water

MDPI

Article

Machine Learning Models for Water Quality Prediction: A Comprehensive Analysis and Uncertainty Assessment in Mirpurkhas, Sindh, Pakistan

Farkhanda Abbas [1,*], Zhihua Cai [1], Muhammad Shoaib [2], Javed Iqbal [3], Muhammad Ismail [4], Arifullah [5], Abdulwahed Fahad Alrefaei [6] and Mohammed Fahad Albeshr [6]

1 School of Computer Science, China University of Geosciences, Wuhan 430074, China; zhcai@cug.edu.cn
2 State Key Laboratory of Hydraulic Engineering, Simulation and Safety, School of Civil Engineering, Tianjin University, Tianjin 300072, China; xs4shoaib@tju.edu.cn
3 School of Environmental Studies, China University of Geosciences, Wuhan 430074, China; javediqbal@cug.edu.cn
4 Department of Computer Science, Karakoram International University, Gilgit 15100, Pakistan; muhammad.ismail@kiu.edu.pk
5 State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, China Institute of Water Resources and Hydropower Research (IWHR), Beijing 100038, China; parviarif@gmail.com
6 Department of Zoology, College of Science, King Saud University, P.O. Box 2455, Riyadh 11451, Saudi Arabia; afrefaei@ksu.edu.sa (A.F.A.); albeshr@ksu.edu.sa (M.F.A.)
* Correspondence: shamin0427@cug.edu.cn

Abstract: Groundwater represents a pivotal asset in conserving natural water reservoirs for potable consumption, irrigation, and diverse industrial uses. Nevertheless, human activities intertwined with industry and agriculture contribute significantly to groundwater contamination, highlighting the critical necessity of appraising water quality for safe drinking and effective irrigation. This research primarily focused on employing the Water Quality Index (WQI) to gauge water's appropriateness for these purposes. However, the generation of an accurate WQI can prove time-intensive owing to potential errors in sub-index calculations. In response to this challenge, an artificial intelligence (AI) forecasting model was devised, aiming to streamline the process while mitigating errors. The study collected 422 data samples from Mirpurkash, a city nestled in the province of Sindh, for a comprehensive exploration of the region's WQI attributes. Furthermore, the study probed into unraveling the interdependencies amidst variables in the physiochemical analysis of water. Diverse machine learning classifiers were employed for WQI prediction, with findings revealing that Random Forest and Gradient Boosting lead with 95% and 96% accuracy, followed closely by SVM at 92%. KNN exhibits an accuracy rate of 84%, and Decision Trees achieve 77%. Traditional water quality assessment methods are time-consuming and error-prone; a transformative approach using artificial intelligence and machine learning addresses these limitations. In addition to WQI prediction, the study conducted an uncertainty analysis of the models using the R-factor, providing insights into the reliability and consistency of predictions. This dual approach, combining accurate WQI prediction with uncertainty assessment, contributes to a more comprehensive understanding of water quality in Mirpurkash and enhances the reliability of decision-making processes related to groundwater utilization.

Keywords: groundwater modeling; Water Quality Index; machine learning algorithms; water quality assessment

Citation: Abbas, F.; Cai, Z.; Shoaib, M.; Iqbal, J.; Ismail, M.; Arifullah; Alrefaei, A.F.; Albeshr, M.F. Machine Learning Models for Water Quality Prediction: A Comprehensive Analysis and Uncertainty Assessment in Mirpurkhas, Sindh, Pakistan. *Water* **2024**, *16*, 941. https://doi.org/10.3390/w16070941

Academic Editors: Elias Dimitriou and Joaquim Sousa

Received: 12 February 2024
Revised: 19 March 2024
Accepted: 20 March 2024
Published: 25 March 2024

1. Introduction

Water, as an indispensable resource, plays a fundamental role in sustaining life and supporting various human activities. Among its many sources, groundwater stands as a crucial reservoir essential for drinking, agriculture, and industrial processes in Pakistan.

However, the escalating impact of human interventions, particularly in industrial and agricultural sectors, poses a substantial threat to the quality of this invaluable resource [1–4]. The condition of groundwater in Pakistan, including areas like Mirpurkhas in the province of Sindh, has faced mounting challenges due to extensive usage, urbanization, and agricultural runoff, leading to contamination concerns and a decline in overall quality. The region's reliance on groundwater for daily consumption and agricultural needs amplifies the urgency for effective water quality assessment measures [2,5–7]. Contamination of groundwater due to these anthropogenic activities has heightened concerns regarding its suitability for consumption and irrigation purposes, necessitating robust methods for accurate evaluation and monitoring [8–11].

Traditionally, assessing water quality, especially the determination of the Water Quality Index (WQI), relied heavily on manual calculations and established formulas based on a set of parameters [9,12–15]. These methods often entail time-consuming processes and are prone to human errors, particularly in complex calculations involving multiple interdependent factors [16–18]. In recent years, the integration of artificial intelligence and machine learning techniques, implemented using programming languages like Python (version 3.10), alongside specialized libraries such as scikit-learn (version 0.24), XGBoost (version 1.5), and pandas (version 1.3), has emerged as a transformative approach to overcome the limitations of traditional methods. Machine learning models, such as Random Forest, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Trees, were developed and trained using these tools. They offer the advantage of learning patterns and relationships from vast datasets, enabling more accurate and efficient prediction of WQI [19–21].

This study, conducted in the geographic area of Mirpurkhas in Sindh, collected an extensive dataset of 422 samples to comprehensively understand the region's water quality characteristics. Leveraging Python and various machine learning libraries such as scikit-learn, XGBoost, and pandas, the research employed these tools to preprocess data, build, train, and evaluate machine learning classifiers for predicting WQI [22,23]. The results indicate that Random Forest and Gradient Boosting outperformed other algorithms, achieving an exceptional accuracy rate of 99%. Following closely were SVM and XGBoost, scoring approximately 95% and 93% accuracy, respectively, while KNN and Decision Trees demonstrated accuracy rates of 88% and 87%, respectively. These findings underscore the efficacy of Python-based machine learning techniques implemented with specialized libraries in accurately predicting WQI, showcasing their potential for advancing water quality assessment methods, particularly in groundwater evaluation [24–26]. In a similar vein, Reza Mohammadpour's article [27] employed Support Vector Machine (SVM) and two artificial neural network (ANN) methods, feed forward back propagation (FFBP) and radial basis function (RBF), for Water Quality Index (WQI) prediction in a constructed wetland. The SVM model outperformed, achieving a high coefficient of correlation (R^2) of 0.9984 and a low mean absolute error (MAE) of 0.0052, demonstrating its effectiveness in streamlining WQI calculations and optimizing computational efforts in free surface constructed wetland environments. Afaq Juna's article [28] supports our experimental findings, where Random Forest (RF) and XGBoost both attain an 80% accuracy for Water Quality Index (WQI). RF demonstrates precision, recall, and an F1 score of 80%, while XGBoost achieves 80% precision and recall, with an F1 score slightly lower at 79%. In contrast, KNN and SGDC exhibit the lowest WQI accuracy at 59%. Mehedi Hassan's article [19] in WQI prediction demonstrates outstanding accuracy, with Kappa, Accuracy Lower, and Accuracy Upper scores reaching 99.83, 99.17, and 99.07, respectively. These results underscore the crucial role of machine learning in precisely categorizing water quality, highlighting its significance for effective water management and corroborating our high accuracy in machine learning models for Water Quality Index prediction.

Through an interdisciplinary approach integrating environmental science and machine learning, this research aims to contribute to the advancement of accurate and efficient water quality assessment methods, utilizing the potential of artificial intelligence and predictive

modeling implemented through Python-based tools and specialized libraries. Beyond WQI prediction, this study integrates an uncertainty analysis using the R-factor, providing a nuanced perspective on the reliability and consistency of our predictive models. The combined approach of accurate WQI prediction and uncertainty assessment contributes to a more holistic understanding of water quality dynamics in Mirpurkash. Ultimately, this research aims to inform robust decision-making processes regarding groundwater utilization, considering both the accuracy of predictions and the inherent uncertainties associated with them. The integration of artificial intelligence (AI) forecasting models, specifically machine learning classifiers, such as Random Forest, Gradient Boosting, SVM, XGBoost, KNN, and Decision Trees, has proven to be instrumental in predicting the Water Quality Index (WQI) with remarkable accuracy [19,29–32]. However, the accuracy of predictions alone does not provide a complete picture, and understanding the structure of these models is essential for a comprehensive assessment of uncertainty.

This research paper is structured to encompass several key sections. Beginning with an Introduction that highlights the significance of groundwater, particularly in the context of Mirpurkhas in Sindh, it emphasizes the challenges of water quality and the need for advanced assessment methods, summarizing previous studies on groundwater quality, traditional Water Quality Index (WQI) determination methods, their limitations, and existing research on applying machine learning in water quality assessment. The Section 2.2 outlines the steps undertaken, including data collection of 422 samples, data preprocessing, feature selection, and the utilization of machine learning algorithms such as Random Forest, Gradient Boosting, SVM, XGBoost, KNN, and Decision Trees and evaluation of uncertainty in the above machine learning algorithms. The subsequent Section 3 presents the performance metrics of these models in predicting WQI accuracy rates. Following this, the Discussion interprets the outcomes, compares model performances, addresses limitations, and suggests further research avenues. Finally, a Conclusion summarizes the key findings, reinforces the significance of employing machine learning in water quality assessment, and suggests future implications.

