water

MDPI

Article

Machine Learning Algorithms for Predicting the Water Quality Index

Enas E. Hussein [1,*], Muhammad Yousuf Jat Baloch [2,*], Anam Nigar [3], Hussain F. Abualkhair [4], Faisal Khaled Aldawood [5] and Elsayed Tageldin [6]

1 National Water Research Center, Shubra El-Kheima 13411, Egypt
2 College of New Energy and Environment, Jilin University, Changchun 130021, China
3 School of Electronics and Information Engineering, Changchun University of Science and Technology, Changchun 130022, China; nigaranam@yahoo.com
4 Department of Mechanical Engineering, Collage of Engineering, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia; habualkhair@tu.edu.sa
5 Department of Mechanical Engineering, College of Engineering, University of Bisha, P.O. Box 001, Bisha 67714, Saudi Arabia; faldawood@ub.edu.sa
6 Faculty of Engineering and Technology, Future University in Egypt, New Cairo 11835, Egypt; elsayed.tageldin@fue.edu.eg
* Correspondence: enas_el-sayed@nwrc.gov.eg (E.E.H.); engr.yousuf@yahoo.com (M.Y.J.B.)

Abstract: Groundwater is one of the water resources used to preserve natural water sources for drinking, irrigation, and several other purposes, especially in industrial applications. Human activities related to industry and agriculture result in groundwater contamination. Therefore, investigating water quality is essential for drinking and irrigation purposes. In this work, the water quality index (WQI) was used to identify the suitability of water for drinking and irrigation. However, generating an accurate WQI requires much time, as errors may be made during the sub-index calculations. Hence, an artificial intelligence (AI) prediction model was built to reduce both time and errors. Eighty data samples were collected from Sakrand, a city in the province of Sindh, to investigate the area's WQI. The classification learners were used with raw data samples and the normalized data to select the best classifier among the following decision trees: support vector machine (SVM), k-nearest neighbors (K-NN), ensemble tree (ET), and discrimination analysis (DA). These were included in the classification learner tool in MATLAB. The results revealed that SVM was the best raw and normalized data classifier. The prediction accuracy levels for the training data were 90.8% and 89.2% for the raw and normalized data, respectively. Meanwhile, the prediction accuracy levels for the testing data were 86.67 and 93.33% for the raw and normalized data, respectively.

Keywords: groundwater; water quality index; classification learners; support vector machine

Citation: Hussein, E.E.; Jat Baloch, M.Y.; Nigar, A.; Abualkhair, H.F.; Aldawood, F.K.; Tageldin, E. Machine Learning Algorithms for Predicting the Water Quality Index. *Water* **2023**, *15*, 3540. https://doi.org/10.3390/w15203540

Academic Editors: Habib Ullah and Asfandyar Shahab

Received: 8 August 2023
Revised: 22 September 2023
Accepted: 26 September 2023
Published: 11 October 2023

1. Introduction

Evaluating groundwater quality is essential for conserving the ecological environment, and it forms the basis for the sustainable use of local groundwater resources [1–3]. Rising populations, industrial development, and modern lifestyles have increased water use [4–6]. The most accessible source of fresh water on the planet is groundwater, and it is essential for human health and industrial and agricultural purposes [7,8]. Chemicals and microbial contamination frequently challenge water safety [9–11]. Groundwater has taken over as the primary water supply; it is the preferred sole source, particularly in arid and semi-arid regions where precipitation is inadequate and surface water resources are lacking [12–17]; also, waste material including sewage sludge on soil can increase crop production but contaminate groundwater [18–25].

Pakistan is facing a problem in terms of groundwater contamination [26,27]. In recent decades, the groundwater sources in Pakistan, especially in the Punjab and Sindh provinces,

have been rendered highly vulnerable due to rapid industrialization, mining activities, and agricultural practices, all of which have made the groundwater sources susceptible to contamination and lowered their quality and quantity. Groundwater contamination has been reported in many regions of Pakistan. Some regions in the Sindh province, such as Jamshoro [28], Manchar Lake and Tharparkar [29,30], Hyderabad [31], and Nawabshah [9], currently suffer from contamination. Some regions in the Punjab province also suffer from this issue, including Multan [32], Lahore [33], Sialkot [34], Jhelum [35], Mandi Bahauddin [36], Gujrat [24], Sheikhupura [37], Vehari [2], and Sargodha [38]. Groundwater studies have, thus far, been limited to investigating the effects of groundwater contamination on human life.

Unfortunately, due to increased human activities and environmental changes, increasing amounts of groundwater have been contaminated in the last decade [9,39–41]. Therefore, the water quality (WQ) has deteriorated significantly due to industrial and urban activities, resulting in more severe diseases [42]. Thus, groundwater quality assessment is essential for addressing the impacts of human activities and environmental change on groundwater worldwide [43]. WQ describes water characteristics, such as chemical, physical, and biological parameters, to identify its suitability for utilization, whether for drinking or irrigation purposes [44]. The water quality index (WQI) is a powerful tool for identifying and assessing WQ. This index transforms many water quality parameters with intercorrelated data into a single value that refers to the state of the water quality [45,46]. The WQI concerns both surface water sources and groundwater, since groundwater is the primary water source for communities, especially in desert areas [47]. Therefore, various countries/organizations have established several WQI techniques for assessing the quality of surface water sources and groundwater [45]. However, WQI computations are time-consuming and are often found to encounter various errors during the sub-index calculation. Therefore, using an artificial intelligence prediction model may reduce both the time needed for computation and the computation errors that lead to inaccurate decision making [48].

Artificial intelligence (AI) techniques have recently been widely used as multifunctional and robust tools for developing WQI models. AI models work to predict and assess WQIs for both surface water and groundwater sources [23,49,50]. Therefore, more research focused on using AI and optimization techniques to present WQI models is required. In [51,52], seven WQI models, which were categorized into four weighted (NSF, SRDD, WJ, and WQM) and three unweighted (RMS, Hanh, and AM) algorithms, were used to construct WQI models. The WQI values were developed using an improved WQI methodology [52]. In [53], classification algorithms were used, such as support vector machines (SVMs), naïve Bayes (NB), random forest (RF), k-nearest neighbor (KNN), and gradient boosting (XGBoost), to select the best option for predicting accurate WQIs. The findings suggested that these models were reliable and effective means of producing highly accurate WQI predictions. In [23], six WQI models were developed to predict the WQIs of collected samples, and the models included the generalized regression neural network (GRNN), the Elman neural network (Elm NN, considered a new-generation learning tool), and the feed forward neural network (FFNN), as well as the support vector machine (SVM), linear regression (LR), and neuro-fuzzy (NF) models. The results indicated that the NF WQI model exhibited higher prediction accuracy than the others. The neural network (NN) model was used in [54] to map the nonlinear relation between pH, total dissolved solids (TDS), total alkalinity (TA), total hardness (TH), calcium hardness (Ca-H), residual chlorine, nitrate levels (as NO_3^-), and chloride levels (Cl^-) and WQI. A total of 710 samples from the Jodhpur Rajasthan region in India were used to design the NN model. The results illustrated that the model was highly efficient at predicting the WQIs of the new samples. The developed framework could be automated in this work to help evaluate WQ for better management. In [48], eight AI algorithms were applied to establish a prediction WQI model for samples collected from the Ikkizi region in southeast Algeria. The AI algorithms utilized were regression (MLR), random forest (RF), M5P tree (M5P), random subspace

(RSS), additive regression (AR), artificial neural network (ANN), support vector regression (SVR), and locally weighted linear regression (LWLR). The study developed two scenarios to reduce the computation time, and the second scenario was used to show the water quality variation in critical cases when the essential analysis was required. The findings indicated that TDS and TH are the most important parameters influencing the WQI. MLR exhibited the highest prediction accuracy for the first scenario, and RF was the best for the second scenario. In [55], due to the cost of the computational approaches for determining the suitability of drinking water, classification techniques were used to construct the prediction model, such as decision trees (DTs), k-nearest neighbors (KNN), discriminants analysis (DA), a support vector machine (SVM), and ensemble trees (ETs). The classification methods were applied to 169 data samples collected from the region of Naâma, Algeria. The results revealed that a trained SVM classifier could accurately predict the WQI for the new data samples that assist WQ control and support decision making.

In this work, classification techniques based on classification learners in MATLAB were used to build a classifier that was utilized to predict the WQI. Using the classifiers to identify the WQI of the data samples can save time and reduce the costs of conducting WQI computations every time. The classification techniques were applied to 80 data samples collected from Sindh province, Pakistan's second-most populous province. The 80 samples were divided up, with 65 used for training, and the remaining 15 used for testing. The results indicate that the linear support vector machine (Linear SVM) constituted the highest prediction accuracy model for the applied data. For the raw and normalized data, the prediction accuracy for the training data was 90.8 and 89.2%, respectively. On the other hand, the prediction accuracy for the raw and normalized testing data was 93.33% and 86.67%, respectively.

