*environmental and earth sciences proceedings*

MDPI

*Proceeding Paper*

# Water Quality Classification in Terms of WQI Using Machine Learning Algorithms in Keenjhar Lake, Pakistan †

**Muhammad Imran [1], Danrong Zhang [1,*], Muhammad Zaman [2], Shazia Parveen [3] and Nur E Jannat Mishu [4]**

[1] College of Hydrology and Water Resources Engineering, Hohai University, Nanjing 210098, China; m.imrankbr@gmail.com
[2] Department of Irrigation and Drainage, University of Agriculture, Faisalabad 38000, Pakistan; muhammad.zaman@uaf.edu.pk
[3] Department of Biochemistry, Bahauddin Zakariya University, Multan 60000, Pakistan; shizakhan5505@gmail.com
[4] College of Information Science and Engineering, Hohai University, Changzhou 213022, China; engrmishu.bd68@yahoo.com
* Correspondence: danrong_zhang@hhu.edu.cn
† Presented at the 8th International Electronic Conference on Water Sciences, 14–16 October 2024; Available online: https://sciforum.net/event/ECWS-8.

**Abstract:** Water quality management is critical for sustainable resource utilization, public health protection, and environmental conservation. This study employs machine learning techniques to enhance the water quality classification (WQC) for Keenjhar Lake, Karachi, Pakistan, using a comprehensive dataset spanning from 1993 to 2022. Three machine learning models, the Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF) algorithms, were evaluated for their performance on a dataset comprising 360 monthly instances and 6 key water quality parameters: temperature, pH, dissolved oxygen, electrical conductivity, BOD5, nitrate, fecal coliform, and total coliform. The models were assessed using statistical metrics, including accuracy, precision, recall, Pearson's correlation, and F1-score. Among the models, the SVM algorithm demonstrated superior performance, achieving a classification accuracy of 99.50% for WQC. This study highlights the integration of machine learning in water quality analysis as a novel approach for precise classification, offering insights into improved decision-making for water resource management and treatment strategies.

**Keywords:** water quality index; water quality classification; machine learning algorithms; Keenjhar lake; Pakistan

check for updates

Academic Editor: Junye Wang

Published: 7 April 2025

**Citation:** Imran, M.; Zhang, D.; Zaman, M.; Parveen, S.; Mishu, N.E.J. Water Quality Classification in Terms of WQI Using Machine Learning Algorithms in Keenjhar Lake, Pakistan. *Environ. Earth Sci. Proc.* **2025**, *32*, 13. https://doi.org/10.3390/eesp2025032013

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Water quality classification (WQC) is essential for monitoring the condition of water bodies. It helps to identify potential contaminants and ensure water source safety. The Water Quality Index (WQI) is an effective tool to assess overall water quality. It considers various parameters such as temperature, pH levels, dissolved oxygen, turbidity, and levels of pollutants. These parameters are assigned specific weights and scores, which are then combined to calculate the overall WQI [1]. Consequently, there is a pressing need to develop a comprehensive and standardized approach to calculating the water quality classification for drinking water, ensuring that it accurately reflects potential health risks and facilitates informed decision-making for both regulatory bodies and consumers [2].

Machine learning (ML) is an essential component of artificial intelligence (AI) that enables systems to automatically learn and improve from experience without the need for

explicit programming [3]. ML techniques use in-depth data analysis to identify patterns and adjust accordingly [4,5]. In water quality (WQ) studies, ML provides significant potential for assessing, classifying, and predicting WQ indicators. For instance, given sufficient data availability, ML models can accurately simulate hydrological processes and contaminant transport [6].

In recent years, researchers have demonstrated the feasibility and effectiveness of artificial intelligence (AI) in estimating WQ. Liao and Sun [7] combined artificial neural networks (ANN) with decision tree algorithms to predict WQ. Other researchers [8] developed a deep learning network model, outperforming supervised learning-based techniques in forecasting water-dissolved oxygen and pH levels. To classify WQ, Shafi et al. [9] applied four machine learning (ML) algorithms—SVM, NN, DNN, and KNN—across various water bodies, finding that the DNN algorithm achieved the highest accuracy at 93%. A recent study explored the application of machine learning models in predicting water quality parameters, focusing on the relationships between the physical attributes of the catchment area and the in-stream water quality. Using data from the Iran Water Resource Management Company, the study examined 11 water quality parameters over 5 years (1998–2002) and evaluated model performance through various metrics, including Pearson's Correlation Coefficient and MAPE. The Random Forest model demonstrated superior performance across the parameters [10].

These studies underscore the increasing adoption of ML for water quality prediction and classification. They also highlight the growing recognition of its capability to handle complex, non-linear relationships and large datasets, making it a valuable tool for real-time water quality monitoring. This research extends the current knowledge by employing three machine learning algorithms, namely the SVM, KNN, and RF algorithms, to predict and classify the water quality in Keenjhar Lake, Karachi, Pakistan. By utilizing data from 1993 to 2022, this study seeks to enhance the accuracy of water quality classification and provide a more reliable solution for water management in the region.

## 2. Material and Methods

This study used three machine learning algorithms, namely the SVM, KNN, and RF algorithms, to predict the classification of drinking water for human beings. Machine learning is one of the most remarkable and fast-growing fields in computer science. Machine learning is built on practical algorithms that use tools and functions to deal with massive and composite datasets. This study divided the WQI into different classes, namely excellent, good, poor, very poor, and unfit. This study used performance evaluation metrics such as sensitivity, specificity, F1-score, and accuracy to find the best-fitting algorithm. The proposed ML algorithms are shown in Figure 1.

### *2.1. Study Area and Dataset Collection*

In this study, we selected Keenjhar Lake as our research area, as it is one of the primary sources of drinking water for approximately 1.8 million people in the city of Karachi and parts of the Thatta district, Sindh, Pakistan [11], as shown in Figure 2.

- Surface area~13,468 ha (33,280 acres);
- Max. L~15 miles and Max. W~3.7 miles;
- Water Volume~0.53 $\times$ $10^6$ acre. ft (650 $hm^3$).

Table 1 shows the data collected from a lake at several locations between 1993 and 2022.