2. Materials and Methods

2.1. Study Area

Mirpurkhas, situated in the Sindh province of Pakistan, experiences an arid to semi-arid climate characterized by scorching summers with temperatures often exceeding 40 degrees Celsius (104 degrees Fahrenheit) from April to September Figure 1. Monsoons, occurring between July and September, bring moderate to heavy rainfall, providing relief from the intense heat. Winters are relatively mild, ranging from around 10 to 20 degrees Celsius (50 to 68 degrees Fahrenheit). Geographically, Mirpurkhas is located near the Indus River in the southern part of Pakistan and is renowned for its agricultural activities (Figure 1) [7,33]. Wells play a crucial role in providing groundwater for various purposes, including drinking water supply and agricultural irrigation, supporting the local livelihoods within this semi-arid region.

Mirpurkhas, a town located in the Sindh province of Pakistan, relies heavily on well water for various purposes. The inhabitants of Mirpurkhas primarily utilize well water for drinking, agricultural irrigation, and domestic needs [34,35]. Wells in the region serve as a primary source of groundwater, supplying water to the local community. Well water in Mirpurkhas is crucial for sustaining daily activities and agricultural practices. However, like many areas reliant on groundwater, the water quality in wells can be susceptible to contamination from various sources such as agricultural runoff, industrial activities, and natural factors [36–38].

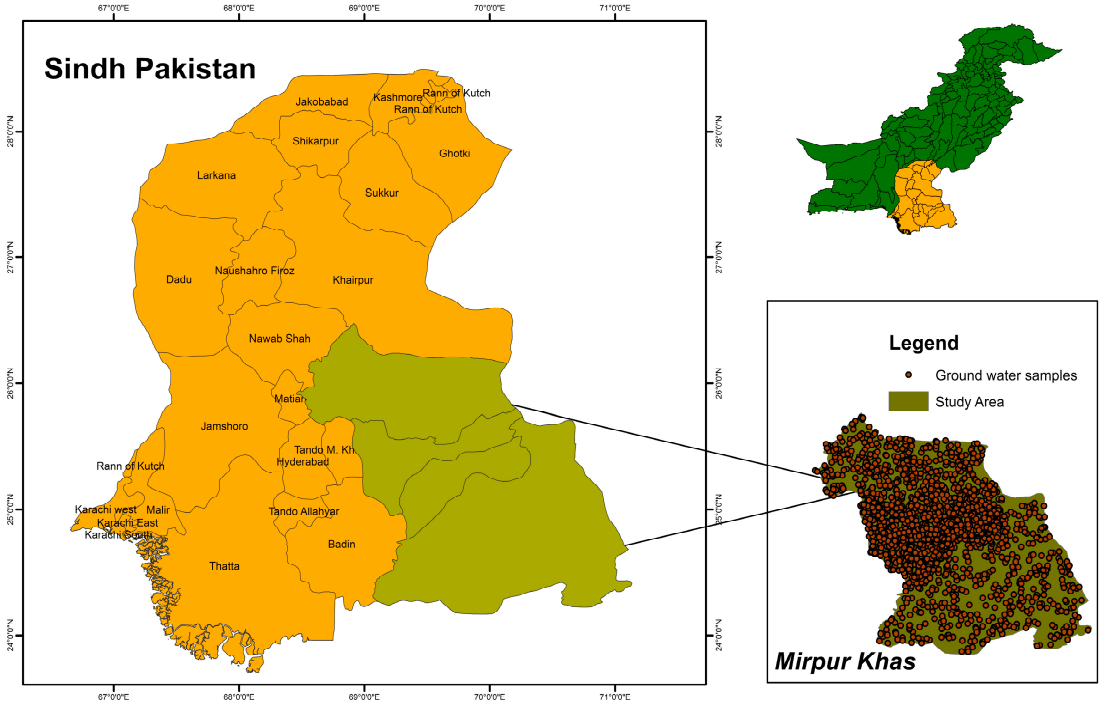

Figure 1. Study area and groundwater sampling points.

2.2. Methodology

The research methodology involved the collection of 422 water samples from multiple sites across Mirpurkhas, Sindh, Pakistan, during the period from April to May 2022, covering various locations deemed significant for groundwater extraction and consumption. Parameters, including pH levels, temperature, dissolved oxygen, turbidity, nitrates, and other physiochemical characteristics, were measured using standardized water testing procedures and equipment [39–41]. The dataset comprises a total of 422 samples, filtered to 0.45 µm for further analysis, with their locations recorded using a global positioning system (GPS). Standard methods outlined by the American Public Health Association [42] were followed for analysis. The well depths vary widely, ranging from 5.7 m to 590 m, indicating a diverse dataset that includes samples from both shallow and deep aquifers. The variation in well depths is essential to consider, as it may influence groundwater characteristics, impacted by geological and hydrological factors associated with different depth ranges. Following data collection, a rigorous preprocessing phase was conducted to ensure data accuracy and suitability for machine learning analysis Figure 2. This stage encompassed handling missing values through imputation methods, outlier removal, and normalization or scaling to ensure uniformity across parameters. Feature engineering was performed to extract pertinent features and reduce dimensionality for enhanced model performance. Feature selection techniques were employed, including Variance Inflation Factor (VIF) and Information Gain (IG), to identify influential parameters affecting water quality. These methods aimed to reduce redundancy and select the most informative features for modeling [43,44].

The evaluation of groundwater suitability for human consumption involved the computation of the Water Quality Index (WQI) based on the standards established by the World Health Organization (WHO). The WQI calculation comprised a three-step procedure. Initially, an individual weight (wi) was assigned to each parameter, encompassing TDS, Sodium, Calcium, Magnesium, Bicarbonate, Sulfate, Chloride, pH, EC, Nitrate (NO^{3-}), Well Depth, and Potassium. Subsequently, the relative weight (W_i) for each parameter was determined. Lastly, quality-rating scales (q_i) and sub-indices (SI_i) were computed for each parameter, and the overall WQI was derived by summing the sub-indices. The resulting classification into five groups, ranging from Group 1 (0–25), indicating Excellent water quality,

to Group 5 (above 100), signifying Very Poor to Unacceptable water quality, was employed in collaboration with machine learning models for a more comprehensive analysis.

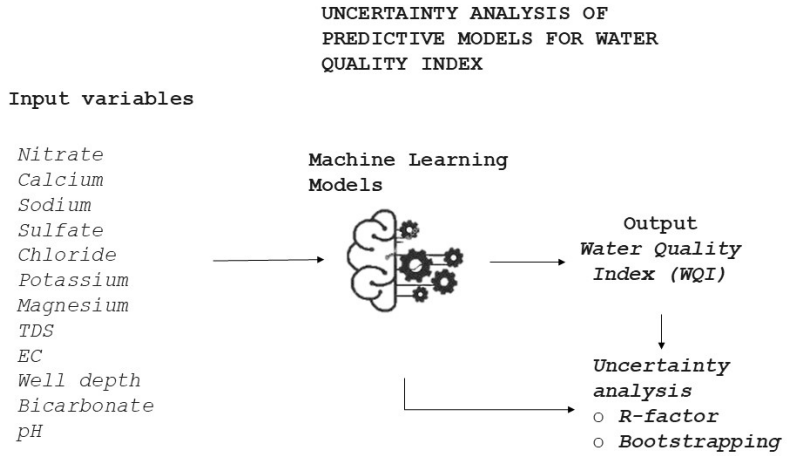

Figure 2. Methodology used for predicting Water Quality Index for given input variables.

For model development, thePython programming language was utilized, along with machine learning libraries such as scikit-learn, XGBoost, pandas, and numpy. Supervised learning algorithms, including Random Forest, Gradient Boosting, Support Vector Machine (SVM), XGBoost, K-Nearest Neighbors (KNN), and Decision Trees, were implemented and trained using the preprocessed dataset [45,46]. Hyperparameter tuning through techniques like grid search and cross-validation optimized the models. The performance of the developed models was evaluated using common metrics such as accuracy, confusion matrix, Friedman test, and Nemenyi test [47,48]. The uncertainty of model predictions has been evaluated using R-factor and bootstrapping.

The data underwent resampling using a cross-validation technique to assess model robustness and generalizability. The training and testing ratios, as evidenced by our confusion matrices across all classifiers, fall within the range of 20% to 33%. XGB, Random Forest, and SVC are generally regarded as robust and less susceptible to overfitting, enabling them to perform effectively with a smaller testing set (20%). In contrast, KNN, Gradient Boosting, and Decision Tree models may exhibit greater sensitivity to the nuances of the training data, suggesting potential benefits from a larger testing set (30%) for a more thorough evaluation [49]. Results interpretation involved comparing and analyzing the outcomes of various machine learning classifiers to identify the most accurate models for predicting the Water Quality Index (WQI). Models demonstrating the highest accuracy rates were further analyzed to understand the impact of different parameters on WQI prediction and water quality assessment.