The materials and methods are presented in Section 2. In Section 3, the classification learner methods are outlined, and we describe results produced by applying various classifiers to the data. The results of using linear SVM on the raw and normalized data are illustrated in Section 4. Finally, the key conclusions are explained in Section 5.

2. Materials and Methods

2.1. Description of the Study Area

The research region is located in Sindh, Pakistan's second most highly populated province, where the study was conducted. Sakrand is a city in the province of Sindh (Figure 1). The research area is 25 m (82 ft) above sea level and has a hot, dry climate. It can be categorized as an arid subtropical zone, meaning that winters are exceedingly cold and dry, while summers are extremely hot, arid, humid, and windy. Winter lows of 1 °C may be experienced, while summer highs of 53 °C are possible. The bulk of the yearly precipitation, which ranges from 200 to 300 mm [56], falls during the monsoon season, which lasts from July through September. The most extensively grown crops in the area are wheat and cotton, which are grown in Sindh's delta plain [57]. The groundwater resources come from the Indus River.

In contrast to the water in other vital regions, which is sometimes relatively salty, the groundwater near the constricting stretch of the Indus River is usually less salty [58]. Low cropping intensity, canal seepage, and lateral channels contribute to salinization. A chemical analysis showed that more than 5.5% of the landmass comprises saline–alkaline soil, and 15% of the investigated region is moderately salinized [25]. The average depth of the groundwater table is 4.53 m, with a range of 1.5 to 12 m. The flow is more westerly, toward the Indus River, at the center and lower parts, and, in the higher and intermediate portions, it is more south-westerly. Typically, sand makes up the majority of the aquifer, which is uniform, does not have artesian water, and flows.

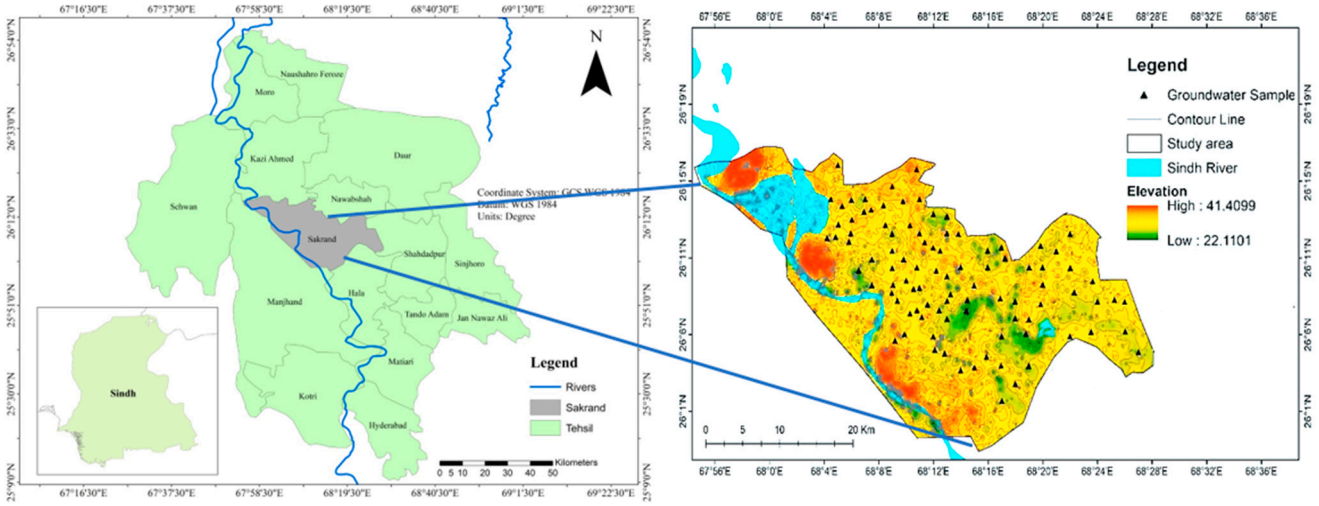


Figure 1. Study area and sampling points.

2.2. Sample Collection and Analysis

In Sakrand, Sindh province, Pakistan, 80 groundwater samples were collected from shallow aquifers (<35 m) from April to May 2022 and filtered to 0.45 µm for further analysis. A global positioning system (GPS) was used to record the groundwater samples' locations. All analyses followed the American Public Health Association's (APHA 2005) standard methods; 250 and 100 mL glass bottles were used to collect samples to analyze cations and anions, respectively. A drop of nitric acid (HNO_3) was immediately added to the samples to adjust the pH (<2.0) for cation analysis [59]. The samples were analyzed for drinking water use and to evaluate the water quality index (WQI). The following tests were used to determine the physicochemical parameters of the water: An EC-TDS-Temp (RS232C/Meter CON 110 m) was used to detect the total dissolved solids (TDS), electrical conductivity (EC), and temperature. The pH was computed using a meter from the pH 720 WTW Series. The process of acid titration determined the total alkalinity of the samples that contained methyl-orange. Cl^- and $HCO_3{}^-$ were determined using titration techniques; however, turbidity was measured using a turbidity meter, as opposed to the main anions such as ($NO_3{}^-$) and sulfate ($SO_4{}^{2-}$), which were determined using a UV–VIS (ultraviolet–visible) spectrophotometer (Analytik Jena, Jena, Germany). The cations, including Mg^{2+}, Ca^{2+}, K^+, Na^+, and Fe^{2+}, were measured using a flame photometer (PFP7, Cambridge shire, UK). Finally, the samples were checked for accuracy to determine % charge balance errors according to Equation (1):

$$\% \text{ CBE} = \frac{[\sum \text{cations} - \sum \text{anions}]}{[\sum \text{cations} + \sum \text{anions}]} \times 100 \quad (1)$$

The physicochemical analysis with % CBE within ± 5% is ideal for further investigation. The unit of cations and anions is meq/L.

2.3. Water Quality Index (WQI) of Groundwater

The water quality index (WQI) determines whether water is fit for human consumption. Water quality assessment is the most thorough approach for evaluating groundwater quality; as a result, the WHO criteria were utilized to assess the groundwater quality. We carefully examined the combined impact of several chemical factors on groundwater. The WQI is determined using three steps.

The first step includes determining the individual parameters' weights. All of the parameters and weights are denoted by the symbol wi. In the order of their relative weight relevance, the parameters were EC, pH, TDS, HCO^{3-}, Cl^-, $SO_4{}^{2-}$, Ca^{2+}, Mg^{2+}, Na^+, K^+,

NO^{3-}, F^{-}, Fe, and As. In step two, the relative weights for each parameter were computed using Equation (2):

$$W_i = \frac{w_i}{\sum_i^n w_i} \quad (2)$$

where the relative weight is denoted by W_i, the weight associated with each parameter is represented by w_i, and the total number of parameters is denoted by n.

In the third step, the quality-rating scale q_i is calculated for each parameter using Equation (3):

$$q_i = \frac{C_i}{S_i} \times 100 \quad (3)$$

where C_i is the concentration of each parameter, S_i is the WHO standard, and qi is the quality index. In addition, sub-indices are computed using Equations (4) and (5):

$$SI_i = W_i \times q_i \quad (4)$$

$$WQI = \sum_{i=1}^{n} SI_i \quad (5)$$

SI_i and W_i are the sub-indices and relative weights for the ith parameter, respectively, whereas q_i is the concentration-based rating.

Overall, we divided the WQI into three groups—(i) <50, excellent; (ii) >50, good; and (iii) >100, poor—and we classified it using the machine learning models.

3. Classification Learner

3.1. Training Results of Classification Learner

Classification is the process of classifying a given dataset; it is performed on both structured and unstructured data. In addition, it predicts the data point classes. Classes are referred to as targets, labels, or categories. The main goal of classification is to identify the class to which the new data will belong. Therefore, the definition of a classification model is that the model predicts or draws a conclusion on the input data provided for training and will predict the class or category of the data.

The classification learner is a MATLAB tool. It includes several classifiers, the most important of which are the decision tree, the support vector machine (SVM), the k-nearest neighbors (KNNs), the ensemble trees, and discrimination analysis [60,61]. In addition, it allows for supervised learning tasks to be performed, such as interactive data exploration, feature selection, defining validation schemes, constructing training models, and evaluating results. Supervised machine learning can be performed using different classifiers by feeding a known dataset (observations) and a known output (responses) in the form of labels of classes. The trained model can be exported to predict the classes of a new dataset.