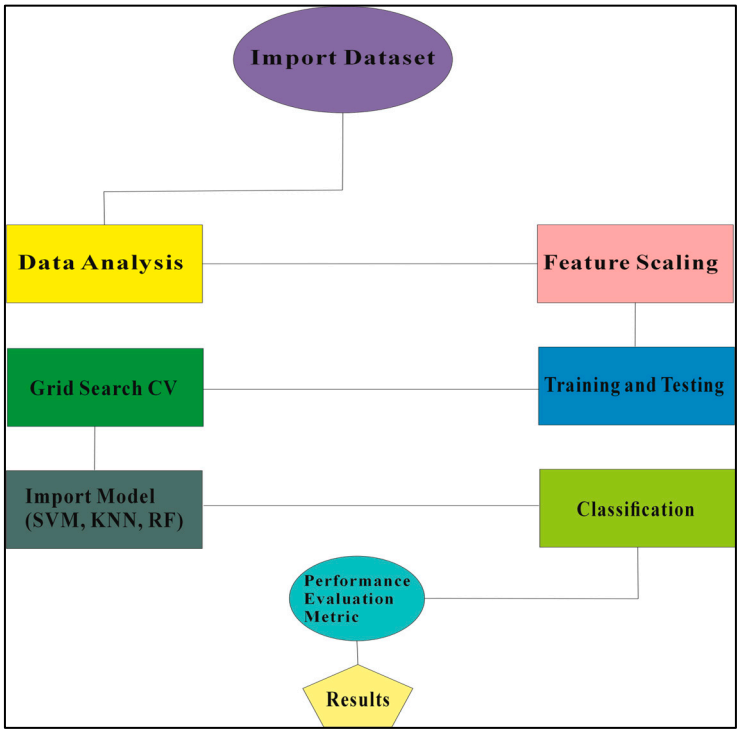



**Figure 1.** Proposed ML algorithms for this study.

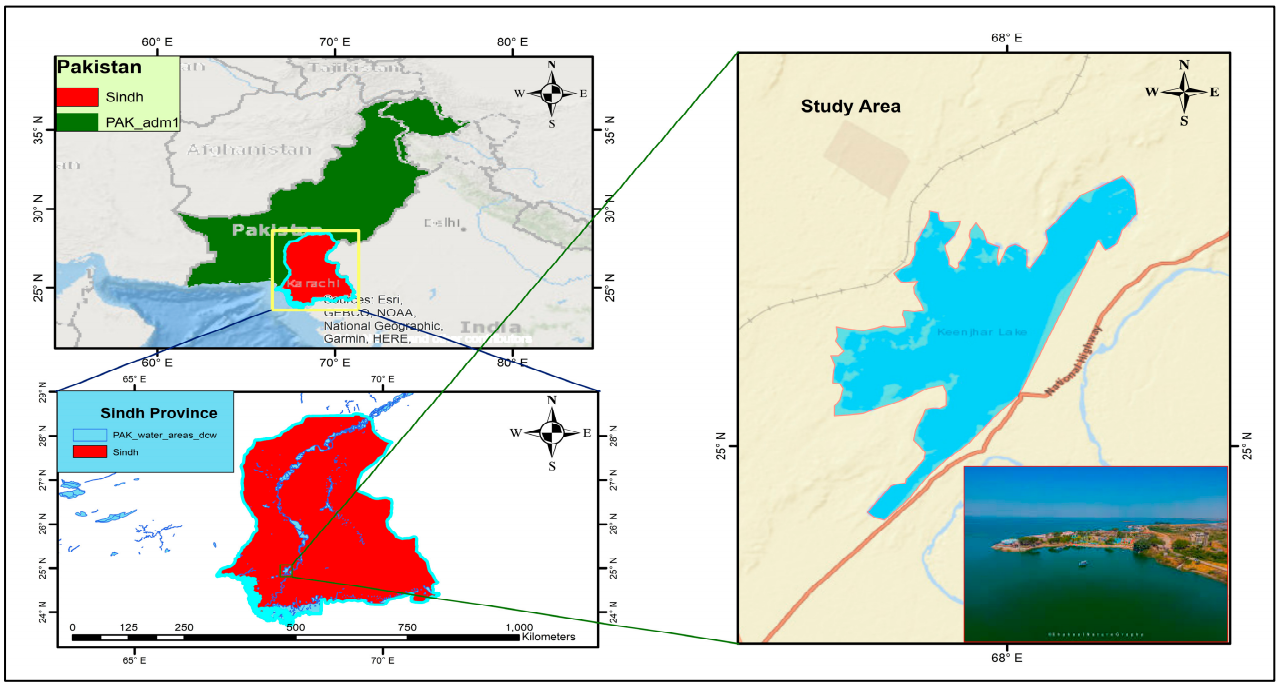



**Figure 2.** Study Area—Keenjhar Lake, Karachi, Pakistan.

The Irrigation Department of Sindh (IDS), Pakistan, collected these data to ensure that the water was safe for drinking.

The dataset was split into two subsets: a training set (80%) and a test set (20%). To evaluate the representativeness of both sets, statistical characteristics such as mean, maximum, minimum, and standard deviation values were calculated for the input variables (e.g., dissolved oxygen, pH, and turbidity) and the output variable (e.g., Water Quality Index). The statistical properties of the training and test sets were compared to ensure that

they were statistically similar, providing a fair representation of the overall dataset. Table 2 presents these statistics for the training and testing sets.

**Table 1.** Description of the input/output dataset.

| Sr. No | Input Data | Unit | Duration | Data Source | Output |
|---|---|---|---|---|---|
| 1 | Temperature | °C | 1993–2022 | IDS | WQC |
| 2 | PH | - | 1993–2022 | IDS | |
| 3 | DO | mg/L | 1993–2022 | IDS | |
| 4 | EC | µS/cm | 1993–2022 | IDS | |
| 5 | BOD5 | mg/L | 1993–2022 | IDS | |
| 6 | Nitrate | mg/L | 1993–2022 | IDS | |

**Table 2.** Description of the dataset.

| | T °C | PH | DO | EC | BOD5 | Nitrate | WQI |
|---|---|---|---|---|---|---|---|
| Mean | 25.95 | 7.51 | 13.77 | 657.53 | 2.20 | 1.93 | 54.05 |
| Std | 1.5 | 0.49 | 1.75 | 62.01 | 1.07 | 2.73 | 27.91 |
| Min | 21.8 | 6.64 | 1.45 | 203 | 0.13 | 0.00 | 4.20 |
| Max | 31 | 9.13 | 17.53 | 1002 | 7.41 | 45.5 | 168.42 |

### *2.2. Water Quality Index (WQI) Calculations*

The Water Quality Index (WQI) is determined by averaging the individual index values of selected parameters [12]. In this study, the WQI scores were used to classify the water samples, with the following formula applied to calculate the WQI, written in Equations (1)–(4):

$$\mathrm{WQI} = \frac{\sum_{i=1}^{N} qi * wi}{\sum_{i=1}^{N} wi} \quad (1)$$

$$qi = 100 * \left( \frac{Vi - Videal}{Si - Videal} \right) \quad (2)$$

$$wi = \frac{K}{Si} \quad (3)$$

$$\mathrm{K} = \frac{1}{\sum_{i=1}^{N} Si} \quad (4)$$

where $wi$ is the weight unit determined for each parameter, as indicated in Equation (3) below, $N$ is the number of parameters employed in the water computation, and $qi$ is the quality rating scale of each parameter $qi$, determined by Equation (2).