The variables that have been used in our research to determine the Water Quality Index are shown in Figure 3.

The VIF analysis Table 1 highlights varying degrees of multicollinearity among the features considered for water quality assessment in Mirpurkhas, Sindh, Pakistan. Notably, certain parameters, such as 'TDS', 'Sodium', 'Calcium', and 'Magnesium', exhibited notably high VIF values, indicative of strong multicollinearity among these variables. Conversely, 'Potassium', 'Well Depth', and 'Nitrate (NO^{3-})' demonstrated relatively lower VIF values, suggesting lower levels of multicollinearity in comparison [50].

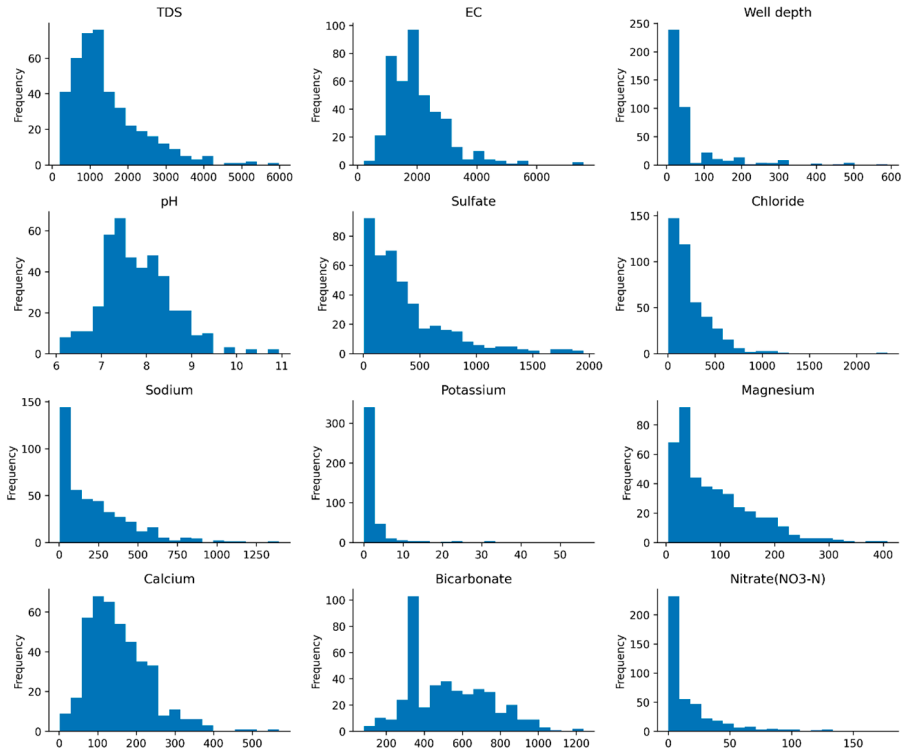

Figure 3. Input parameters used for Water Quality Index prediction and assessment.

Table 1. Variance Inflation Factor (VIF) values indicating multicollinearity among water quality assessment features in Mirpurkhas, Sindh, Pakistan.

Feature	VIF
TDS (mg/L)	4209.78
Sodium (mg/L)	1137.34
Calcium (mg/L)	425.13
Magnesium (mg/L)	380.55
Bicarbonate (mg/L)	58.74
Sulfate (mg/L)	39.68
Chloride (mg/L)	31.69
pH	20.16
EC (us/cm)	10.20
Nitrate (NO_3-N) (mg/L)	5.45
Well Depth (m)	1.70
Potassium (mg/L)	1.43

The Variance Inflation Factor (VIF) values, obtained from the assessment of water quality parameters in Mirpurkhas, Sindh, reveal varying degrees of multicollinearity among features considered for predicting the Water Quality Index (WQI). Features such as 'TDS', 'Sodium', 'Calcium', and 'Magnesium' exhibit notably high VIF values, suggesting strong interdependencies among these variables. This significant multicollinearity potentially impacts the accuracy of predictive models developed for water quality assessment [51]. Parameters with lower VIF values, including 'Potassium', 'Well Depth', and 'Nitrate (NO^{3-})', indicate weaker correlations, potentially posing less influence on multicollinearity issues within predictive models. Addressing high multicollinearity, particularly among variables with elevated VIF values, becomes crucial in enhancing the reliability and precision of predictive models for more accurate water quality assessment in the Mirpurkhas region.

Tree-based models (Decision Trees, Random Forest, Gradient Boosting, XGBoost) and K-Nearest Neighbors (KNN) are generally less sensitive to multicollinearity compared to linear models like linear regression or logistic regression. Support Vector Machine (SVM)

can be sensitive to multicollinearity to some extent, depending on the kernel used; therefore, we use a linear kernel with SVM [52]. The linear kernel computes the dot product between two observations. It is less sensitive to multicollinearity because it effectively works in the original feature space without introducing non-linear transformations.

Although elevated VIF values may signal multicollinearity and potential challenges in linear models, opting to include all variables based on Information Gain remains a viable strategy, particularly when employing tree-based models such as KNN, RF, Gradient Boosting, XGBoost, and Decision Trees. In addition, to make SVM less sensitive to multicollinearity, we use a linear kernel in our research. Nevertheless, it is crucial to empirically validate this decision by evaluating the model's performance on independent datasets or employing robust cross-validation techniques.

The Information Gain (IG) analysis Table 2 highlights the relevance of various features in predicting the Water Quality Index (WQI) in Mirpurkhas, Sindh, Pakistan. Features such as 'Nitrate (NO_3-N)', 'Calcium', 'Sodium', 'Sulfate', 'Chloride', 'Potassium', and 'Magnesium' exhibit higher IG values, indicating their considerable relevance in predicting WQI. Conversely, 'pH', 'Bicarbonate', 'Well Depth', 'EC', and 'TDS' present relatively lower IG values, suggesting comparatively lesser impact in predicting the WQI. Understanding the relevance of these features assists in selecting the most influential variables for the development of accurate predictive models for water quality assessment.

Table 2. Information Gain (IG) values indicating corresponding information gain for each water quality assessment feature in Mirpurkhas, Sindh, Pakistan.

Feature	IG
Nitrate (NO_3-N) (mg/L)	0.876
Calcium (mg/L)	0.869
Sodium (mg/L)	0.869
Sulfate (mg/L)	0.869
Chloride (mg/L)	0.869
Potassium (mg/L)	0.869
Magnesium (mg/L)	0.869
TDS (mg/L)	0.816
EC (us/cm)	0.784
Well Depth (m)	0.525
Bicarbonate (mg/L)	0.520
pH	0.509

However, it is important to note that while IG values help identify influential features, the absolute value of IG alone might not necessarily determine the direct impact or importance of a feature in predicting the WQI [53]. Other factors, such as domain knowledge, the nature of the dataset, and the specific context of the water quality assessment, should also be considered when selecting influential variables for building accurate predictive models. Therefore, while IG values provide valuable insights, the selection of the most influential variables should involve a comprehensive analysis that integrates multiple factors beyond IG values alone.

2.3. Uncertainty Analysis

2.3.1. R-Factor

While various factors contribute to the uncertainty in predicting Water Quality Index (WQI), including modeling, sampling errors, data preparation, and pre-processing, this study specifically addresses the uncertainty linked to individual model structures and input parameter selection. To assess model structure uncertainty, the analysis involves examining a set of three predicted WQI values during the testing phase for each observed WQI. These predictions are generated by the aforementioned predictive models.

The mean and standard deviation are computed for each predicted set, serving as parameters for a designated normal distribution function. Employing the 'Monte Carlo'

simulation method, 1000 WQI values are generated for each observed value based on this distribution. While other methods like Latin Hypercube [54], Lagged Average [55], and Multimodal Nesting [56] are utilized for sample generation, the Monte Carlo technique has demonstrated greater applicability, especially in hydrology and water-related sciences [57]. To quantify the uncertainty associated with WQI prediction, the 95% prediction confidence interval (i.e., the interval between the 97.5% and 2.5% quantiles), known as the prediction uncertainty of 95% (95PPI), is determined using the generated WQI values for each observed WQI. Specifically, the uncertainty is computed using the defined R-factor (Equation (1)).

The formula for the calculation of the R-factor is expressed as:

$$\text{R-factor} = \frac{s_p}{s_x} \tag{1}$$

Here, s_x represents the standard deviation of the observed values, and s_p is determined using Equation (2):

$$s_p = \frac{\sum_{i=1}^{J}(U_{Li} - L_{Li})}{J} \tag{2}$$

In this equation, J denotes the number of observed data points, while U_{Li} and L_{Li} correspond to the i-th values of the upper quartile (97.5%) and lower quartile (2.5%) of the 95% prediction confidence interval band (95PPI).