The data to be classified were collected from Sakrand, Sindh, Pakistan, and each sample's water quality index (WQI) was computed. A total of 80 samples were used in this study. The data were categorized as training and testing data. Sixty-five data samples were used to train the classification model, and fifteen data samples were used to test the constructed model and determine its prediction accuracy. The WQI was classified into three classes (excellent, good, and poor), as shown in Table 1. The distribution of the sample data based on the WQI class limit is categorized in Table 2. The data, including 65 training samples, were distributed based on WQI computation as follows: 13 samples were excellent, 40 were good, and 12 were poor in terms of the WQI state. The 15 testing samples were categorized as follows: 2 exhibited an excellent WQI state, 2 had a poor WQI state, and the 11 remaining samples exhibited good WQI states. Table 3 illustrates some of the data used to construct the training model of classification. EC, pH, TDS, HCO^{3-}, Cl^{-}, SO_4^{2-}, Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, NO^{3-}, F^{-}, Fe, and As were used as the input parameters to the classifiers, and the code related to the WQI value is the output of the model. The water quality index (WQI) is the last column, and the code refers to each sample class based on

the WQI class limit in Table 1, where code 1 refers to an excellent WQI, 2 refers to a good WQI, and 3 refers to a poor WQI.

Table 1. WQI class limit.

Water Quality Index (WQI)	WQI Class	WQI Code
0 < WQI < 50	Excellent	1
50 < WQI < 100	Good	2
WQI > 100	Poor	3

Table 2. WQI data sample distribution.

WQI State	Number of Samples	
	Training	**Testing**
Excellent	13	2
Good	40	11
Poor	12	2
Total	65	15

Table 3. Training data samples.

Model Inputs																	Model Output	
EC	**PH**	**NTU**	**TDS**	**ALK.**	**TH**	**HCO_3^-**	**Cl^-**	**SO_4^{2-}**	**Ca^{2+}**	**Mg^{2+}**	**Na^+**	**K^+**	**NO_3^-**	**F^-**	**Fe**	**As**	**Code**	**WQI**
1306	7.6	2.1	474	2.4	220	120	89	81	40	29	68	2.1	0.89	0.14	0.01	0	1	34.947
1307	7.7	2.4	358	2	150	100	88	48	24	22	58	1	0.6	0.18	0.03	10	1	31.974
1308	7.4	2.8	401	2.4	180	120	89	58	32	24	60	2	0.77	0.14	0.07	0	1	32.144
1309	7.7	2.9	656	4.2	310	210	130	89	52	44	90	3.2	0.66	0.19	0.03	0	1	46.7
1876	8.1	2.2	1078	13.6	560	280	195	169	68	53	171	3.3	0.76	0.29	0.04	0	2	69.322
1356	8	3	884	5.8	410	450	168	139	180	51	120	4.6	0.67	0.23	0.03	0	2	69.837
2700	6.9	2.9	1201	9.2	820	280	280	228	72	58	139	4.6	0.92	0.37	0.04	10	2	83.405
1399	7.2	3.9	868	6.6	430	680	150	175	160	90	120	3.9	0.99	0.32	0.07	0	2	78.639
1949	6.9	3	1228	9.2	600	290	189	180	74	61	167	6.1	1	0.44	0.06	10	2	77.049
1566	7.2	3.2	895	28	480	460	200	144	159	49	158	3.8	0.86	0.37	0.09	0	2	73.765
2440	6.8	2.5	1600	4.2	750	850	238	200	120	115	211	6	0.97	0.27	0.09	0	3	107.43
3220	6.6	2.1	1312	9.6	890	350	395	320	62	27	326	5.4	0.99	0.44	0.03	0	3	100.59
3080	6.7	2	710	20.2	900	780	311	280	130	134	281	13.3	0.78	0.42	0.23	0	3	118.38
3880	6.8	3.9	1952	6	850	700	333	265	210	128	309	14	2	0.55	0.02	0	3	137.66

Data samples were used in their raw form without any data transformation. In addition, data were normalized by dividing each value in each column by the maximum value of the column. The normalization process was performed according to Equation (6):

$$X_{normalized} = \frac{X_i}{max(X_n)} \quad (6)$$

where $X_{normalized}$ refers to the normalized value of X, X_i expresses the parameter's value in sample i, and n refers to the total number of samples in each parameter's column.

When using the raw data and the normalized data samples for all classifiers in the classification learner MATLAB tools, the classification accuracy of each classifier was recorded, as shown in Table 4. Table 4 illustrates the prediction accuracies of each classifier. The results indicate that the linear SVM developed the highest prediction accuracies, whether the data were in their raw form or normalized. For example, the prediction accuracy was 90.8% for the training data in their raw form and 89.2% for the normalized data. On the other hand, the KNN classifier developed the second-best prediction accuracy for the training data. The prediction accuracies were the same for the raw and standardized data, with a value of 87.7%. Based on the prediction accuracy results, the trained model

developed using linear SVM was used as the classifier model to test the new data and determine their class.

Table 4. Comparison between different classifiers.

Classifier	Training Data	
	Raw Data	**Normalization**
1. Decision Tree (DT)		
Fine tree	72.3	80.0
Medium tree	72.3	80.0
Coarse tree	69.2	76.9
Linear discriminant	86.2	86.2
Quadratic discriminant	Fail	Fail
2. Support Vector Machine (SVM)		
Linear SVM	**90.8**	**89.2**
Quadratic SVM	87.7	83.1
Cubic SVM	89.2	86.2
Fine Gaussian SVM	61.5	61.5
Medium Gaussian SVM	84.6	80.0
Coarse Gaussian SVM	66.2	66.2
3. K-Nearest Neighbors (KNN)		
Fine KNN	87.7	87.7
Medium KNN	78.5	80.0
Coarse KNN	61.5	61.5
Cosine KNN	75.4	73.8
Cubic KNN	87.5	76.9
Weighted KNN	84.6	83.1
4. Ensemble Trees		
Ensemble Boosted Trees	61.5	61.5
Ensemble Bagged Trees	83.1	86.2
Ensemble Subspace Discriminant	84.6	84.6
Ensemble Subspace KNN	86.2	86.2
Ensemble RUSBoosted Trees	78.5	80.0

Inaccurate predictions can arise for various reasons:

- There are not enough relevant and informative features. Thus, the classifiers struggle to make accurate predictions.
- The training data are imbalanced, leading to one class having significantly more instances than the others.
- When a classifier learns too much from the training data and cannot generalize effectively to new data; this might occur if the model is complicated or the training set contains noise.
- When a classifier is too simple and fails to capture the underlying patterns in the data; this can happen if the model is not complex enough or if there are insufficient training data.

3.2. Linear Support Vector Machine (LSVM)

SVM is a linear model for classification and regression problems. It can deal with many practical problems, both linear and nonlinear. Its premise is straightforward: developing a hyperline or plane that separates the data into classes. Therefore, the new data point can easily be placed in the accurate class. This method can be categorized into two types: linear SVM and nonlinear SVM. Linear SVM can be used for linear data when the data set can be classified into two classes using a single straight line. For nonlinear SVM, the data cannot be classified using a straight line [60,62].

For linear SVM, it helps to find the best decision line or boundary of each class, as shown in Figure 2. This boundary or better region is called a hyperplane. Next, it explores the nearest point of the lines of both classes, which are called support vectors. The distance

between the vectors and the hyperplane is called the margin. SVM is used to maximize this margin.

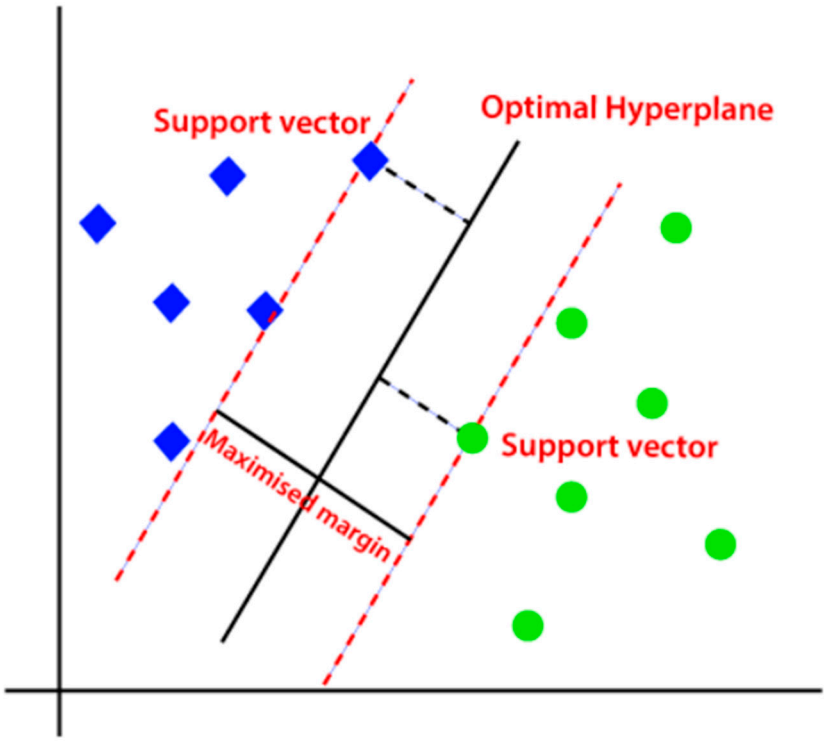


Figure 2. Support vector machine algorithm (blue square is class 1 and green circle is class 2).