Where $Vi$ stands for the measured value of the parameter $i$ in the water samples under investigation, $Si$ is the suggested standard value for the parameter $i$. Clean water with all parameters set to 0, except pH, which is set to 7.0, and dissolved oxygen (DO), which is set to 14.6 mg/L, is ideal. $K$, which stands for the proportionality constant, is described in Equation (4).

Water quality classification (WQC) is essential in predicting the precise results from the Water Quality Index. The WQI values were categorized into five water quality classes: a range of 0–25 represents excellent water quality, 26–50 indicates good water quality, 51–75 corresponds to poor water quality, 76–100 denotes very poor water quality, and values above 100 indicate water that is unfit for drinking purposes [13]. Figure 3 shows the WQC in our dataset.

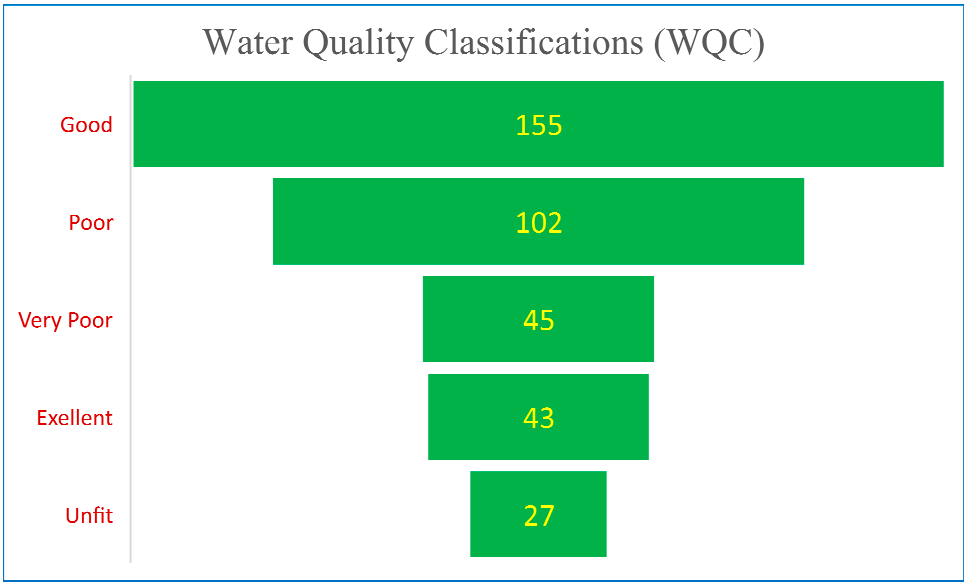



**Figure 3.** Count of WQC based on the WQI of the dataset.

### *2.3. Methods*

#### 2.3.1. Support Vector Machine

The Support Vector Machine (SVM) algorithm is a supervised machine learning algorithm primarily used for classification and regression tasks. This study applied the SVM algorithm to classify water quality by finding the optimal hyperplane that separates the data points of different classes in a multidimensional feature space. The algorithm aims to maximize the margin, which is the distance between the hyperplane and the nearest data points (support vectors) from each class, as shown in Figure 4.

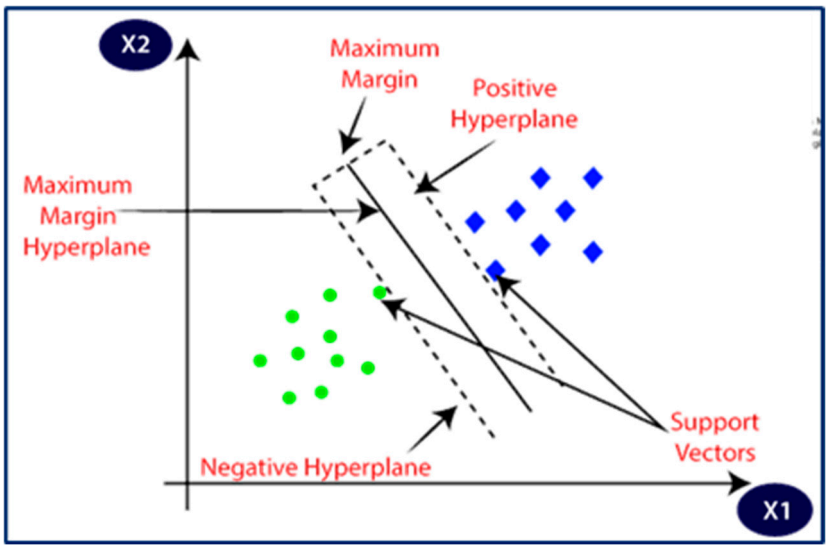



**Figure 4.** Support Vector Machine algorithm sketch.

Mathematically, for a given dataset $(x_i, y_i)$, where $(x_i \epsilon \mathbb{R}^n)$ represents the input features and $y_i \in \{-1, 1\}$ represents the class labels, the SVM algorithm solves the following optimization problem:

$$min_{w,b} \frac{1}{2} \| w \|^2$$

This equation is subject to the following:

$$y = y_i\left(w^T x_i + b\right) \geq 1, \forall_i$$

Here, $w$ is the weight vector defining the orientation of the hyperplane and is also the bias term. The goal is to minimize $\frac{1}{2}\| w \|^2$, which maximizes the margin between classes.

