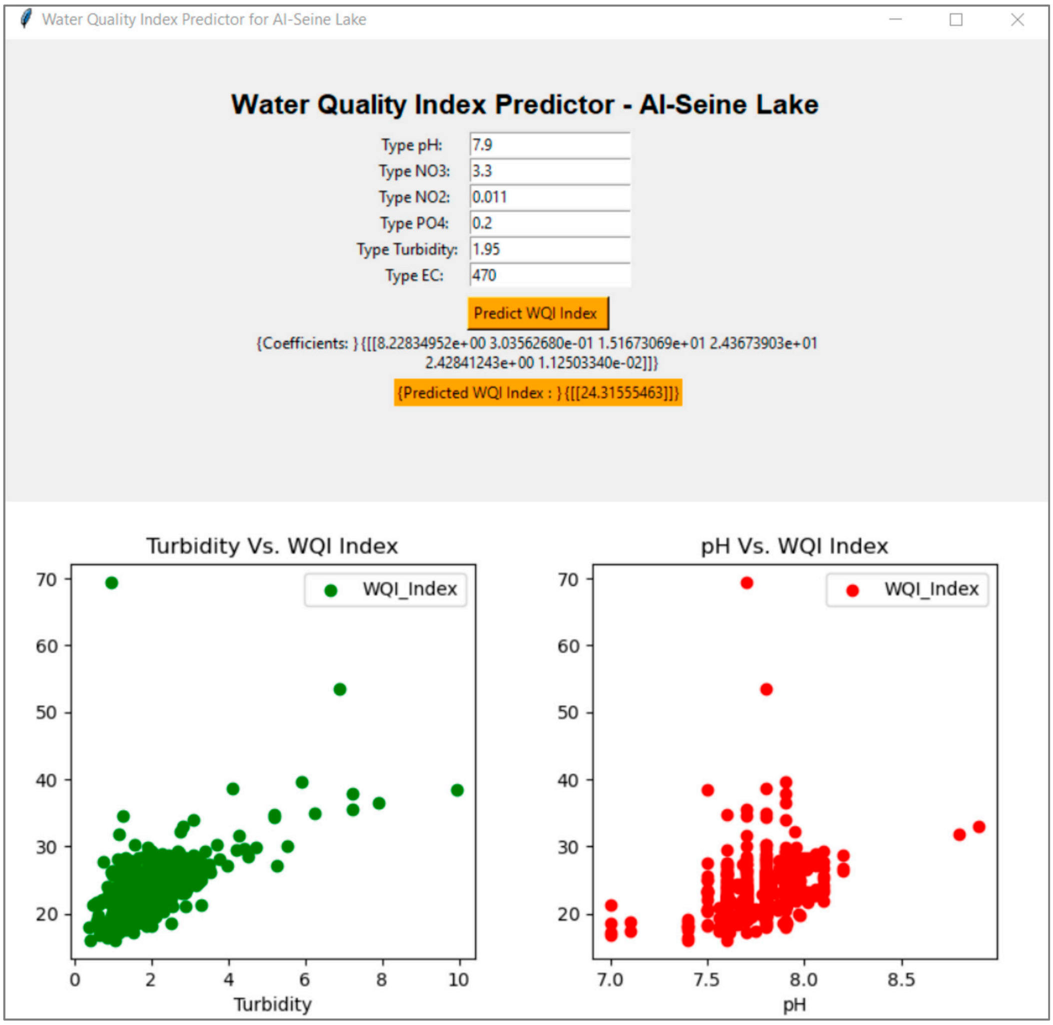


Figure A4. Changes in electrical conductivity values (µS/cm) in the Lattakia drinking water intake during the study period (2021–2022).



Type pH: 7.9
Type NO3: 3.3
Type NO2: 0.011
Type PO4: 0.2
Type Turbidity: 1.95
Type EC: 470

```
{Coefficients: }{[[8.22834952e+00 3.03562680e-01 1.51673069e+01 2.43673903e+01
2.42841243e+00 1.12503340e-02]]}
{Predicted WQI Index : }{[[24.31555463]]}
```

Figure A5. Graphical user interface (GUI) for calculating the water quality index for Lattakia drinking water intake in Al-Seine Lake.