Therefore, the hyperplanes are essential in multidimensional space to distinguish between classes, and they are used to discriminate between two classes (y_i and y_i) for two input vectors (X_i and X_j). The classification process is an optimization problem, since the margin between the two classes must be maximized. The hyperplane can be expressed as orthogonal weight vector ω as in Equation (7) [62]:

$$\omega = \begin{bmatrix} \omega_1 & \omega_2 \ldots\ldots\ldots & \omega_k \end{bmatrix} \tag{7}$$

This orthogonal vector can be used with the input vector X_i to identify the function of the hyperplane, h, as in Equation (8):

$$h(X_i) = \omega^T \cdot X_i + \omega_0 = \omega_0 + \sum_{i=1}^{k} \omega_i \cdot x_i \tag{8}$$

where the term ω_0 refers to the basis that is required to define the hyperplane position where $h(X) = 0$.

If two points X_i and X_j refer to the two closest points on each side of the hyperplane, then the $h(X_i)$ and $h(X_j)$ values can be expressed as in Equations (9) and (10), respectively:

$$h(X_i) = \omega^T \cdot X_i + \omega_0 = 1 \tag{9}$$

and

$$h(X_i) = \omega^T \cdot X_i + \omega_0 = -1 \tag{10}$$

where X_i refers to class 1 if $h(X_i) \geq 0$ and class -1 in other cases.

The distance between the two hyperplanes $X_i - X_j$ can be identified using Equation (11):

$$X_i - X_j = \frac{2}{\|\omega\|^2} \tag{11}$$

where ω is the weight vector.

Minimizing ω leads to the maximization of the distance between the two hyperplanes $X_i - X_j$. Therefore, this optimization problem can be identified, as in Equation (12):

$$\begin{cases} Min: \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{k}\zeta_i \\ Such\ that: y_i(\omega.x_i + \omega_0) \geq 1 - \zeta_i \\ \forall i,\ 0 < \zeta_i < 1,\ i = 1, 2, \ldots\ldots, k \end{cases} \quad (12)$$

where ζ_i refers to the slack non-negative variables, and C refers to the margin parameters.

4. Results and Discussion

4.1. Results According to the Training of the Raw Data

Table 4 shows that the linear SVM was the best classifier, with the highest prediction accuracy of 90.8%. The accurate and inaccurate predicted points are shown in the scatter plot in Figure 3. The colored circles refer to the accurate predictions, and the crosses represent the inaccurate predictions. The history section depicts all the classifier results that explained each classifier's prediction accuracy, and the bold number refers to the accuracy of the linear SVM. The number of predictors that expressed all parameters used to compute the WQI was 17 (EC, pH, TDS, HCO_3^-, Cl^-, SO_4^{2-}, Ca^{2+}, Mg^{2+}, Na^+, K^+, NO_3^-, F^-, Fe, and As). Therefore, the total number of training samples represents 65 data samples. There were three response classes (1 for excellent WQI, 2 for good WQI, and 3 for poor WQI, as shown in Table 1). Tenfold cross-validation was applied to ensure that the training process and its results were exemplary. The data were divided into ten groups: one of the ten was used as a validation group, and the other nine were training sets. The prediction accuracy was computed for the nine groups. The validation group was inserted into the nine groups, and another one was selected as a data validation group, and so on. At the end of the training process, the accuracy was determined as the average of the ten prediction accuracies. As shown in Figure 3, the training time was 7.667 s, and the prediction speed was 470 obs/s.

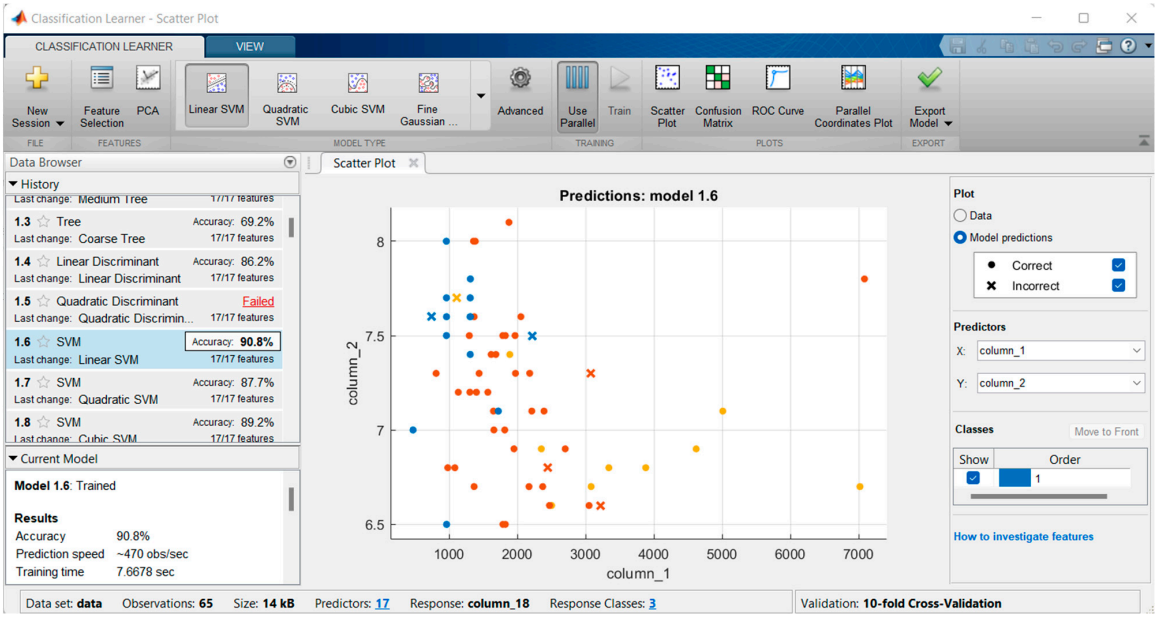
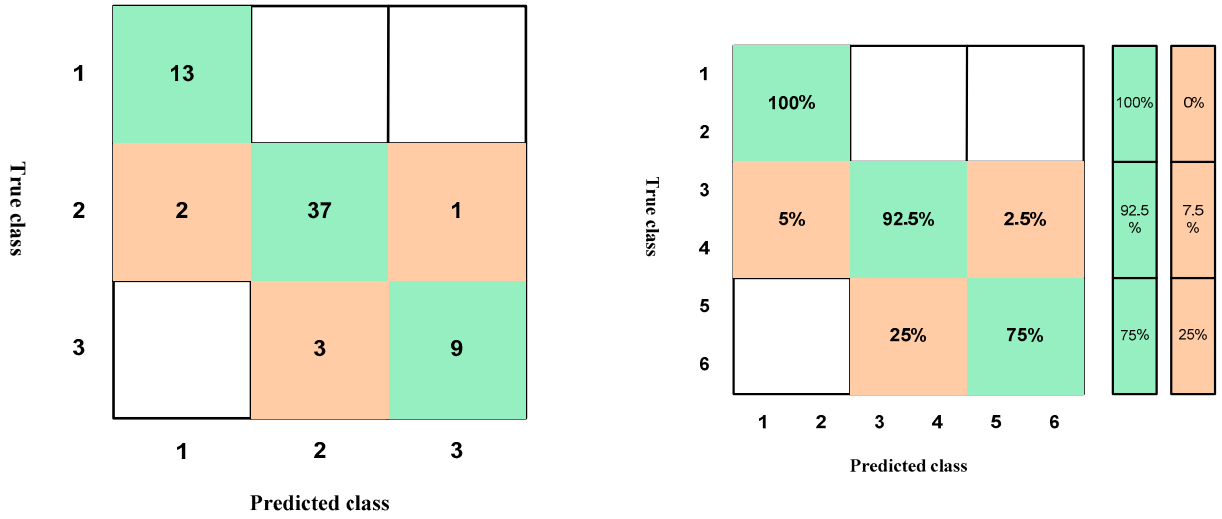


Figure 3. Scatter plot of the training of the raw data.

Figure 4a,b show the confusion matrix of the training data in their raw form. It illustrates the number of accurate predictions in green cells and the number of inaccurate predictions in red cells. For class 1 (excellent WQI), the trained linear SVM produced

accurate predictions for all samples, the same as the actual WQI states, and it accurately predicted the 13 samples with excellent WQI states (13/13, 100%). A total of 37 out of 40 samples were inaccurately predicted as having good WQI, with a prediction accuracy of 92.5%. The prediction accuracy of the poor WQI samples was 75%, with 9 out of 12 samples being accurately predicted.