In cases where the data are not linearly separable, a kernel function ϕ(x) is introduced to map the input features into a higher-dimensional space where a linear separation is feasible. Commonly used kernel functions include the following:

- Linear kernel = K $(x_i, y_i) = x_i^T x_j$;
- Polynomial kernel = K $(x_i, y_i) = \left(x_i^T x_j + c\right)^d$;
- Radial Basis Function (RBF) kernel = K $(x_i, y_i) = exp\left(-\gamma \| x_i - x_j \|^2\right)$.

The RBF kernel was selected for this study because it can handle non-linear relationships between the water quality parameters. The regularization parameter C was tuned to balance the trade-off between maximizing the margin and minimizing the classification errors.

y = 0 (x lies on hyperplane)

y < 0 (x lies on negative side of hyperplane)

y > 0 (x lies on positive side of hyperplane)

For all green points: $w \times x + b \leq -1$.
For all blue points: $w \times x + b \geq 1$.

2.3.2. K-Nearest Neighbor

The K-Nearest Neighbor (KNN) algorithm is a non-parametric and instance-based supervised machine learning algorithm for classification and regression tasks. In this study, the KNN algorithm was employed to classify water quality by comparing new data points with existing labeled data points based on similarity. The algorithm classifies a new data point by assigning it to the class most frequently represented among its K-Nearest neighbors, where k is a user-defined parameter.

The similarity between data points is measured using the Euclidean distance, which quantifies the straight-line distance between two points in an n-dimensional feature space.

$$D = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (5)$$

$x_i$= Coordinate point A.
$y_i$ = Coordinate point B.
D = Distance b/w two points A and B.

As shown in Equation (5), this formula determines the distance between a new data point and all training data points. The k closest data points are then identified, and the majority class among these neighbors is assigned to the new data point.

In this study, the value of k was optimized by evaluating the model's performance across different values, and the optimal k was selected based on accuracy and other performance metrics.

- Figure 5a illustrates how the KNN algorithm groups new variables into classes based on their similarity to existing data points.

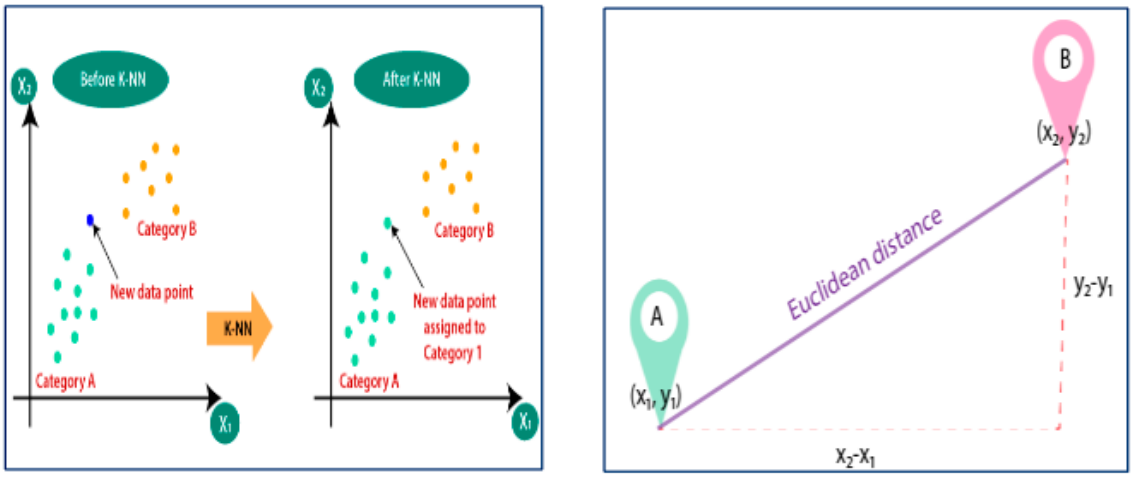



**Figure 5.** (**a**) KNN algorithm sketch. (**b**) Euclidean distance.

- Figure 5b demonstrates the calculation of Euclidean distance, visually representing the process of identifying the closest neighbors in the feature space.

The KNN algorithm is computationally simple and intuitive; its performance is highly dependent on the choice of k, the distance metric, and the scale of the data. To ensure optimal performance, the dataset was standardized to avoid biases introduced by differing scales among the features.

#### 2.3.3. Random Forest

The Random Forest (RF) algorithm is a supervised learning algorithm widely used for classification and regression tasks, due to its robustness and ability to handle high-dimensional data. In this study, the RF algorithm was applied to classify water quality by creating an ensemble of decision trees, where each tree contributed to the final classification based on majority voting.

The algorithm works by constructing multiple decision trees during the training phase. Each tree is built using a subset of the training data and a random subset of features, a process known as bootstrapping, and feature randomness is shown in Figure 6. This reduces the likelihood of overfitting and ensures diversity among the trees in the forest, thereby improving the overall model accuracy.

- Number of Estimators (n_estimators): This parameter specifies the number of decision trees in the forest. In this study, n_estimators were set to 100 after tuning, which provided a balance between computational efficiency and model accuracy.
- Maximum Features (max_features): This parameter determines the number of features considered for splitting at each node. Setting max_features to $\sqrt{m}$, where m is the total number of features, ensures that each tree is trained on a diverse subset of features, improving model generalization.

The RF algorithm classifies data by aggregating the outputs of all decision trees. For a given input x, the prediction is determined as follows:

$$y' = majority\ vote\ (y_1, y_2, \ldots\ldots\ldots\ldots, y_k)$$

where $y_1, y_2, \ldots\ldots\ldots\ldots, y_k$ are the predictions from individual trees.

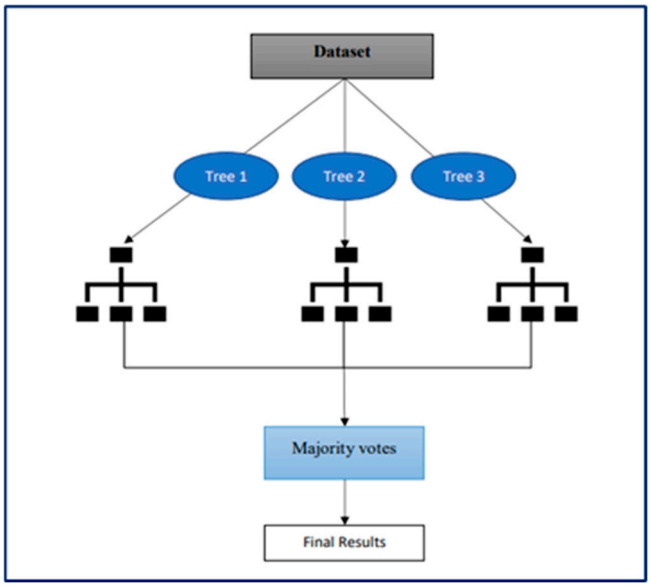



**Figure 6.** Random Forest algorithm sketch.

### *2.4. Performance Evaluation Matrix*

The performance of the classification models was assessed using confusion matrices, which are instrumental in evaluating the predictive accuracy of machine learning algorithms. A confusion matrix provides a detailed breakdown of the model's predictions by showing the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) in comparison to the actual labels in the test dataset.