References

1. World Health Organization. *Guidelines for Drinking-Water Quality: First Addendum to the Fourth Edition*; WHO: Geneva, Switzerland, 2017.
2. Nouraki, A.; Alavi, M.; Golabi, M.; Albaji, M. Prediction of water quality parameters using machine learning models: A case study of the Karun River, Iran. *Environ. Sci. Pollut. Res.* **2021**, *28*, 57060–57072. [CrossRef] [PubMed]
3. UN Environment Programme. *A Snapshot of the World's Water Quality: Towards a Global Assessment*; United Nations Environment Programme: Nairobi, Kenya, 2016.
4. Asadollah, S.B.H.S.; Sharafati, A.; Motta, D.; Yaseen, Z.M. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *J. Environ. Chem. Eng.* **2021**, *9*, 104599. [CrossRef]
5. Mishra, B.K.; Regmi, R.K.; Masago, Y.; Fukushi, K.; Kumar, P.; Saraswat, C. Assessment of Bagmati river pollution in Kathmandu Valley: Scenario-based modeling and analysis for sustainable urban development. *Sustain. Water Qual. Ecol.* **2017**, *9*, 67–77. [CrossRef]
6. Ewaid, S.H.; Abed, S.A. Water quality index for Al-Gharraf river, southern Iraq. *Egypt. J. Aquat. Res.* **2017**, *43*, 117–122. [CrossRef]
7. Ramakrishnaiah, C.; Sadashivaiah, C.; Ranganna, G. Assessment of water quality index for the groundwater in Tumkur Taluk, Karnataka State, India. *E-J. Chem.* **2009**, *6*, 523–530. [CrossRef]
8. Ewaid, S.H.; Abed, S.A. Water quality assessment of Al-Gharraf River, South of Iraq using multivariate statistical techniques. *Al-Nahrain J. Sci.* **2017**, *20*, 114–122. [CrossRef]
9. Tung, T.M.; Yaseen, Z.M. A survey on river water quality modelling using artificial intelligence models: 2000–2020. *J. Hydrol.* **2020**, *585*, 124670.
10. Nearing, G.S.; Kratzert, F.; Sampson, A.K.; Pelissier, C.S.; Klotz, D.; Frame, J.M.; Prieto, C.; Gupta, H.V. What role does hydrological science play in the age of machine learning? *Water Resour. Res.* **2021**, *57*, e2020WR028091. [CrossRef]
11. Jafar, R. Assessment of surface water quality by using multivariate statistical techniques. *Tishreen Univ. J. Eng. Sci. Ser.* **2022**, *44*, 11–31.
12. Abbasi, T.; Abbasi, S.A. *Water Quality Indices*; Elsevier: Amsterdam, The Netherlands, 2012.
13. Ahmed, M.; Mumtaz, R.; Hassan Zaidi, S.M. Analysis of water quality indices and machine learning techniques for rating water pollution: A case study of Rawal Dam, Pakistan. *Water Supply* **2021**, *21*, 3225–3250. [CrossRef]
14. Bedi, S.; Samal, A.; Ray, C.; Snow, D. Comparative evaluation of machine learning models for groundwater quality assessment. *Environ. Monit. Assess.* **2020**, *192*, 776. [CrossRef] [PubMed]
15. Gupta, S.; Gupta, S.K. Evaluation of River Health Status Based on Water Quality Index and Multiple Linear Regression Analysis. In *Sustainable Environmental Engineering and Sciences: Select Proceedings of SEES 2021*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 77–85.
16. Wu, Z.; Wang, X.; Chen, Y.; Cai, Y.; Deng, J. Assessing river water quality using water quality index in Lake Taihu Basin, China. *Sci. Total Environ.* **2018**, *612*, 914–922. [CrossRef] [PubMed]
17. Nair, J.P.; Vijaya, M. River Water Quality Prediction and index classification using Machine Learning. *Proc. J. Phys. Conf. Ser.* **2022**, *2325*, 012011. [CrossRef]
18. Malek, N.H.A.; Wan Yaacob, W.F.; Md Nasir, S.A.; Shaadan, N. Prediction of Water Quality Classification of the Kelantan River Basin, Malaysia, Using Machine Learning Techniques. *Water* **2022**, *14*, 1067. [CrossRef]
19. Nguyen, D.P.; Ha, H.D.; Trinh, N.T.; Nguyen, M.T. Application of artificial intelligence for forecasting surface quality index of irrigation systems in the Red River Delta, Vietnam. *Environ. Syst. Res.* **2023**, *12*, 24. [CrossRef]
20. Rezaie-Balf, M.; Attar, N.F.; Mohammadzadeh, A.; Murti, M.A.; Ahmed, A.N.; Fai, C.M.; Nabipour, N.; Alaghmand, S.; El-Shafie, A. Physicochemical parameters data assimilation for efficient improvement of water quality index prediction: Comparative assessment of a noise suppression hybridization approach. *J. Clean. Prod.* **2020**, *271*, 122576. [CrossRef]
21. Kouadri, S.; Elbeltagi, A.; Islam, A.R.M.T.; Kateb, S. Performance of machine learning methods in predicting water quality index based on irregular data set: Application on Illizi region (Algerian southeast). *Appl. Water Sci.* **2021**, *11*, 190. [CrossRef]
22. Irwan, D.; Ali, M.; Ahmed, A.N.; Jacky, G.; Nurhakim, A.; Ping Han, M.C.; AlDahoul, N.; El-Shafie, A. Predicting Water Quality with Artificial Intelligence: A Review of Methods and Applications. *Arch. Comput. Methods Eng.* **2023**, 1–20. [CrossRef]
23. Jafar, R. Application of the Water Quality Index (NSFWQI) on the Al-Sain Lake. *Tishreen Univ. J. Eng. Sci. Ser.* **2016**, *38*, 20.
24. Yadav, A.K.; Khan, P.; Sharma, S.K. Water Quality Index Assessment ofGroundwater in Todaraisingh Tehsil of Rajasthan State, India—A Greener Approach. *E-J. Chem.* **2010**, *7*, S428–S432. [CrossRef]
25. World Health Organization. *Guidelines for Drinking-Water Quality: Incorporating the First and Second Addenda*; World Health Organization: Geneva, Switzerland, 2022.
26. Meride, Y.; Ayenew, B. Drinking water quality assessment and its effects on residents health in Wondo genet campus, Ethiopia. *Environ. Syst. Res.* **2016**, *5*, 1. [CrossRef]
27. Pooja, D.; Kumar, P.; Singh, P.; Patil, S. *Sensors in Water Pollutants Monitoring: Role of Material*; Springer: Berlin/Heidelberg, Germany, 2020.

28. Singh, A.L.; Tripathi, A.K.; Kumar, A.; Singh, V. Nitrate and phosphate contamination in ground water of Varanasi, Uttar Pradesh, India. *J. Ind. Res. Technol.* **2012**, *2*, 26–32.
29. Bui, D.T.; Khosravi, K.; Tiefenbacher, J.; Nguyen, H.; Kazakis, N. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Sci. Total Environ.* **2020**, *721*, 137612. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.