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

Figure 4. Confusion matrix of the training of the raw data: (**a**) accurate predictions in relation to inaccurate predictions; (**b**) the prediction accuracy of the accurate and inaccurate predictions.

The positive and negative predictive values can be determined as shown in Figure 5. The predicted class 3, for poor WQI, appeared ten times: nine were accurate, with an accuracy of 90%, and one sample was predicted as having a good WQI (class 2), with a 10% error rate. The excellent WQI state appeared 15 times, 13 times accurately, with 87% positive predictive value, and twice as an inaccurate prediction or false prediction, with an error rate of 13%. The two false samples were classified as having a good WQI.

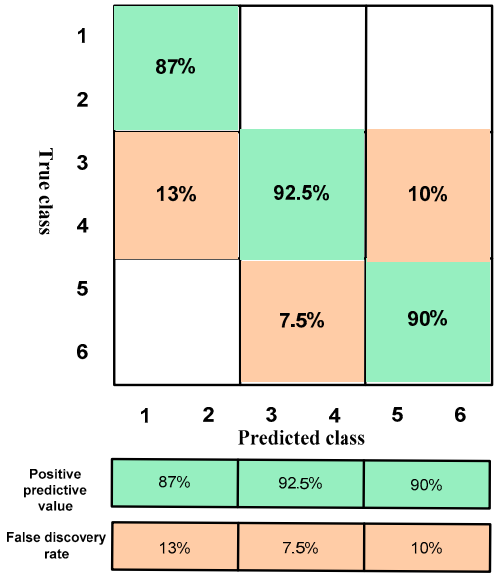


Figure 5. Confusion matrix for positive predictive values and negative predictive values.

Figure 6 illustrates the receiver operating characteristics (ROC). It depicts the current classifier performance with the true positive rate (TPR) and false positive rate (FPR). In

total, 4% of the observations were inaccurately assigned to the positive category based on FPR, and 100% were accurately classified as TPR. Therefore, a classifier's accuracy can be measured according to its area under the curve (AUC), which refers to a prediction accuracy of 98%. Hence, the classifier performed better than expected.

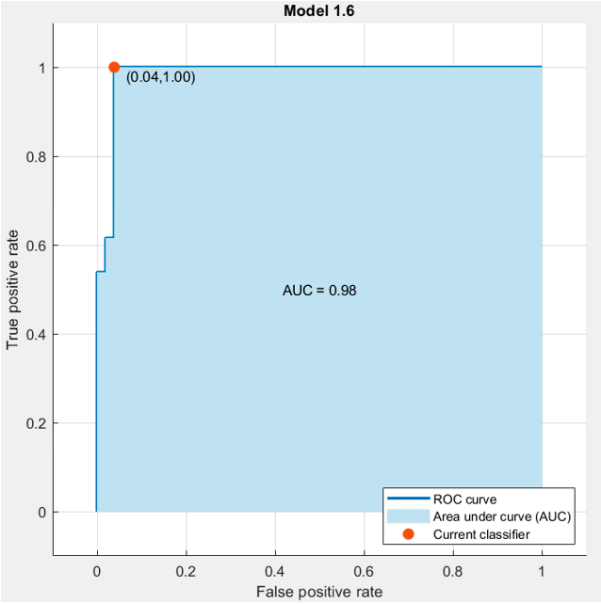


Figure 6. ROC curve between the positive class (class 1) and the negative class (class 3).

As shown in Table 5, a total of 15 samples were used to test the constructed model via the training process. These samples were selected randomly to measure the prediction accuracy of the constructed model. The trained model can be exported to the workspace in MATLAB to create a new file that contains the new sample data without its actual response, and only the input data. When testing the model via the new data samples, the results indicated that the constructed model predicted 13 samples with the accurate class out of the 15 testing samples, with a prediction accuracy of 86.667%. However, two good WQI states were wrongly predicted, one as an excellent WQI and the other as a poor WQI.

Table 5. The testing sample results of the constructed linear SVM.

Model Inputs																		Actual WQI Code	Model Output
EC	PH	NTU	TDS	ALK.	TH	HCO	Cl^-	SO_4^{2-}	Ca^{2+}	Mg^{2+}	Na^+	K^+	NO_3^-	F^-	Fe	As	WQI	Code	Prediction
1304	7	2	669	5.8	330	290	99	98	72	36	77	4.1	0.77	0.17	0.06	0	48.85	1	1
1305	8	2.3	564	4.4	350	220	70	62	62	47	39	2	0.76	0.18	0.04	5	41.74	1	1
1401	6.9	2	1000	6.8	460	340	188	135	88	58	158	2.1	0.87	0.32	0.05	0	64.44	2	2
1617	7.3	1.3	897	6.4	450	320	147	139	78	62	113	1.6	0.72	0.27	0.06	5	62.48	2	2
1920	6.5	3.1	1000	6.7	450	335	186	134	78	62	113	1.6	0.88	0.22	0.05	5	66.49	2	2
2470	6.5	2.1	1228	10.1	610	505	177	186	110	81	158	3.1	0.9	0.25	0.03	5	84.71	2	2
1684	7.2	3.1	1581	11.6	730	580	263	195	150	86	226	8.8	0.98	0.22	0.43	0	99.26	2	3
2380	7.3	3.2	1078	12	480	370	189	177	92	61	162	4.5	0.97	0.23	0.06	5	78.95	2	2
1772	7.1	2.4	1523	12	770	600	236	239	140	102	187	6.5	0.77	0.23	0.05	10	96.27	2	2
1693	7.2	3	1134	7.8	560	390	225	144	94	79	146	4.2	0.75	0.1	0	0	72.21	2	2
1405	6.8	3.6	1054	4.6	580	430	170	155	84	84	53	12.4	0.89	0.19	0.03	5	72.67	2	2
1102	7	2.4	1084	5	430	250	132	179	72	53	124	2.5	0.78	0.16	0.03	5	57.75	2	2
1103	7.1	1	899	4.1	240	204	127	148	40	63	48	3.4	0.88	0.13	0.02	0	48.65	2	1

Table 5. *Cont.*

Model Inputs																	Actual WQI Code	Model Output
EC	PH	NTU	TDS	ALK.	TH	HCO	Cl^-	SO_4^{2-}	Ca^{2+}	Mg^{2+}	Na^+	K^+	NO_3^-	F^-	Fe	As	WQI	Code	Prediction
4620	6.9	2.7	1971	14	1111	900	415	370	260	142	371	5.2	3	0.21	0.03	0	153.7	3	3
1888	7.4	3.2	2483	3	1260	1010	533	388	100	148	472	5.2	0.77	0.43	0.22	0	144.5	3	3
																			86.67

4.2. Results According to the Training of the Normalization of the Data

The effect of data normalization was investigated, to determine whether it enhances the prediction accuracy or leads to a decrease. Thus, data normalization was performed as in (1). The new data were trained for all classifiers using the classification learner tool in MATLAB, and the same result was obtained. The linear SVM produced the highest prediction accuracy, with a value of 89.2%, as shown in Figure 7. The confusion matrix of the prediction results of the trained model referred to some changes rather than the trained model results with raw data. The number of accurate predictions in each class was changed. The results of the trained model indicated that 12 out of 13 excellent WQI samples were accurate, with a prediction accuracy of 92%. For the good WQI samples, 38 were accurately predicted with a prediction accuracy of 95%, and 2 were inaccurate predictions, with a 5% prediction accuracy. Finally, the prediction accuracy of the poor WQI state was 67%, where 8 out of 12 samples were predicted accurately. Figure 8 illustrates the prediction accuracy for each class, where the green cells refer to the accurate predictions, while the red cells refer to the incorrect predictions.

The testing data samples were normalized based on the maximum value of each parameter column. Then, the constructed linear SVM was applied to these testing data to develop the model's prediction accuracy with the new data samples. The prediction accuracy of the constructed model based on the normalized data is higher than that generated by the model constructed with the raw data. As shown in Table 6, the prediction accuracy of the constructed model can be computed according to the agreement between the actual WQI code and the resultant code from the model. The results indicate that the model can accurately predict 14 of the 15 new samples. As shown in Table 6, there was only one incorrect prediction, which occurred in the thirteen samples with a good WQI state; the one incorrectly predicted sample was predicted as having an excellent WQI state.

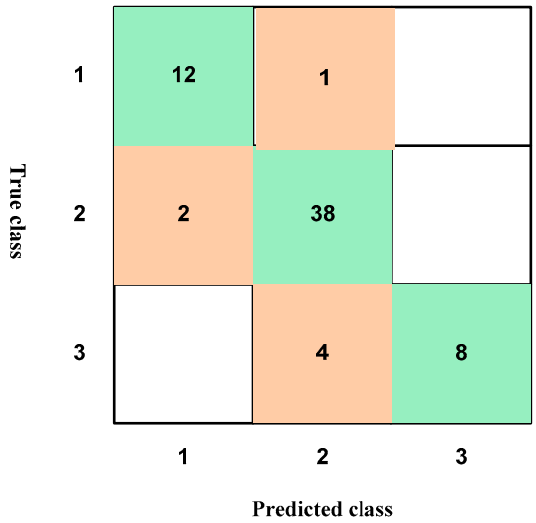


Figure 7. Confusion matrix for the trained model with normalized data.