Other approaches, such as the Coefficient of Variation (CV), Prediction Interval Coverage Probability (PICP), and Prediction Interval Normalized Root-mean-square Width (PINRW), have been proposed as substitutes for the R-factor method [58]. Nevertheless, these alternative methods solely rely on either observed or predicted data. In contrast, the R-factor method takes into account both observed and predicted data, making it a more comprehensive metric for characterizing prediction uncertainty [59,60]. The inherent uncertainty in predictive models arises from various sources, including the complexity of the underlying data and the dynamic nature of water quality parameters. The structure of machine learning models contributes significantly to this uncertainty, and exploring their characteristics sheds light on the reliability of predictions.

2.3.2. Bootstrapping

In the uncertainty analysis of predictive models for Water Quality Index, generating prediction intervals is crucial for understanding the range of possible values for each prediction. This step involves using bootstrapping, a resampling technique that provides a measure of the uncertainty associated with the model's predictions. Bootstrapping involves creating multiple bootstrap samples by randomly drawing observations with replacements from the original dataset. For each bootstrap sample, the model is trained, and predictions are made on the test set. This process is repeated numerous times (in our case, 1000 iterations), resulting in a distribution of predicted values for each data point Figure 4.

The Mean Squared Error on the test set (0.108) indicates the average squared difference between the actual Water Quality Index values and the predicted values. A lower MSE generally suggests better model performance, demonstrating that the model's predictions are, on average, close to the true values. However, the MSE alone may not provide a complete picture, as it does not account for the uncertainty in the predictions. This is where prediction intervals come into play. The generated prediction intervals using bootstrapping offer insights into the variability and uncertainty associated with the model's predictions. The lower and upper bounds of the intervals (calculated at the 2.5th and 97.5th percentiles, respectively) represent the plausible range within which the true Water Quality Index values are likely to fall. The scatter plot (Figure 4) of actual versus predicted values, along with the shaded gray area representing the prediction intervals, provides a clear visualization of the model's performance and the associated uncertainty. The narrower the prediction intervals, the more confident we can be in the model's predictions.

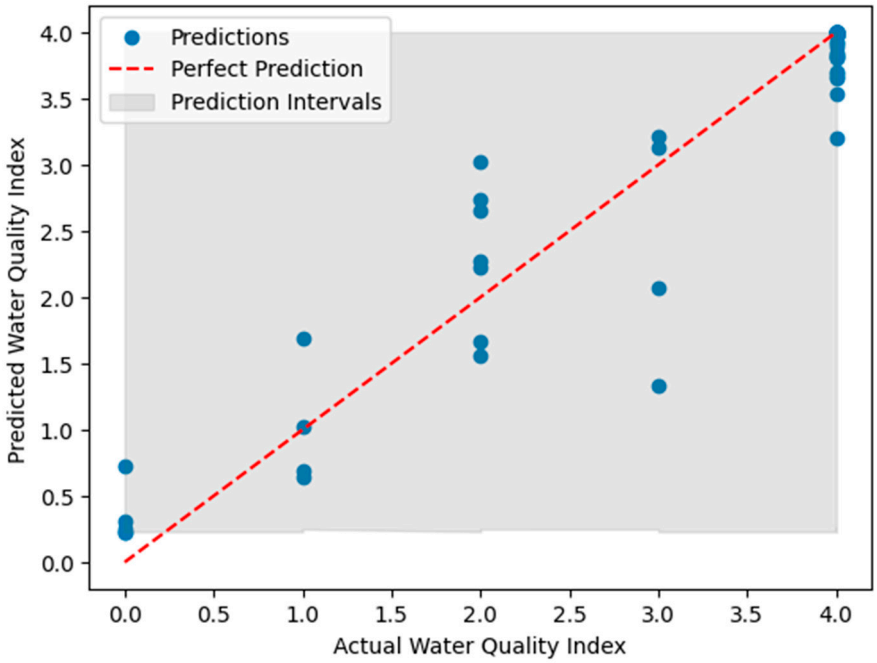

Figure 4. Predicted values, along with the prediction intervals, using bootstrapping.

A narrow prediction interval suggests that the model has a high degree of certainty in its predictions. A wider prediction interval indicates higher uncertainty, emphasizing the need for caution when relying on specific predictions in these regions. By incorporating bootstrapping to generate prediction intervals, we not only assess the model's accuracy through MSE but also gain a comprehensive understanding of the uncertainty inherent in the Water Quality Index predictions. This holistic approach enhances the reliability and robustness of the predictive modeling process, making it more applicable and informative for water quality management and decision-making.

2.3.3. Random Forest and Gradient Boosting

These ensemble methods aggregate predictions from multiple decision trees, which individually capture different patterns in the data. The robustness of Random Forest and Gradient Boosting lies in their ability to mitigate overfitting and enhance predictive accuracy. However, the ensemble nature introduces uncertainty due to the variability in individual tree predictions [61,62].

2.3.4. Support Vector Machine (SVM) and XGBoost

SVM focuses on finding the hyperplane that best separates data into classes, while XGBoost optimizes the performance of weak learners through boosting. The structural complexity of SVM and the iterative refinement process of XGBoost contribute to their predictive power but also introduce uncertainty, particularly in capturing non-linear relationships and intricate patterns [63].

2.3.5. K-Nearest Neighbors (KNN) and Decision Trees

KNN relies on proximity-based classification, and Decision Trees partition the data based on feature splits. These models are interpretable and less complex, but their simplicity can lead to uncertainty when faced with intricate relationships in the data. KNN's reliance on neighbors introduces variability, while Decision Trees' sensitivity to data changes may affect stability [64].

Understanding the interplay between model structure and uncertainty is crucial for reliable water quality assessments. The ensemble nature of Random Forest and Gradient Boosting, along with the iterative optimization in SVM and XGBoost, contributes to their

robust performance but introduces variability. Simpler models like KNN and Decision Trees may be more interpretable but can exhibit uncertainty in capturing complex relationships. The uncertainty associated with each model's structure emphasizes the importance of a nuanced approach to water quality prediction. Integrating uncertainty analysis, such as the R-factor, alongside accurate predictions allows for a more informed and cautious interpretation of water quality assessments, fostering a holistic understanding for effective decision-making, as shown in Table 3.

Table 3. R-factor obtained for all the machine learning algorithms in WQI prediction.

Classifier	R-Factor
K-Nearest Neighbors	0.83
Decision Trees	0.77
Gradient Boosting	0.83
Random Forest	0.83
SVM	0.83
XGBoost	0.83

3. Results

3.1. AUC-Based Performance Evaluation

The AUC values, as presented in Table 4 and Figure 5, offer valuable insights into the performance of various machine learning models in predicting the Water Quality Index. Decision Trees (DTs) exhibit reasonable discriminatory power with an AUC of 0.77, while the Random Forest (RF) and XGBoost models outperform, showcasing high AUC values of 0.95 and 0.96, respectively. These results underscore their robust performance in accurately categorizing water quality. The Gradient Boosting model also demonstrates excellent discriminatory power, with an AUC of 0.95. The Support Vector Machine (SVM) performs admirably with an AUC of 0.92, indicating effective classification. K-Nearest Neighbors (KNN) exhibits good discriminatory power, though slightly lower compared to some other models, with an AUC of 0.84. These varying AUC values emphasize the importance of selecting models with superior discriminatory capabilities when predicting the Water Quality Index, contributing to informed decision-making in environmental management.

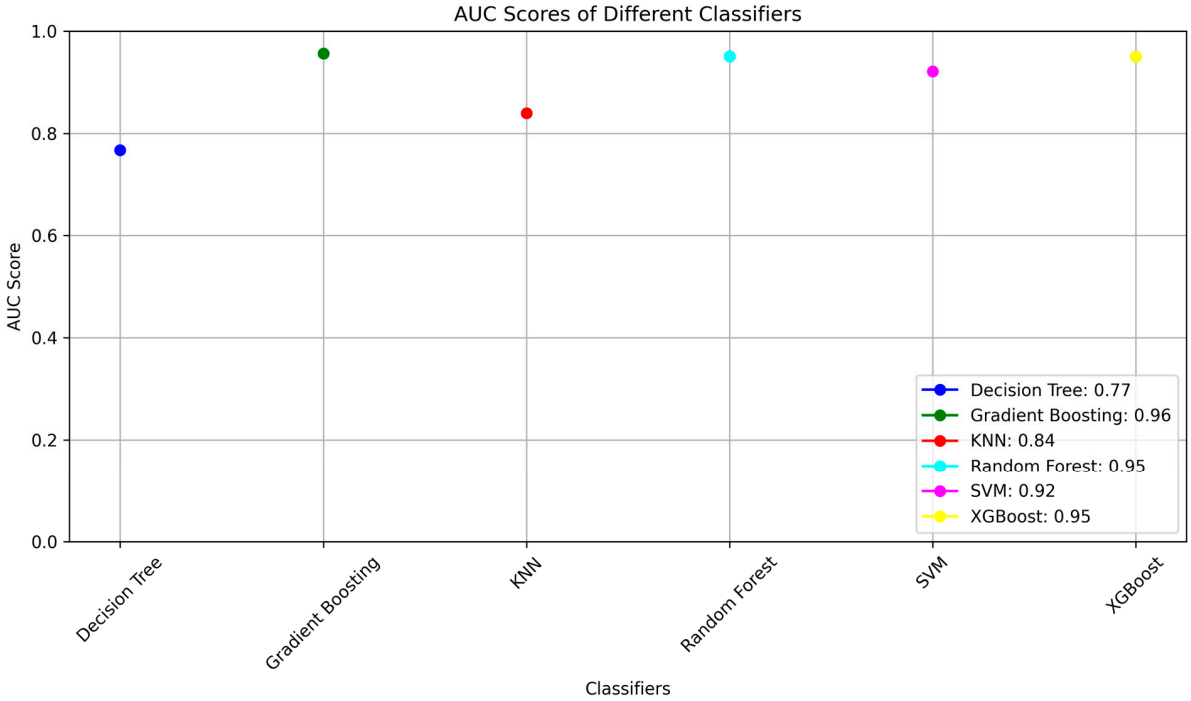

Figure 5. The AUC score for different classifiers used for Water Quality Index prediction and assessment.