Several key performance metrics were calculated from the confusion matrix, as outlined in Table 3.

**Table 3.** Results of proposed algorithms.

| | SVM (%) | KNN (%) | RF (%) |
|---|---|---|---|
| Precision | 94.57 | 45.72 | 91.99 |
| Recall | 93.73 | 45.09 | 94.28 |
| F1-score | 93.63 | 44.79 | 91.66 |
| Accuracy | 99.5 | 58 | 95 |

#### 2.4.1. Precision (Specificity)

Precision is calculated by dividing the number of true positives by the total number of positives. Mathematically, it is expressed as follows:

$$\text{Precision (P)} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

#### 2.4.2. Recall (Sensitivity)

Recall shows how well our algorithm can pick out true positives. Mathematically, it can be written as follows:

$$\text{Recall (R)} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

#### 2.4.3. F1-Score

The F1-score is calculated as the weighted average of precision and recall. Unlike accuracy, which may not effectively account for false positives and false negatives, the

F1-score provides a more balanced measure, especially when the class distribution is uneven. This makes it more useful in many scenarios. It is expressed as follows:

$$\text{F1-Score} = 2 * \frac{\text{Precisoin} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

2.4.4. Accuracy

A machine learning algorithm's accuracy can be measured by comparing it to the training data to determine which is most effective at identifying correlations and patterns among the variables in the dataset. It is written as follows:

$$\text{Accuracy} = \frac{\text{True Negative} + \text{True Positive}}{\text{Total}}$$

2.4.5. Pearson's Correlation Matrix Coefficient

The Pearson's correlation matrix coefficient method analyzes the correlation between the significant dataset characteristics utilized to predict the WQI values.

$$\text{R} = \frac{\boldsymbol{n} * \sum(\boldsymbol{u} * \boldsymbol{v}) - (\sum \boldsymbol{u})(\sum \boldsymbol{v})}{[\boldsymbol{n}\sum(\boldsymbol{u}^2) - \sum(\boldsymbol{u}^2)] * [\boldsymbol{n}\sum(\boldsymbol{v}^2) - \sum(\boldsymbol{v}^2)]} \times \mathbf{100}$$

where

R: Pearson's correlation coefficient approach;
$\boldsymbol{u}$: Input values in the first set of the training data;
$\boldsymbol{v}$: Input values of the second set of the training data;
$\boldsymbol{n}$: Total number of input variables.

## 3. Results

Pearson's correlation coefficient (R) was used to evaluate the linear relationships between the Water Quality Index (WQI) and all input variables, including dissolved oxygen, pH, conductivity, biochemical oxygen demand (BOD5), nitrate, fecal coliform, and total coliform. This analysis helped identify the most influential variables for predicting the WQI values. The results revealed strong correlations between all variables and the WQI, confirming their critical importance in determining the water quality. Figure 7 illustrates the correlation of each variable with the WQI.

The dataset was split into two subsets: 80% for training and 20% for testing. The training dataset was used to develop the machine learning models, while the testing dataset was used to evaluate the models' predictive performance. This approach ensures that the models are trained effectively while maintaining a separate dataset to assess their generalization capabilities.

The purpose of using this 80/20 split is twofold. First, it ensures that the models are trained on a sufficiently large dataset, capturing the underlying patterns and relationships between variables. Second, by keeping 20% of the data exclusively for testing, the study provides an unbiased evaluation of the model's predictive accuracy on unseen data, minimizing the risk of overfitting.

The strong correlations between the WQI and the selected variables reinforce their significance for water quality classification. The 80/20 split strategy further strengthens the methodological framework, providing a reliable basis for evaluating the performance of machine learning classifiers in real-world applications.

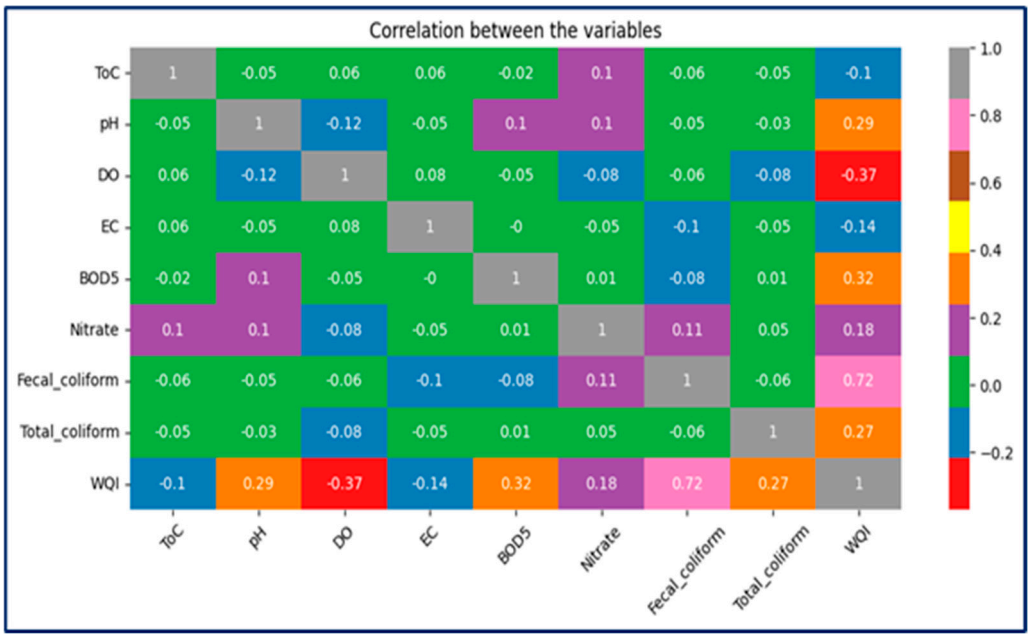



**Figure 7.** Correlation between all variables and WQI.

Figure 8a–c illustrates the 5 × 5 confusion matrices for the Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF) classifiers, respectively. Each confusion matrix is color-coded to visually represent the distribution of correctly and incorrectly classified instances for the respective models. These matrices were utilized to compute the key performance metrics in Table 3, including accuracy, precision, recall, and F1-score.