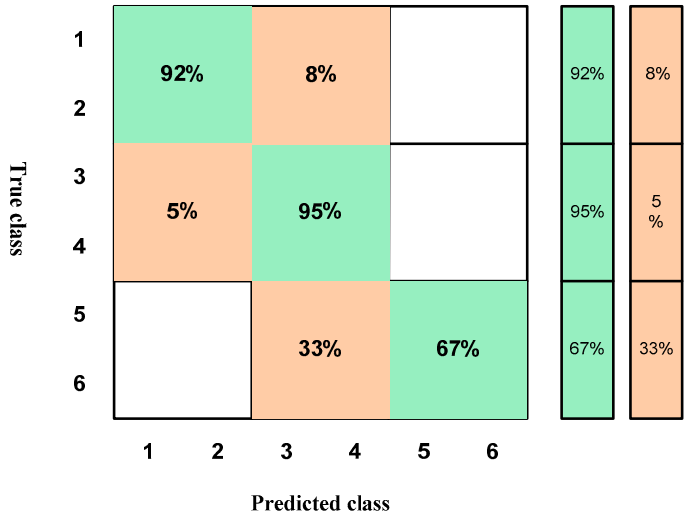


Figure 8. The prediction accuracy of the linear SVM with data normalization.

Table 6. The prediction accuracy of the constructed model.

Inputs Data																Actual WQI	Predicted WQI
EC	PH	NTU	TDS	ALK.	TH	HCO_3^-	Cl^-	SO_4^{2-}	Ca^{2+}	Mg^{2+}	Na^+	K^+	NO_3^-	F^-	Fe	As	Code	Prediction
0.18	0.9	0.5	0.15	0.2	0.18	0.2	0.12	0.1	0.2	0.1	0.1	0.2	0.2	0.2	0.17	0	1	1
0.18	1	0.6	0.13	0.2	0.19	0.2	0.09	0.1	0.2	0.1	0.1	0.1	0.2	0.2	0.11	0.5	1	1
0.2	0.9	0.5	0.22	0.2	0.26	0.2	0.24	0.2	0.3	0.1	0.2	0.1	0.3	0.4	0.14	0	2	2
0.23	0.9	0.3	0.2	0.2	0.25	0.2	0.18	0.2	0.2	0.1	0.1	0.1	0.2	0.4	0.17	0.5	2	2
0.27	0.8	0.8	0.22	0.2	0.25	0.2	0.23	0.2	0.2	0.1	0.1	0.1	0.3	0.3	0.14	0.5	2	2
0.35	0.8	0.5	0.27	0.4	0.34	0.4	0.22	0.2	0.3	0.2	0.2	0.1	0.3	0.3	0.09	0.5	2	2
0.24	0.9	0.8	0.35	0.4	0.41	0.4	0.33	0.2	0.4	0.2	0.3	0.3	0.3	0.3	1.23	0	2	2
0.34	0.9	0.8	0.24	0.4	0.27	0.3	0.24	0.2	0.3	0.1	0.2	0.2	0.3	0.3	0.17	0.5	2	2
0.25	0.9	0.6	0.34	0.4	0.43	0.4	0.3	0.3	0.4	0.2	0.2	0.3	0.2	0.3	0.14	1	2	2
0.24	0.9	0.8	0.25	0.3	0.31	0.3	0.28	0.2	0.3	0.2	0.2	0.2	0.2	0.1	0	0	2	2
0.2	0.8	0.9	0.23	0.2	0.32	0.3	0.21	0.2	0.2	0.2	0.1	0.5	0.3	0.2	0.09	0.5	2	2
0.16	0.9	0.6	0.24	0.2	0.24	0.2	0.17	0.2	0.2	0.1	0.2	0.1	0.2	0.2	0.09	0.5	2	2
0.16	0.9	0.3	0.2	0.1	0.13	0.1	0.16	0.2	0.1	0.1	0.1	0.1	0.3	0.2	0.06	0	2	1
0.65	0.9	0.7	0.44	0.5	0.62	0.6	0.52	0.4	0.8	0.3	0.5	0.2	0.9	0.3	0.09	0	3	3
0.27	0.9	0.8	0.55	0.1	0.7	0.7	0.67	0.5	0.3	0.3	0.6	0.2	0.2	0.6	0.63	0	3	3
																		93.33

5. Conclusions

In this work, the classification learner tool in MATLAB was used to construct a prediction model to reduce the computation time necessary to determine the WQI state for water samples. The data samples were used in their raw and normalized forms to investigate which one can produce the highest prediction accuracy with the classifiers in the classification learner tool. The results based on the training data indicate that the linear SVM is the best classifier with the two data forms. The prediction accuracies with the raw and normalized forms were 90.8 and 89.2%, respectively. On the other hand, the prediction accuracy for the testing data samples with normalized data developed higher accuracy than with raw testing data. The prediction accuracy of the testing data was 93.33% (14/15), while it was 86.67% with the testing of the raw data. The prediction accuracy of the testing data shows that the constructed linear SVM can predict reasonably well and can be used for other new data. Finally, the results indicate that the linear SVM model can predict the WQI code for new data samples with reasonable precision.

Author Contributions: Conceptualization, E.E.H. and M.Y.J.B.; data curation, E.E.H., M.Y.J.B. and A.N.; formal analysis, E.E.H. and M.Y.J.B.; funding acquisition, E.E.H., H.F.A., F.K.A. and E.T.; investigation, E.E.H., M.Y.J.B., H.F.A., F.K.A. and E.T.; methodology, E.E.H., M.Y.J.B. and A.N.; project administration, E.E.H. and M.Y.J.B.; resources, E.E.H., M.Y.J.B., A.N., H.F.A. F.K.A. and E.T.; software, E.E.H. and M.Y.J.B.; supervision, M.Y.J.B.; validation, E.E.H., M.Y.J.B., A.N., H.F.A., F.K.A. and E.T.; visualization, E.E.H., M.Y.J.B., A.N., H.F.A., F.K.A. and E.T.; writing—original draft preparation, E.E.H. and M.Y.J.B.; writing—review and editing, E.E.H., M.Y.J.B., A.N., H.F.A., F.K.A. and E.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Deanship of Scientific Research, Taif University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the first or corresponding author.

Acknowledgments: The authors would like to acknowledge the Deanship of Scientific Research, Taif University, for funding this work.