Table 4. Performance Evaluation of Machine Learning Algorithms in WQI Prediction.

Algorithm	AUC
Decision Trees (DTs)	0.77
Random Forest (RF)	0.95
Gradient Boosting	0.96
K-Nearest Neighbors (KNN)	0.84
Support Vector Machine (SVM)	0.92
XGBoost	0.95

The ROC curve for Class 5 closely resembles that of Class 1, positioned near the ideal top-left corner Figure 6. The AUC score of 0.99 highlights exceptional performance in distinguishing Class 5 from other classes. The ROC curves provide a visual representation of the trade-off between sensitivity and specificity for each class, showcasing that all the classifiers performed well to make accurate predictions regarding WQI predictions in our experiment.

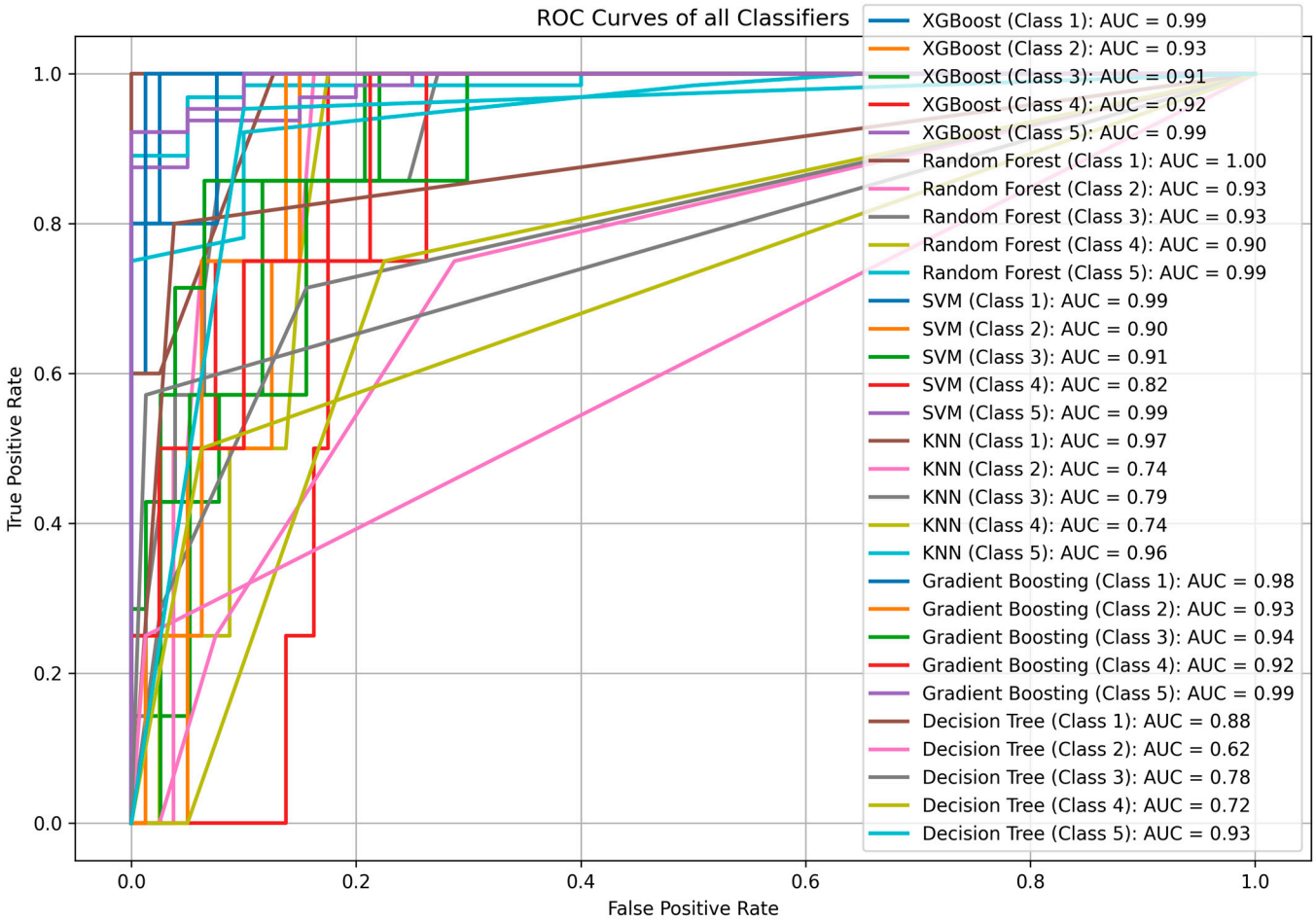

Figure 6. The ROC Curves for different classifiers used for Water Quality Index prediction and assessment.

3.2. Statistical Analysis Using Friedman Test

The Friedman Test was employed to assess the overall performance variation among multiple machine learning algorithms utilized for predicting the Water Quality Index (WQI) in the Mirpurkhas region of Sindh, Pakistan Table 5. The computed Friedman Test statistic yielded an F-value of 5.0 with a corresponding p-value of 0.4159 [65]. This analysis examines whether there exists a statistically significant difference in the performance of the various

machine learning models employed for water quality assessment. The obtained *p*-value of 0.4159, exceeding the conventional significance level of 0.05, indicates insufficient evidence to reject the null hypothesis. Therefore, based on this statistical test, there appears to be no significant difference observed in the predictive performance of the machine learning algorithms utilized for WQI prediction in the Mirpurkhas region [66].

Table 5. The results of the Friedman Test, indicating an F-value of 5.0 and a corresponding *p*-value of 0.4159.

Test	Value
Friedman Test—F-value	5.0
Friedman Test—*p*-value	0.4159

3.3. Nemenyi Test for Pairwise Comparisons

Each value in the matrix represents the critical distance between pairs of algorithms. In this matrix [67], rows and columns correspond to the XGB Classifier, Random Forest Classifier, Support Vector Classifier, K-Nearest Neighbors Classifier, Gradient Boosting Classifier, and Decision Tree Classifier, respectively (labeled 1 to 6) Table 6. The value of 1.0 along the diagonal signifies the comparison of an algorithm with itself, showing a critical distance of zero (as expected). The NaN (Not a Number) values outside the diagonal indicate that there is no significant difference between those pairs of algorithms based on the Nemenyi Test at a specific significance level.

Table 6. The critical distance values obtained from the Nemenyi Test for pairwise comparisons among six machine learning classifiers (XGB Classifier, Random Forest Classifier, Support Vector Classifier, K-Nearest Neighbors Classifier, Gradient Boosting Classifier, Decision Tree Classifier).

	XGB Classifier	Random Forest Classifier	Support Vector Classifier	K-Nearest Neighbors Classifier	Gradient Boosting Classifier	Decision Tree Classifier
1	1.0	NaN	NaN	NaN	NaN	NaN
2	NaN	1.0	NaN	NaN	NaN	NaN
3	NaN	NaN	1.0	NaN	NaN	NaN
4	NaN	NaN	NaN	1.0	NaN	NaN
5	NaN	NaN	NaN	NaN	1.0	NaN
6	NaN	NaN	NaN	NaN	NaN	1.0

The critical distance values are used in post hoc tests like the Nemenyi Test to compare the average ranks of different algorithms and determine which pairs of algorithms exhibit statistically significant differences in performance [68]. If the difference in the average ranks of two algorithms exceeds the critical distance value, this suggests a statistically significant difference in their performance.

3.4. Confusion Matrix

A confusion matrix serves as a tabular tool employed in machine learning and classification endeavors to assess the efficacy of a classification algorithm. It condenses the model's predictions on a dataset, contrasting them with the actual labels [69]. This matrix proves instrumental in discerning the nature of errors made by a model, including instances of false positives and false negatives. Tables 7–12 in our study showcase the confusion matrices for all machine learning algorithms utilized, offering a comprehensive overview of their performance.

These tables exhibit the counts of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) for each class predicted by each respective classifier in a five-class classification problem for predicting the Water Quality Index (WQI) in the specified region.

Table 7. The table represents a confusion matrix detailing the classification results for a multi-class classification for XGB Classifier.