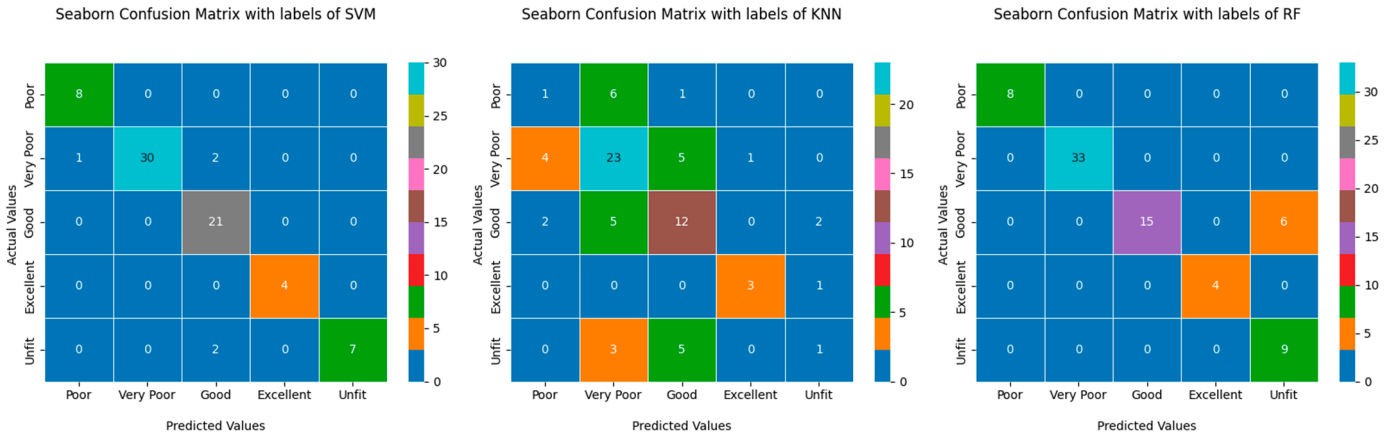



**Figure 8.** (**a**) Confusion matrix of SVM; (**b**) confusion matrix of KNN; and (**c**) confusion matrix of RF.

The Support Vector Machine (SVM) classifier utilized a Radial Basis Function (RBF) kernel, known for effectively handling non-linear data patterns, to achieve optimal performance. The regularization parameter (C) was set to 1000, balancing the trade-off between achieving a low error rate on the training data and maintaining a more straightforward decision boundary. In contrast, the kernel parameter ($\gamma$) was tuned to 0.1 to control the influence of individual data points. For the K-Nearest Neighbor (KNN) classifier, the number of nearest neighbors (k) was optimized and set to 4, with the Euclidean distance metric used to determine the similarity between data points. This configuration allowed the KNN model to classify instances effectively based on the features' proximity. The Random

Forest (RF) classifier consisted of an ensemble of 500 decision trees (n_estimators=), where the maximum number of features considered at each split was set to the square root of the total features, to ensure diversity among trees and reduce the likelihood of overfitting. Additional hyperparameters, such as the maximum tree depth and minimum samples per leaf, were tuned to enhance the RF model's performance further. These carefully selected parameters for each classifier underline the methodological rigor of this study, and contribute significantly to the reliability of the results.

The performance comparison shows that in Table 3, the SVM model outperformed the other classifiers, achieving a remarkably high accuracy of 99.5%. The RF model followed at 95%, while the KNN model lagged, with an accuracy of 58%. The superior performance of the SVM classifier can be attributed to its ability to handle complex, non-linear data patterns effectively using the RBF kernel.

Figure 9a–c presents the Receiver Operating Characteristic (ROC) curves for the classifiers, displaying both individual class curves and micro and macro-average labels. The Area Under the Curve (AUC) quantifies the overall performance of the classifiers, with a higher AUC indicating better discrimination capability. In this study, the SVM algorithm exhibited superior ROC–AUC values compared to the other classifiers, highlighting its effectiveness in distinguishing between classes.