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

References

1. Dilpazeer, F.; Munir, M.; Baloch, M.Y.J.; Shafiq, I.; Iqbal, J.; Saeed, M.; Abbas, M.M.; Shafique, S.; Aziz, K.H.H.; Mustafa, A.; et al. A Comprehensive Review of the Latest Advancements in Controlling Arsenic Contaminants in Groundwater. *Water* **2023**, *15*, 478. [CrossRef]
2. Jat Baloch, M.Y.; Zhang, W.; Zhang, D.; Al Shoumik, B.A.; Iqbal, J.; Li, S.; Chai, J.; Farooq, M.A.; Parkash, A. Evolution Mechanism of Arsenic Enrichment in Groundwater and Associated Health Risks in Southern Punjab, Pakistan. *Int. J. Environ. Res. Public Health* **2022**, *19*, 13325. [CrossRef]
3. Li, S.; Zhang, W.; Zhang, D.; Xiu, W.; Wu, S.; Chai, J.; Ma, J.; Baloch, M.Y.J.; Sun, S.; Yang, Y. Migration risk of *Escherichia coli* O157: H7 in unsaturated porous media in response to different colloid types and compositions. *Environ. Pollut.* **2023**, *323*, 121282. [CrossRef] [PubMed]
4. Baloch, M.Y.J.; Talpur, S.A.; Talpur, H.A.; Iqbal, J.; Mangi, S.H.; Memon, S. Effects of Arsenic Toxicity on the Environment and Its Remediation Techniques: A Review. *J. Water Environ. Technol.* **2020**, *18*, 275–289. [CrossRef]
5. Tariq, A.; Mumtaz, F.; Zeng, X.; Baloch, M.Y.J.; Moazzam, M.F.U. Spatio-temporal variation of seasonal heat islands mapping of Pakistan during 2000–2019, using day-time and night-time land surface temperatures MODIS and meteorological stations data. *Remote Sens. Appl. Soc. Environ.* **2022**, *27*, 100779. [CrossRef]
6. Tariq, A.; Ali, S.; Basit, I.; Jamil, A.; Farmonov, N.; Khorrami, B.; Khan, M.M.; Sadri, S.; Baloch, M.Y.J.; Islam, F. Terrestrial and groundwater storage characteristics and their quantification in the Chitral (Pakistan) and Kabul (Afghanistan) river basins using GRACE/GRACE-FO satellite data. *Groundw. Sustain. Dev.* **2023**, *23*, 100990. [CrossRef]
7. Jat Baloch, M.Y.; Su, C.; Talpur, S.A.; Iqbal, J.; Bajwa, K. Arsenic Removal from Groundwater Using Iron Pyrite: Influence Factors and Removal Mechanism. *J. Earth Sci.* **2023**, *34*, 857–867. [CrossRef]
8. Stojanović Bjelić, L.; Ilić, P.; Nešković Markić, D.; Ilić, S.; Popović, Z.; Mrazovac Kurilić, S.; Mihajlović, D.; Farooqi, Z.; Jat Baloch, M.; Mohamed, M. Contamination in water and ecological risk of heavy metals near a coal mine and a thermal power plant (republic of srpska, bosnia and herzegovina). *Appl. Ecol. Environ. Res.* **2023**, *21*, 3807–3822.
9. Jat Baloch, M.Y.; Zhang, W.; Chai, J.; Li, S.; Alqurashi, M.; Rehman, G.; Tariq, A.; Talpur, S.A.; Iqbal, J.; Munir, M.; et al. Shallow groundwater quality assessment and its suitability analysis for drinking and irrigation purposes. *Water* **2021**, *13*, 3361. [CrossRef]
10. Baloch, M.Y.J.; Mangi, S.H. Treatment of synthetic greywater by using banana, orange and sapodilla peels as a low cost activated carbon. *J. Mater. Environ. Sci.* **2019**, *10*, 966–986.
11. Kumar, R.; Parkash, A.; Almani, S.; Baloch, M.Y.J.; Khan, R.; Soomro, S.A. Synthesis of porous cobalt oxide nanosheets: Highly sensitive sensors for the detection of hydrazine. *Funct. Compos. Struct.* **2022**, *4*, 035002. [CrossRef]
12. Zhang, W.; Chai, J.; Li, S.; Wang, X.; Wu, S.; Liang, Z.; Baloch, M.Y.J.; Silva, L.F.; Zhang, D. Physiological characteristics, geochemical properties and hydrological variables influencing pathogen migration in subsurface system: What we know or not? *Geosci. Front.* **2022**, *13*, 101346. [CrossRef]
13. Zhang, W.; Zhu, Y.; Gu, R.; Liang, Z.; Xu, W.; Jat Baloch, M.Y. Health Risk Assessment during In Situ Remediation of Cr (VI)-Contaminated Groundwater by Permeable Reactive Barriers: A Field-Scale Study. *Int. J. Environ. Res. Public Health* **2022**, *19*, 13079. [CrossRef] [PubMed]
14. Iqbal, J.; Su, C.; Wang, M.; Abbas, H.; Baloch, M.Y.J.; Ghani, J.; Ullah, Z.; Huq, M.E. Groundwater fluoride and nitrate contamination and associated human health risk assessment in South Punjab, Pakistan. *Environ. Sci. Pollut. Res.* **2023**, *30*, 61606–61625. [CrossRef] [PubMed]

15. Iqbal, J.; Amin, G.; Su, C.; Haroon, E.; Baloch, M.Y.J. Assessment of landcover impacts on the groundwater quality using hydrogeochemical and geospatial techniques. *Environ. Sci. Pollut. Res.* 2023, *in press*. [CrossRef] [PubMed]
16. Khalid, W.; Jat Baloch, M.Y.; Ali, A.; Ngata, M.R.; Alrefaei, A.F.; Rashid, A.; Ilić, P.; Almutairi, M.H.; Siddique, J. Groundwater Contamination and Risk Assessment in Greater Palm Springs. *Water* **2023**, *15*, 3099. [CrossRef]
17. Khan, Z.; Ali, S.A.; Mohsin, M.; Shamim, S.K.; Mankovskaya, E.; Parvin, F.; Bano, N.; Ahmad, A.; Jat Baloch, M.Y. Estimating Photosynthetically Active Euphotic Layer in Major Lakes of Kumaun Region Using Secchi Depth. *Water Air Soil Pollut.* **2023**, *234*, 597. [CrossRef]
18. Baloch, H.; Kandhro, B.; Channa, A.; Memon, S.A.; Jat Baloch, M.Y. Enhancement of Biogas Production from Fixed Dome Biogas Plant through Recycling of Digested Slurry. *Int. J. Environ. Sci. Nat. Resour.* **2022**, *29*, 556274.
19. Jat Baloch, M.Y.; Talpur, S.A.; Iqbal, J.; Munir, M.; Bajwa, K.; Baidya, P.; Talpur, H.A. Review Paper Process Design for Biohydrogen Production from Waste Materials and Its Application. *Sustain. Environ.* **2022**, *7*, 2022. [CrossRef]
20. Basharat, U.; Zhang, W.; Baloch, M.Y.J.; Abbasi, A.; Ali, B.; Khan, S.M.; Khan, S.H.; Niaz, A.; Irshad, M. Review Paper Presence and Dispersion of Organic and Inorganic Contaminants in Groundwater. *Sustain. Environ.* **2023**, *8*, 71. [CrossRef]
21. Jat Baloch, M.Y.; Zhang, W.; Sultana, T.; Akram, M.; Al Shoumik, B.A.; Khan, M.Z.; Farooq, M.A. Utilization of sewage sludge to manage saline-alkali soil and increase crop production: Is it safe or not? *Environ. Technol. Innov.* **2023**, *32*, 103266. [CrossRef]
22. Asghar, I.; Ahmed, M.; Farooq, M.A.; Ishtiaq, M.; Arshad, M.; Akram, M.; Umair, A.; Alrefaei, A.F.; Jat Baloch, M.Y.; Naeem, A. Characterizing indigenous plant growth promoting bacteria and their synergistic effects with organic and chemical fertilizers on wheat (*Triticum aestivum*). *Front. Plant Sci.* **2023**, *14*, 1232271. [CrossRef] [PubMed]
23. Manzar, M.S.; Benaafi, M.; Costache, R.; Alagha, O.; Mu'azu, N.D.; Zubair, M.; Abdullahi, J.; Abba, S. New generation neurocomputing learning coupled with a hybrid neuro-fuzzy model for quantifying water quality index variable: A case study from Saudi Arabia. *Ecol. Inform.* **2022**, *70*, 101696. [CrossRef]
24. Masood, N.; Farooqi, A.; Zafar, M.I. Health risk assessment of arsenic and other potentially toxic elements in drinking water from an industrial zone of Gujrat, Pakistan: A case study. *Environ. Monit. Assess.* **2019**, *191*, 1–15. [CrossRef] [PubMed]
25. Qureshi, A.; Lashari, B.; Kori, S.; Lashari, G. Hydro-salinity behavior of shallow groundwater aquifer underlain by salty groundwater in Sindh Pakistan. In Proceedings of the Fifteenth International Water Technology Conference, Alexandria, Egypt, 28–30 May 2011; pp. 1–15.
26. Iqbal, J.; Su, C.; Rashid, A.; Yang, N.; Jat Baloch, M.Y.; Talpur, S.A.; Ullah, Z.; Rahman, G.; Rahman, N.U.; Sajjad, M.M. Hydrogeochemical assessment of groundwater and suitability analysis for domestic and agricultural utility in Southern Punjab, Pakistan. *Water* **2021**, *13*, 3589. [CrossRef]
27. Talpur, S.A.; Noonari, T.M.; Rashid, A.; Ahmed, A.; Jat Baloch, M.Y.; Talpur, H.A.; Soomro, M.H. Hydrogeochemical signatures and suitability assessment of groundwater with elevated fluoride in unconfined aquifers Badin district, Sindh, Pakistan. *SN Appl. Sci.* **2020**, *2*, 1038. [CrossRef]
28. Baig, J.A.; Kazi, T.G.; Arain, M.B.; Afridi, H.I.; Kandhro, G.A.; Sarfraz, R.A.; Jamal, M.K.; Shah, A.Q. Evaluation of arsenic and other physico-chemical parameters of surface and ground water of Jamshoro, Pakistan. *J. Hazard. Mater.* **2009**, *166*, 662–669. [CrossRef]
29. Arain, M.; Kazi, T.; Baig, J.; Jamali, M.; Afridi, H.; Shah, A.; Jalbani, N.; Sarfraz, R.A. Determination of arsenic levels in lake water, sediment, and foodstuff from selected area of Sindh, Pakistan: Estimation of daily dietary intake. *Food Chem. Toxicol.* **2009**, *47*, 242–248. [CrossRef]
30. Brahman, K.D.; Kazi, T.G.; Afridi, H.I.; Naseem, S.; Arain, S.S.; Ullah, N. Evaluation of high levels of fluoride, arsenic species and other physicochemical parameters in underground water of two sub districts of Tharparkar, Pakistan: A multivariate study. *Water Res.* **2013**, *47*, 1005–1020. [CrossRef]
31. Arain, M.; Kazi, T.; Jamali, M.; Jalbani, N.; Afridi, H.; Shah, A. Total dissolved and bioavailable elements in water and sediment samples and their accumulation in Oreochromis mossambicus of polluted Manchar Lake. *Chemosphere* **2008**, *70*, 1845–1856. [CrossRef]
32. Gul, M.; Mashhadi, A.F.; Iqbal, Z.; Qureshi, T.I. Monitoring of arsenic in drinking water of high schools and assessment of carcinogenic health risk in Multan, Pakistan. *Hum. Ecol. Risk Assess. Int. J.* **2020**, *26*, 2129–2141. [CrossRef]
33. Sultana, J.; Farooqi, A.; Ali, U. Arsenic concentration variability, health risk assessment, and source identification using multivariate analysis in selected villages of public water system, Lahore, Pakistan. *Environ. Monit. Assess.* **2014**, *186*, 1241–1251. [CrossRef] [PubMed]
34. Shehzad, M.T.; Sabir, M.; Zia-ur-Rehman, M.; Zia, M.A.; Naidu, R. Arsenic concentrations in soil, water, and rice grains of rice-growing areas of Punjab, Pakistan: Multivariate statistical analysis. *Environ. Monit. Assess.* **2022**, *194*, 346. [CrossRef] [PubMed]
35. Rasheed, H.; Iqbal, N.; Ashraf, M.; ul Hasan, F. Groundwater quality and availability assessment: A case study of District Jhelum in the Upper Indus, Pakistan. *Environ. Adv.* **2022**, *7*, 100148. [CrossRef]
36. Abbas, M.; Shen, S.-L.; Lyu, H.-M.; Zhou, A.; Rashid, S. Evaluation of the hydrochemistry of groundwater at Jhelum Basin, Punjab, Pakistan. *Environ. Earth Sci.* **2021**, *80*, 300. [CrossRef]
37. Abbas, M.; Cheema, K.J. Arsenic levels in drinking water and associated health risk in district Sheikhupura, Pakistan. *J. Anim. Plant Sci.* **2015**, *25*, 719–724.