True\Predicted	Class 1	Class 2	Class 3	Class 4	Class 5
Class 1	2 (TP)	3 (FN)	0	0	0
Class 2	0	3 (TP)	0	1 (FP)	0
Class 3	0	3 (FP)	1 (TP)	2 (FP)	1 (TN)
Class 4	0	2 (FP)	0	1 (TP)	1 (TN)
Class 5	0	0	0	2 (FP)	62 (TP)

Table 8. The table represents a confusion matrix detailing the classification results for a multi-class classification for Random Forest Classifier.

True\Predicted	Class 1	Class 2	Class 3	Class 4	Class 5
Class 1	4 (TP)	1 (FN)	0	0	0
Class 2	2 (FP)	1 (TP)	0	0	1 (FN)
Class 3	0	1 (FP)	1 (TP)	2 (FP)	3 (TN)
Class 4	0	1 (FP)	0	0	3 (TN)
Class 5	0	0	0	0	64 (TP)

Table 9. The table represents a confusion matrix detailing the classification results for a multi-class classification for SVC (Support Vector Classifier with probability = True).

True\Predicted	Class 1	Class 2	Class 3	Class 4	Class 5
Class 1	4 (TP)	1 (FN)	0	0	0
Class 2	2 (FP)	1 (TP)	0	0	1 (FN)
Class 3	1 (FP)	0	0	0	6 (FN)
Class 4	0	0	0	0	4 (FN)
Class 5	0	0	0	0	64 (TP)

Table 10. The table represents a confusion matrix detailing the classification results for a multi-class classification for KNN classifier.

True\Predicted	Class 1	Class 2	Class 3	Class 4	Class 5
Class 1	3 (TP)	2 (FN)	0	0	0
Class 2	3 (TP)	0	0	0	1 (FN)
Class 3	0	2 (FP)	1 (TP)	1 (FP)	3 (TN)
Class 4	0	0	0	0	4 (TN)
Class 5	0	1 (FP)	1 (TP)	2 (FP)	60 (TP)

Table 11. The table represents a confusion matrix detailing the classification results for a multi-class classification for Gradient Boosting Classifier.

True\Predicted	Class 1	Class 2	Class 3	Class 4	Class 5
Class 1	4 (TP)	1 (FN)	0	0	0
Class 2	2 (FP)	1 (TP)	0	1 (FP)	0
Class 3	0	2 (FP)	2 (TP)	2 (FP)	1 (TN)
Class 4	0	2 (FP)	0	1 (TP)	1 (TN)
Class 5	0	0	0	0	64 (TP)

The Water Quality Index (WQI) ranges in Table 13 were calculated based on a general classification scheme. These ranges are commonly used in water quality assessments, and the specific values may vary depending on the guidelines or standards adopted by environmental agencies.

Table 12. The table represents a confusion matrix detailing the classification results for a multi-class classification for Decision Tree Classifier.

True\Predicted	Class 1	Class 2	Class 3	Class 4	Class 5
Class 1	4 (TP)	1 (FN)	0	0	0
Class 2	2 (FP)	2 (TP)	0	0	0
Class 3	0	1 (FP)	3 (TP)	2 (FP)	1 (TN)
Class 4	0	1 (FP)	1 (FP)	1 (TP)	1 (TN)
Class 5	0	1 (FP)	0 (FP)	2 (FP)	61 (TP)

Table 13. Water quality ranges and their corresponding classes.

Classes	WQI Range	Water Quality
Class 1	0–25	Excellent water quality
Class 2	26–50	Good water quality
Class 3	51–75	Fair water quality
Class 4	76–100	Poor water quality
Class 5	Above 100	Very poor to unacceptable water quality

Total count of Class 5 instances among all classifiers:

$$62 + 64 + 64 + 60 + 64 + 61 = 375 \tag{3}$$

Calculating the percentage of WQI values that belong to Class 5:

$$\text{Percentage of Class 5 WQI} = \frac{\text{Total count of instances}}{\text{Count of Class 5 instances}} \times 100 \tag{4}$$

$$\text{Percentage of Class 5 WQI} = \frac{375}{422} \times 100 \approx 88.63\% \tag{5}$$

Approximately 88.63% of the Water Quality Index (WQI) values fall into Class 5, which represents very poor to unacceptable water quality. Table 14 below shows the testing sample results for the RF classifier.

Table 14. Testing sample results for RF classifier.

TDS	EC	Well Depth	pH	Sulfate	Chloride	Sodium	Potassium	Magnesium	Calcium	Bicarbonate	Nitrate (NO_3-N)	Rescaled_ WQI	Model Predicted_ WQI_Code
281	2892	24	8.5	17.12571	30.90157	17.18087	0.380653	6.112794	57.36129	82.53007	23.6361	134	5
285	2455	33	7.99	10.50744	25.27189	15.79175	0.880966	10.69318	73.59643	233.2372	5.283148	104	5
293	1266	47.1	7.39	15.33482	14.97501	35.97985	1.158209	14.11058	54.82109	275.1002	3.197757	6	1
302	2111	25	7.56	30.84742	22.94148	12.19949	1.193796	12.96383	78.09345	358.8264	5.686334	80	4
310	2000	19	8.37	67.98364	36.64451	20.84443	2.2859	15.15061	69.56399	358.8264	0.239171	76	4
370	1712	33	7.37	50.69821	23.40693	9.955348	0.890684	9.830066	103.6547	358.8264	10.54824	43	2
405	1122	300	8.21	768.8918	27.44628	15.49567	1.111356	14.22923	98.52735	358.8264	2.862759	66	3
406	1242	60	6.91	54.48075	152.4355	14.49937	1.267128	16.98422	106.119	358.8264	27.40252	26	2
406	1242	40	6.91	55.73943	30.83817	8.220457	1.118785	21.9048	105.0737	358.8264	7.057716	17	1
409	1829	101	10.94	59.5349	12.51032	5.80349	0.708562	10.05297	125.124	358.8264	8.928545	67	3
479	1972	265	9.18	54.17818	62.93912	23.93549	0.597271	21.97534	119.45	358.8264	3.133473	98	4

4. Discussion

This study focused on predicting the Water Quality Index (WQI) in the Mirpurkhas region, Sindh, Pakistan, utilizing various machine learning algorithms and exploring the significance of model structures and variable importance. The extensive analysis encompassed data collection, preprocessing, model development, performance evaluation, statistical tests, and uncertainty analysis, providing a comprehensive understanding of the water quality assessment process.

The AUC-based performance evaluation shed light on the efficacy of machine learning models in discriminating between different water quality classes. Notably, Random Forest

and XGBoost demonstrated high AUC values of 0.99 and 0.95, respectively, indicating robust discriminatory power. Gradient Boosting and SVM also exhibited excellent performance, with AUC values of 0.95 and 0.93. Decision Trees, while showing reasonable discriminatory power (AUC of 0.87), stood out as a viable model. The findings of this study shed light on the effectiveness of machine learning models in predicting the Water Quality Index (WQI) for the Mirpurkhas region in Sindh, Pakistan. Notably, XGBoost and Gradient Boosting demonstrated remarkable accuracy rates of 95%, outperforming other models. Random Forest closely followed suit, showcasing its effectiveness in WQI prediction. These outcomes align with the growing body of literature emphasizing the potential of machine learning in water quality assessment. The high accuracy of XGBoost, Gradient Boosting, and Random Forest models suggests their robust performance in capturing the intricate relationships among water quality parameters. Such findings resonate with studies conducted in various regions, where ensemble methods and tree-based models have shown superiority in water quality prediction [70]. The ability of these models to handle non-linear relationships and complex patterns in water quality data enhances their utility in environmental monitoring. Support Vector Machine (SVM) also exhibited commendable performance, with an AUC of 0.93, indicating effective classification. This aligns with studies that have highlighted the versatility of SVM in handling diverse datasets and its efficacy in water quality modeling [71,72]. However, it is essential to acknowledge the variations in model sensitivity to multicollinearity, as indicated by the Variance Inflation Factor (VIF) analysis. Features such as 'TDS', 'Sodium', 'Calcium', and 'Magnesium' exhibited high VIF values, suggesting strong interdependencies among these variables. While tree-based models are generally less sensitive to multicollinearity, addressing high VIF values remains crucial for enhancing the reliability of predictive models, especially in linear models like SVM. Information Gain analysis highlighted the relevance of specific physiochemical variables, such as 'Nitrate (NO_3-N)', 'Calcium', and 'Sodium', in WQI prediction. The study recommended future research to address limitations, including dataset size and variable scope. Advanced strategies like feature engineering, ensemble methods, and integration of remote sensing data were proposed to enhance predictive accuracy and provide a nuanced understanding of water quality dynamics. The Information Gain (IG) analysis provided insights into the relevance of different features in predicting WQI. Variables such as 'Nitrate (NO_3-N)', 'Calcium', 'Sodium', 'Sulfate', 'Chloride', 'Potassium', and 'Magnesium' exhibited higher IG values, underscoring their considerable impact on water quality assessment. This finding aligns with existing literature emphasizing the importance of specific physiochemical variables in influencing water quality [73,74].