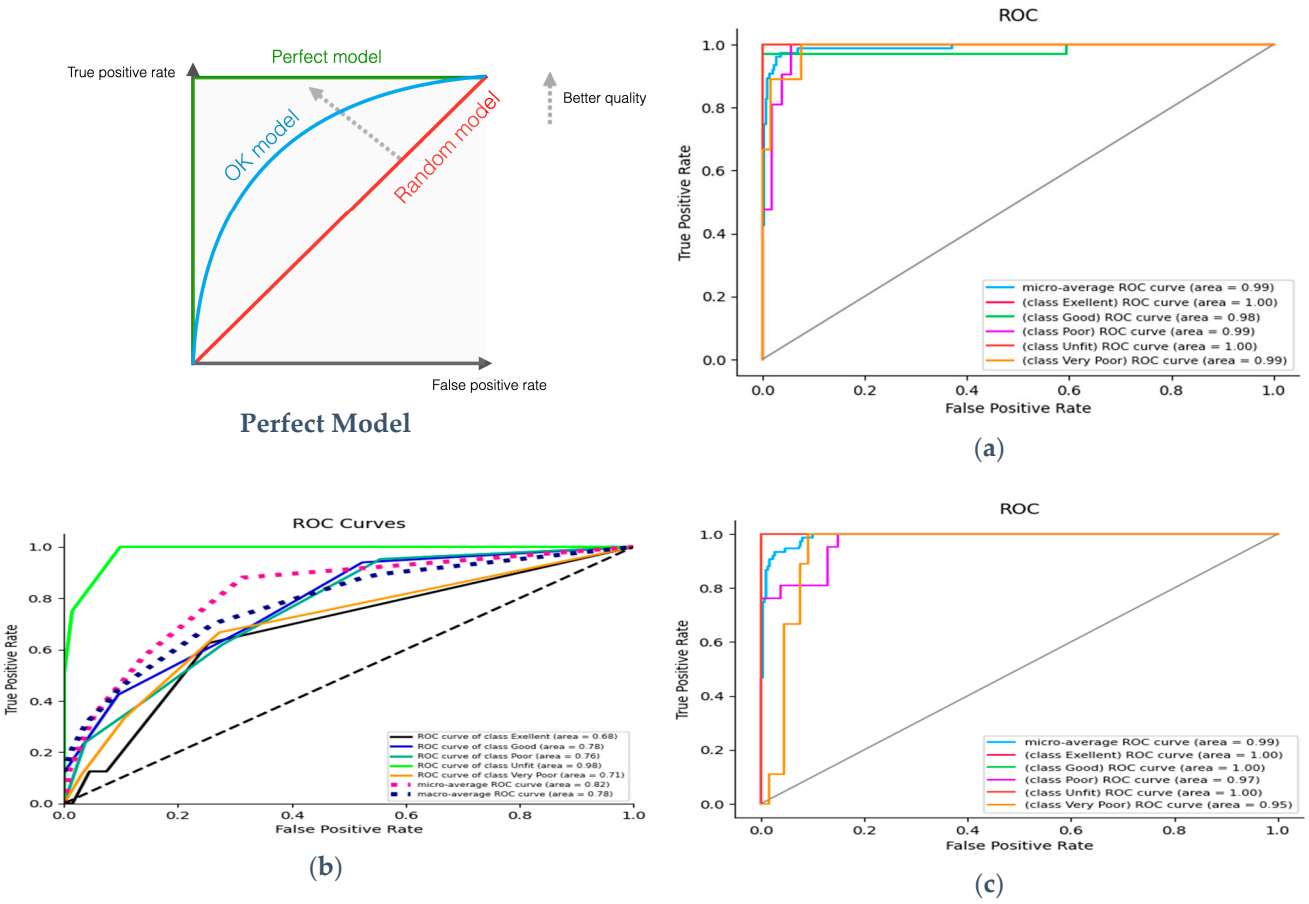
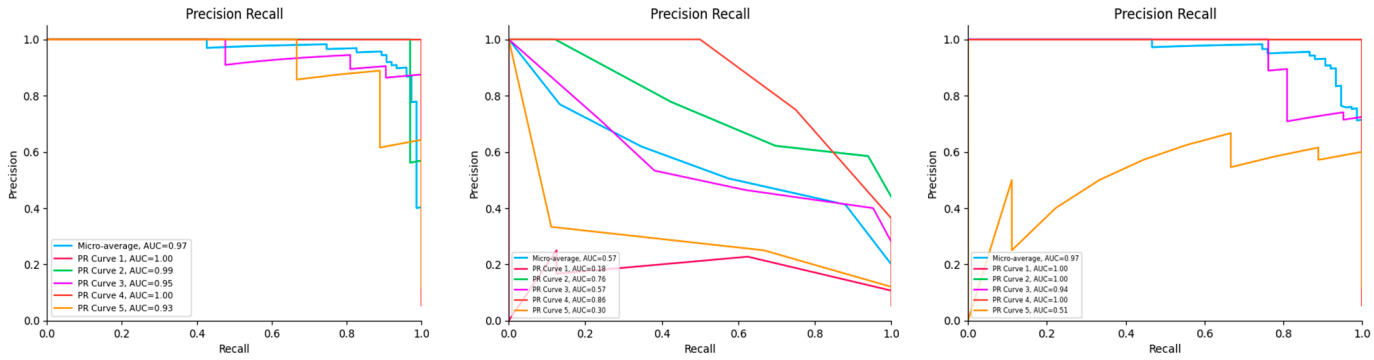



**Figure 9.** ROC curve for perfect model; (**a**) ROC curve for SVM; (**b**) ROC curve for KNN; and (**c**) ROC curve for RF.

Figure 10a–c shows the precision-recall (PR) curves for all classifiers, which provide an additional perspective on classifier performance, especially in scenarios with imbalanced class distributions. The results demonstrate that the SVM algorithm outperformed the other models, maintaining higher precision and recall values. These curves provide critical

insights into the models' predictive performance, reinforcing the conclusion that the SVM algorithm is the most reliable classifier for this dataset.




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

**Figure 10.** (**a**) Precision–recall for SVM; (**b**) precision-recall for KNN; and (**c**) precision-recall for RF.

This study found that the SVM classifier achieved the highest accuracy score of 99.5% on the actual dataset, outperforming both the KNN and RF classifiers, which attained 58% and 95% accuracy scores, respectively. These results are detailed in Table 3 and visualized in Figure 11.

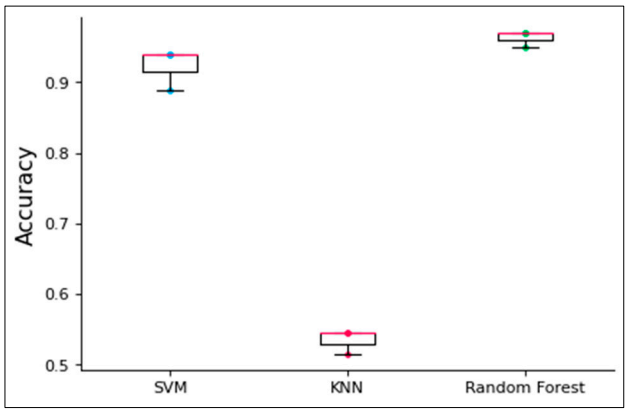



**Figure 11.** Accuracy of ML algorithms.

## 4. Discussion

This study presents an intelligent, real-time water quality monitoring framework using machine learning techniques, focusing on Keenjhar Lake, Pakistan. The dataset, spanning from 1993 to 2022, included 360 samples with key water quality parameters, such as dissolved oxygen, pH, conductivity, biochemical oxygen demand (BOD5), nitrate, fecal coliform, and total coliform. These features influence aquatic ecosystems, making them essential for water quality classification.

Three machine learning classifiers were used in this study, namely the Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF) algorithms, and their performance was assessed using metrics such as precision, recall, F1-score, and accuracy, derived from confusion matrices. Among these models, the SVM algorithm demonstrated superior performance, with an accuracy of 99.5%, significantly outperforming the RF (95%) and KNN (58%) algorithms. The SVM classifier also showed consistent reliability across the precision–recall, ROC, and average precision curves, confirming its robustness for water quality classification tasks.