38. Baloch, M.Y.J.; Zhang, W.; Al Shoumik, B.A.; Nigar, A.; Elhassan, A.A.; Elshekh, A.E.; Bashir, M.O.; Ebrahim, A.F.M.S.; Iqbal, J. Hydrogeochemical mechanism associated with land use land cover indices using geospatial, remote sensing techniques, and health risks model. *Sustainability* **2022**, *14*, 16768. [CrossRef]
39. Daud, M.; Nafees, M.; Ali, S.; Rizwan, M.; Bajwa, R.A.; Shakoor, M.B.; Arshad, M.U.; Chatha, S.A.S.; Deeba, F.; Murad, W. Drinking water quality status and contamination in Pakistan. *BioMed Res. Int.* **2017**, *2017*, 7908183. [CrossRef]
40. Hussein, E.E.; Fouad, M.; Gad, M.I. Prediction of the pollutants movements from the polluted industrial zone in 10th of Ramadan city to the Quaternary aquifer. *Appl. Water Sci.* **2019**, *9*, 20. [CrossRef]
41. Slukovskii, Z.; Dauvalter, V.; Guzeva, A.; Denisov, D.; Cherepanov, A.; Siroezhko, E. The hydrochemistry and recent sediment geochemistry of small lakes of Murmansk, Arctic Zone of Russia. *Water* **2020**, *12*, 1130. [CrossRef]
42. Islam, N.; Irshad, K. Artificial ecosystem optimization with Deep Learning Enabled Water Quality Prediction and Classification model. *Chemosphere* **2022**, *309*, 136615. [CrossRef]
43. Durango-Cordero, J.; Saqalli, M.; Ferrant, S.; Bonilla, S.; Maurice, L.; Arellano, P.; Elger, A. Risk assessment of unlined oil pits leaking into groundwater in the Ecuadorian Amazon: A modified GIS-DRASTIC approach. *Appl. Geogr.* **2022**, *139*, 102628. [CrossRef]
44. Gidey, A. Geospatial distribution modeling and determining suitability of groundwater quality for irrigation purpose using geospatial methods and water quality index (WQI) in Northern Ethiopia. *Appl. Water Sci.* **2018**, *8*, 82. [CrossRef]
45. Abdessamed, D.; Jodar-Abellan, A.; Ghoneim, S.S.; Almaliki, A.; Hussein, E.E.; Pardo, M.Á. Groundwater quality assessment for sustainable human consumption in arid areas based on GIS and water quality index in the watershed of Ain Sefra (SW of Algeria). *Environ. Earth Sci.* **2023**, *82*, 510. [CrossRef]
46. Uddin, M.G.; Nash, S.; Rahman, A.; Olbert, A.I. A novel approach for estimating and predicting uncertainty in water quality index model using machine learning approaches. *Water Res.* **2023**, *229*, 119422. [CrossRef]
47. Aralu, C.C.; Okoye, P.A.; Abugu, H.O.; Eze, V.C. Pollution and water quality index of boreholes within unlined waste dumpsite in Nnewi, Nigeria. *Discov. Water* **2022**, *2*, 14. [CrossRef]
48. 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]
49. El Bilali, A.; Taleb, A.; Brouziyne, Y. Groundwater quality forecasting using machine learning algorithms for irrigation purposes. *Agric. Water Manag.* **2021**, *245*, 106625. [CrossRef]
50. Elbeltagi, A.; Pande, C.B.; Kouadri, S.; Islam, A.R.M.T. Applications of various data-driven models for the prediction of groundwater quality index in the Akot basin, Maharashtra, India. *Environ. Sci. Pollut. Res.* **2022**, *29*, 17591–17605. [CrossRef]
51. Uddin, M.G.; Nash, S.; Olbert, A.I. A review of water quality index models and their use for assessing surface water quality. *Ecol. Indic.* **2021**, *122*, 107218. [CrossRef]
52. Uddin, M.G.; Nash, S.; Rahman, A.; Olbert, A.I. A comprehensive method for improvement of water quality index (WQI) models for coastal water quality assessment. *Water Res.* **2022**, *219*, 118532. [CrossRef] [PubMed]
53. Uddin, M.G.; Nash, S.; Rahman, A.; Olbert, A.I. Performance analysis of the water quality index model for predicting water state using machine learning techniques. *Process Saf. Environ. Prot.* **2023**, *169*, 808–828. [CrossRef]
54. Mohd Zebaral Hoque, J.; Ab Aziz, N.A.; Alelyani, S.; Mohana, M.; Hosain, M. Improving Water Quality Index Prediction Using Regression Learning Models. *Int. J. Environ. Res. Public Health* **2022**, *19*, 13702. [CrossRef] [PubMed]
55. Derdour, A.; Jodar-Abellan, A.; Pardo, M.Á.; Ghoneim, S.S.; Hussein, E.E. Designing Efficient and Sustainable Predictions of Water Quality Indexes at the Regional Scale Using Machine Learning Algorithms. *Water* **2022**, *14*, 2801. [CrossRef]
56. Salma, S.; Rehman, S.; Shah, M.A. Rainfall trends in different climate zones of Pakistan. *Pak. J. Meteorol.* **2012**, *9*, 37–47.
57. Qureshi, A.S.; McCornick, P.G.; Qadir, M.; Aslam, Z. Managing salinity and waterlogging in the Indus Basin of Pakistan. *Agric. Water Manag.* **2008**, *95*, 1–10. [CrossRef]
58. Shahab, A.; Shihua, Q.; Rashid, A.; Hasan, F.U.; Sohail, M.T. Evaluation of water quality for drinking and agricultural suitability in the lower Indus plain in Sindh province, Pakistan. *Pol. J. Environ. Stud.* **2016**, *25*, 2563–2574. [CrossRef]
59. Federation, W.E. *American Public Health Association. Standard Methods for the Examination of Water and Wastewater*; American Public Health Association: Washington, DC, USA, 2005; Volume 21.
60. Halder, S.; Das, S.; Basu, S. Use of support vector machine and cellular automata methods to evaluate impact of irrigation project on LULC. *Environ. Monit. Assess.* **2023**, *195*, 3. [CrossRef]
61. Park, J.; Choi, Y.; Byun, J.; Lee, J.; Park, S. Efficient differentially private kernel support vector classifier for multi-class classification. *Inf. Sci.* **2023**, *619*, 889–907. [CrossRef]
62. Benmahamed, Y.; Kherif, O.; Teguar, M.; Boubakeur, A.; Ghoneim, S.S. Accuracy improvement of transformer faults diagnostic based on DGA data using SVM-BA classifier. *Energies* **2021**, *14*, 2970. [CrossRef]