The Friedman Test, employed to assess overall performance variation, yielded an F-value of 5.0 with a p-value of 0.4159. The non-significant p-value suggests consistent performance across machine learning algorithms, emphasizing their similarity in predictive accuracy for WQI in the Mirpurkhas region. The lack of significant differences supports the reliability and consistency of the models. The Nemenyi Test for pairwise comparisons provided critical distance values, offering insights into statistically significant differences in algorithm performance. The absence of significant differences between certain pairs of algorithms highlighted their comparable performance. While the critical distance values can guide algorithm ranking, the overall consistency observed in the Friedman Test aligns with the notion that various algorithms perform similarly in WQI prediction. Confusion matrices detailed the classification results for each machine learning algorithm, presenting true positives, false positives, false negatives, and true negatives for each water quality class. These matrices provide a granular view of model errors and successes, aiding in the interpretation of classification performance. The consistently high true-positive rates and low false-positive rates across classifiers reflect the models' abilities to accurately predict water quality classes. Defined water quality ranges and corresponding classes facilitated the interpretation of model predictions. Approximately 88.63% of WQI values fell into Class 5, representing very poor to unacceptable water quality. This distribution underscores the predominance of deteriorated water quality in the Mirpurkhas region, emphasizing the

urgency of effective water resource management. The uncertainty analysis, incorporating the R-factor and bootstrapping, added a crucial layer of insight into the reliability of model predictions. The R-factor addressed structural uncertainty, while bootstrapping provided prediction intervals, aiding in understanding the range of possible values for each prediction. The approach acknowledged and quantified uncertainty, contributing to a more informed interpretation of water quality assessments. Uncertainty analysis, including the R-factor and bootstrapping, contributed to a nuanced understanding of predictive model reliability. The Monte Carlo simulation method provided a robust approach to assessing the uncertainty associated with WQI predictions. The incorporation of bootstrapping not only assessed model accuracy through Mean Squared Error (MSE) but also provided valuable insights into prediction intervals, offering a more comprehensive understanding of uncertainty in the models.

5. Conclusions

In conclusion, this study demonstrates the efficacy of machine learning models, particularly XGBoost, Gradient Boosting, Random Forest, and SVM, in predicting the Water Quality Index for the Mirpurkhas region. The high accuracy rates of these models underscore their potential for precise water quality assessment. Feature importance analysis highlights the critical role of specific variables, emphasizing the need for targeted monitoring and management. This study's findings contribute to the broader discourse on machine learning applications in environmental science. The identified variables and models can serve as valuable tools for water resource management, aiding in informed decision-making. Despite the promising results, it is crucial to acknowledge the study's limitations, including dataset size and variable scope. Future research should explore advanced strategies and incorporate additional parameters for a more comprehensive understanding of water quality dynamics in the region. Overall, this study not only showcases the capabilities of machine learning in water quality prediction but also underscores the importance of considering uncertainties for robust environmental assessments.

Author Contributions: Conceptualization, F.A. and M.S.; Methodology, F.A. and M.S.; Software, F.A. and M.S.; Validation, Z.C., M.I., A. and A.F.A.; Formal analysis, A.F.A., M.F.A. and J.I.; Investigation, M.I.; Resources, J.I., M.F.A. and A.F.A.; Data curation, F.A., A. and M.S.; Writing—original draft, F.A.; Writing—review & editing, F.A., A. and Z.C.; Supervision, Z.C., A.F.A. and M.F.A.; Project administration, M.I., Z.C., M.F.A. and A.F.A. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by Researchers Supporting Project number (RSP2024R436), King Saud University, Riyadh, Saudi Arabia.

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

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

References

1. Rao, E.P.; Puttanna, K.; Sooryanarayana, K.; Biswas, A.; Arunkumar, J. Assessment of nitrate threat to water quality in India. In *The Indian Nitrogen Assessment*; Elsevier: Amsterdam, The Netherlands, 2017; pp. 323–333.
2. Wanke, H.; Nakwafila, A.; Hamutoko, J.; Lohe, C.; Neumbo, F.; Petrus, I.; David, A.; Beukes, H.; Masule, N.; Quinger, M. Hand dug wells in Namibia: An underestimated water source or a threat to human health? *Phys. Chem. Earth Parts A/B/C* **2014**, *76*, 104–113. [CrossRef]
3. Brown, T.C.; Froemke, P. Nationwide assessment of nonpoint source threats to water quality. *BioScience* **2012**, *62*, 136–146. [CrossRef]
4. Lapworth, D.; Boving, T.; Kreamer, D.; Kebede, S.; Smedley, P. Groundwater quality: Global threats, opportunities and realising the potential of groundwater. *Sci. Total Environ.* **2022**, *811*, 152471. [CrossRef] [PubMed]
5. Memon, A.H.; Lund, G.M.; Channa, N.A.; Younis, M.; Ali, S.; Shah, F.B. Analytical Study of Drinking Water Quality Sources of Dighri Sub-division of Sindh, Pakistan. *J. Environ. Agric. Sci.* **2016**, *8*, 38–44.
6. Khan, S.; Aziz, T.; Noor-Ul-Ain, A.K.; Ahmed, I.; Nida, A. Drinking water quality in 13 different districts of Sindh, Pakistan. *Health Care Curr. Rev.* **2018**, *6*, 1000235.

7. Akhan, F.; Siddqui, I.; USMANI, T. of Larkana and Mirpurkhas Districts of Sind. *J. Chem. Soc. Pak. Vol.* **2006**, *28*, 131.
8. Hayder, G.; Kurniawan, I.; Mustafa, H.M. Implementation of machine learning methods for monitoring and predicting water quality parameters. *Biointerface Res. Appl. Chem.* **2020**, *11*, 9285–9295.
9. Avila, R.; Horn, B.; Moriarty, E.; Hodson, R.; Moltchanova, E. Evaluating statistical model performance in water quality prediction. *J. Environ. Manag.* **2018**, *206*, 910–919. [CrossRef]
10. Ashwini, K.; Vedha, J.; Priya, M. Intelligent model for predicting water quality. *Int. J. Adv. Res. Ideas Innov. Technol. ISSN* **2019**, *5*, 70–75.
11. Kalin, L.; Isik, S.; Schoonover, J.E.; Lockaby, B.G. Predicting water quality in unmonitored watersheds using artificial neural networks. *J. Environ. Qual.* **2010**, *39*, 1429–1440. [CrossRef] [PubMed]
12. McGrane, S.J. Impacts of urbanisation on hydrological and water quality dynamics, and urban water management: A review. *Hydrol. Sci. J.* **2016**, *61*, 2295–2311. [CrossRef]
13. Dutt, V.; Sharma, N. Potable water quality assessment of traditionally used springs in a hilly town of Bhaderwah, Jammu and Kashmir, India. *Environ. Monit. Assess.* **2022**, *194*, 30. [CrossRef]
14. Lermontov, A.; Yokoyama, L.; Lermontov, M.; Machado, M.A.S. River quality analysis using fuzzy water quality index: Ribeira do Iguape river watershed, Brazil. *Ecol. Indic.* **2009**, *9*, 1188–1197. [CrossRef]
15. De Pauw, N.; Vanhooren, G. Method for biological quality assessment of watercourses in Belgium. *Hydrobiologia* **1983**, *100*, 153–168. [CrossRef]
16. Zhang, Y.; Guo, F.; Meng, W.; Wang, X.-Q. Water quality assessment and source identification of Daliao river basin using multivariate statistical methods. *Environ. Monit. Assess.* **2009**, *152*, 105–121. [CrossRef]
17. Lenat, D.R. Water quality assessment of streams using a qualitative collection method for benthic macroinvertebrates. *J. N. Am. Benthol. Soc.* **1988**, *7*, 222–233. [CrossRef]
18. Behmel, S.; Damour, M.; Ludwig, R.; Rodriguez, M. Water quality monitoring strategies—A review and future perspectives. *Sci. Total Environ.* **2016**, *571*, 1312–1329. [CrossRef]
19. Hassan, M.M.; Hassan, M.M.; Akter, L.; Rahman, M.M.; Zaman, S.; Hasib, K.M.; Jahan, N.; Smrity, R.N.; Farhana, J.; Raihan, M. Efficient prediction of water quality index (WQI) using machine learning algorithms. *Hum.-Centric Intell. Syst.* **2021**, *1*, 86–97. [CrossRef]
20. Lap, B.Q.; Du Nguyen, H.; Hang, P.T.; Phi, N.Q.; Hoang, V.T.; Linh, P.G.; Hang, B.T.T. Predicting water quality index (WQI) by feature selection and machine learning: A case study of An Kim Hai irrigation system. *Ecol. Inform.* **2023**, *74*, 101991. [CrossRef]
21. Ding, F.; Zhang, W.; Cao, S.; Hao, S.; Chen, L.; Xie, X.; Li, W.; Jiang, M. Optimization of water quality index models using machine learning approaches. *Water Res.* **2023**, *243*, 120337. [CrossRef] [PubMed]
22. Van Rossum, G. Python Programming Language. In Proceedings of the USENIX Annual Technical Conference, Santa Clara, CA, USA, 17–22 June 2007; pp. 1–36.
23. Saabith, A.S.; Vinothraj, T.; Fareez, M. Popular python libraries and their application domains. *Int. J. Adv. Eng. Res. Dev.* **2020**, *7*, 18–26.
24. Bansal, S.; Ganesan, G. Advanced evaluation methodology for water quality assessment using artificial neural network approach. *Water Resour. Manag.* **2019**, *33*, 3127–3141. [CrossRef]
25. Gevrey, M.; Rimet, F.; Park, Y.S.; Giraudel, J.L.; Ector, L.; Lek, S. Water quality assessment using diatom assemblages and advanced modelling techniques. *Freshw. Biol.* **2004**, *49*, 208–220. [CrossRef]
26. Uddin, M.G.; Olbert, A.I.; Nash, S. *Assessment of Water Quality Using Water Quality Index (WQI) Models and Advanced Geostatistical Technique*; Civil Engineering Research Association of Ireland (CERAI): Dublin, Ireland, 2020; pp. 594–599. Available online: https://aran.library.nuigalway.ie/bitstream/handle/10379/16427/CERI2020_Uddin_EBK_final.pdf?sequence=1 (accessed on 1 March 2024).
27. Mohammadpour, R.; Shaharuddin, S.; Chang, C.K.; Zakaria, N.A.; Ghani, A.A.; Chan, N.W. Prediction of water quality index in constructed wetlands using support vector machine. *Environ. Sci. Pollut. Res.* **2015**, *22*, 6208–6219. [CrossRef]
28. Juna, A.; Umer, M.; Sadiq, S.; Karamti, H.; Eshmawi, A.A.; Mohamed, A.; Ashraf, I. Water quality prediction using KNN imputer and multilayer perceptron. *Water* **2022**, *14*, 2592. [CrossRef]
29. Nasir, N.; Kansal, A.; Alshaltone, O.; Barneih, F.; Sameer, M.; Shanableh, A.; Al-Shamma'a, A. Water quality classification using machine learning algorithms. *J. Water Process Eng.* **2022**, *48*, 102920. [CrossRef]
30. 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. [CrossRef]
31. Khoi, D.N.; Quan, N.T.; Linh, D.Q.; Nhi, P.T.T.; Thuy, N.T.D. Using machine learning models for predicting the water quality index in the La Buong River, Vietnam. *Water* **2022**, *14*, 1552. [CrossRef]
32. Asadollah, S.B.H.S.; Sharafati, A.; Motta, D.; Yaseen, Z.M. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *J. Environ. Chem. Eng.* **2021**, *9*, 104599. [CrossRef]
33. Soomro, A.; Mangrio, M.; Bharchoond, Z.; Mari, F.; Pirzada, P.; Lashari, B.; Bhatti, M.; Skogerboe, G. *Maintenance Plans for Irrigation Facilities of Pilot Distributaries in Sindh Province, Pakistan. Volume 3—Bareji Distributary, Mirpurkhas District*; IWMI: Colombo, Sri Lanka, 1997.
34. Van der Hoek, W.; Boelee, E.; Konradsen, F. *Irrigation, Domestic Water Supply and Human Health*; Encyclopedia of Life Support Systems (EOLSS): Paris, France, 2002.