The results emphasize that the SVM algorithm's strong performance stems from its ability to effectively handle complex, multidimensional datasets. Its kernel-based approach allows it to model non-linear relationships between water quality parameters and their classification outcomes [14]. This aligns well with the physical reality of water quality dynamics, where interactions between parameters such as the dissolved oxygen, BOD5, and pH are non-linear and interdependent. For example, higher levels of BOD5 often result in decreased dissolved oxygen, adversely affecting aquatic life, a pattern effectively captured by the SVM model [15].

From a practical standpoint, the findings have significant implications for water quality monitoring. The high accuracy of the SVM classifier supports its deployment in real-time water monitoring systems, enabling stakeholders to make timely decisions for environmental management and public health protection. The classification system can aid in identifying contamination events, optimizing water treatment strategies, and ensuring compliance with environmental regulations [16].

This study also highlights the potential for interdisciplinary collaboration in water quality research. By integrating machine learning expertise with domain knowledge in hydrology and environmental science, more robust and adaptive solutions can be developed for monitoring and managing water resources.

Future research directions include the hardware implementation of the proposed system for real-time applications, integration with Internet of Things (IoT) platforms, and expansion into other domains, such as biomedical studies and industrial wastewater management. Furthermore, these studies could explore ensemble models or hybrid approaches to enhance classification performance. Expanding the dataset to include seasonal variations and anthropogenic influences could also provide deeper insights into the temporal dynamics of water quality.

In conclusion, this study underscores the potential of machine learning techniques, particularly the SVM algorithm, for accurate water quality classification. The proposed framework provides a valuable sustainable water resource management and environmental conservation tool by bridging the gap between computational methods and physical processes.

**Author Contributions:** Conceptualization, M.I. and D.Z.; methodology, M.I.; software, N.E.J.M.; validation, D.Z., M.Z. and S.P.; data curation, S.P. and M.I.; writing—original draft preparation, M.I.; writing—review and editing, D.Z.; visualization, M.Z.; supervision, D.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data will be available on request.

**Acknowledgments:** We are grateful to the Irrigation Department of Sindh, Pakistan, for providing the dataset.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Damo, R.; Icka, P. Evaluation of water quality index for drinking water. *Pol. J. Environ. Stud.* **2013**, *22*, 1045–1051.
2. Kumar, P. Simulation of Gomti River (Lucknow City, India) future water quality under different mitigation strategies. *Heliyon* **2018**, *4*, e01074. [CrossRef] [PubMed]
3. Sun, A.Y.; Scanlon, B.R. How can Big Data and machine learning benefit environment and water management: A survey of methods, applications, and future directions. *Environ. Res. Lett.* **2019**, *14*, 073001. [CrossRef]

4. Bagheri, M.; Akbari, A.; Mirbagheri, S.A. Advanced control of membrane fouling in filtration systems using artificial intelligence and machine learning techniques: A critical review. *Process Saf. Environ. Prot.* **2019**, *123*, 229–252. [CrossRef]
5. Hassanpour, F.; Sharifazari, S.; Ahmadaali, K.; Mohammadi, S.; Sheikhalipour, Z. Development of the FCM-SVR Hybrid Model for Estimating the Suspended Sediment Load. *KSCE J. Civ. Eng.* **2019**, *23*, 2514–2523. [CrossRef]
6. Ehteram, M.; Ghotbi, S.; Kisi, O.; Najah Ahmed, A.; Hayder, G.; Ming Fai, C.; Krishnan, M.; Abdulmohsin Afan, H.; EL-Shafie, A. Investigation on the potential to integrate different artificial intelligence models with metaheuristic algorithms for improving river suspended sediment predictions. *Appl. Sci.* **2019**, *9*, 4149. [CrossRef]
7. Liao, H.; Sun, W. Forecasting and evaluating water quality of Chao Lake based on an improved decision tree method. *Procedia Environ. Sci.* **2010**, *2*, 970–979. [CrossRef]
8. Solanki, A.; Agrawal, H.; Khare, K. Predictive Analysis of Water Quality Parameters using Deep Learning. *Int. J. Comput. Appl.* **2015**, *125*, 29–34. [CrossRef]
9. Shafi, U.; Mumtaz, R.; Anwar, H.; Qamar, A.M.; Khurshid, H. Surface Water Pollution Detection using Internet of Things. In Proceedings of the 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT), Islamabad, Pakistan, 8–10 October 2018; pp. 92–96. [CrossRef]
10. Kovačević, M.; Amiri, B.J.; Lozančić, S.; Hadzima-Nyarko, M.; Radu, D.; Nyarko, E.K. Application of Machine Learning in Modeling the Relationship between Catchment Attributes and Instream Water Quality in Data-Scarce Regions. *Toxics* **2023**, *11*, 996. [CrossRef] [PubMed]
11. Lashari, K.H.; Habib Naqvi, S.; Palh, Z.A.; Laghari, Z.A.; Mastoi, A.A.; Sahato, G.A.; Mastoi, G.M. The Effects of Physiochemical Parameters on Planktonic Species Population of Keenjhar Lake, District Thatta, Sindh, Pakistan. *Am. J. Biosci.* **2014**, *2*, 38–44.
12. Kangabam, R.D.; Bhoominathan, S.D.; Kanagaraj, S.; Govindaraju, M. Development of a water quality index (WQI) for the Loktak Lake in India. *Appl. Water Sci.* **2017**, *7*, 2907–2918. [CrossRef]
13. Robert, G.K.; Onyari, C.N.; Mbaka, J.G. Development of a Water Quality Assessment Index for the Chania River, Kenya. *Afr. J. Aquat. Sci.* **2021**, *46*, 142–152. [CrossRef]
14. Leong, W.C.; Bahadori, A.; Zhang, J.; Ahmad, Z. Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM). *Int. J. River Basin Manag.* **2021**, *19*, 149–156. [CrossRef]
15. Ma, J.; Ding, Y.; Cheng, J.C.P.; Jiang, F.; Xu, Z. Soft detection of 5-day BOD with sparse matrix in city harbor water using deep learning techniques. *Water Res.* **2020**, *170*, 115350. [CrossRef] [PubMed]
16. Zhu, M.; Wang, J.; Yang, X.; Zhang, Y.; Zhang, L.; Ren, H.; Wu, B.; Ye, L. A review of the application of machine learning in water quality evaluation. *Eco-Environ. Health* **2022**, *1*, 107–116. [CrossRef] [PubMed]


**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.