35. Van der Hoek, W.; Konradsen, F.; Ensink, J.H.; Mudasser, M.; Jensen, P.K. Irrigation water as a source of drinking water: Is safe use possible? *Trop. Med. Int. Health* **2001**, *6*, 46–54. [CrossRef]
36. Akhtar, N.; Syakir Ishak, M.I.; Bhawani, S.A.; Umar, K. Various natural and anthropogenic factors responsible for water quality degradation: A review. *Water* **2021**, *13*, 2660. [CrossRef]
37. Khatri, N.; Tyagi, S. Influences of natural and anthropogenic factors on surface and groundwater quality in rural and urban areas. *Front. Life Sci.* **2015**, *8*, 23–39. [CrossRef]
38. Burri, N.M.; Weatherl, R.; Moeck, C.; Schirmer, M. A review of threats to groundwater quality in the anthropocene. *Sci. Total Environ.* **2019**, *684*, 136–154. [CrossRef]
39. Udhayakumar, R.; Manivannan, P.; Raghu, K.; Vaideki, S. Assessment of physico-chemical characteristics of water in Tamilnadu. *Ecotoxicol. Environ. Saf.* **2016**, *134*, 474–477. [CrossRef]
40. Patil, P.; Sawant, D.; Deshmukh, R. Physico-chemical parameters for testing of water—A review. *Int. J. Environ. Sci.* **2012**, *3*, 1194–1207.
41. Brusseau, M.; Walker, D.; Fitzsimmons, K. Physical-chemical characteristics of water. In *Environmental and Pollution Science*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 23–45.
42. Beutler, M.; Wiltshire, K.; Meyer, B.; Moldaenke, C.; Luring, C.; Meyerhofer, M.; Hansen, U. APHA (2005), Standard Methods for the Examination of Water and Wastewater, Washington DC: American Public Health Association. Ahmad, SR, and DM Reynolds (1999), Monitoring of water quality using fluorescence technique: Prospect of on-line process control. *Dissolved Oxyg. Dyn. Model. Case Study A Subtrop. Shallow Lake* **2014**, *217*, 95.
43. Kroll, C.N.; Song, P. Impact of multicollinearity on small sample hydrologic regression models. *Water Resour. Res.* **2013**, *49*, 3756–3769. [CrossRef]
44. Sulaiman, M.S.; Abood, M.M.; Sinnakaudan, S.K.; Shukor, M.R.; You, G.Q.; Chung, X.Z. Assessing and solving multicollinearity in sediment transport prediction models using principal component analysis. *ISH J. Hydraul. Eng.* **2021**, *27*, 343–353. [CrossRef]
45. Iliou, T.; Anagnostopoulos, C.-N.; Nerantzaki, M.; Anastassopoulos, G. A novel machine learning data preprocessing method for enhancing classification algorithms performance. In Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS), Rhodes, Greece, 25–28 September 2015; pp. 1–5.
46. Werner de Vargas, V.; Schneider Aranda, J.A.; dos Santos Costa, R.; da Silva Pereira, P.R.; Victória Barbosa, J.L. Imbalanced data preprocessing techniques for machine learning: A systematic mapping study. *Knowl. Inf. Syst.* **2023**, *65*, 31–57. [CrossRef]
47. Veček, N.; Črepinšek, M.; Mernik, M. On the influence of the number of algorithms, problems, and independent runs in the comparison of evolutionary algorithms. *Appl. Soft Comput.* **2017**, *54*, 23–45. [CrossRef]
48. Liang, G.; Zhang, C. A comparative study of sampling methods and algorithms for imbalanced time series classification. In Proceedings of the AI 2012: Advances in Artificial Intelligence: 25th Australasian Joint Conference, Sydney, Australia, 4–7 December 2012; pp. 637–648.
49. Browne, M.W. Cross-validation methods. *J. Math. Psychol.* **2000**, *44*, 108–132. [CrossRef]
50. Daoud, J.I. Multicollinearity and regression analysis. *J. Phys. Conf. Ser.* **2017**, *949*, 012009. [CrossRef]
51. Akram, P.; Solangi, G.S.; Shehzad, F.R.; Ahmed, A. Groundwater Quality Assessment using a Water Quality Index (WQI) in Nine Major Cities of Sindh, Pakistan. *Int. J. Res. Environ. Sci. IJRES* **2020**, *6*, 18–26.
52. Abbas, F.; Zhang, F.; Ismail, M.; Khan, G.; Iqbal, J.; Alrefaei, A.F.; Albeshr, M.F. Optimizing machine learning algorithms for landslide susceptibility mapping along the Karakoram Highway, Gilgit Baltistan, Pakistan: A comparative study of baseline, bayesian, and metaheuristic hyperparameter optimization techniques. *Sensors* **2023**, *23*, 6843. [CrossRef]
53. Wijaya, D.R.; Sarno, R.; Zulaika, E. Information Quality Ratio as a novel metric for mother wavelet selection. *Chemom. Intell. Lab. Syst.* **2017**, *160*, 59–71. [CrossRef]
54. Singhee, A.; Rutenbar, R.A. Why quasi-Monte Carlo is better than Monte Carlo or Latin hypercube sampling for statistical circuit analysis. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2010**, *29*, 1763–1776. [CrossRef]
55. Hoffman, R.N.; Kalnay, E. Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus A Dyn. Meteorol. Oceanogr.* **1983**, *35*, 100–118. [CrossRef]
56. Feroz, F.; Hobson, M.P. Multimodal nested sampling: An efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses. *Mon. Not. R. Astron. Soc.* **2008**, *384*, 449–463. [CrossRef]
57. Noori, R.; Karbassi, A.; Moghaddamnia, A.; Han, D.; Zokaei-Ashtiani, M.; Farokhnia, A.; Gousheh, M.G. Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. *J. Hydrol.* **2011**, *401*, 177–189. [CrossRef]
58. Pan, M.; Li, C.; Liao, J.; Lei, H.; Pan, C.; Meng, X.; Huang, H. Design and modeling of PEM fuel cell based on different flow fields. *Energy* **2020**, *207*, 118331. [CrossRef]
59. Pirmohamed, M.; Burnside, G.; Eriksson, N.; Jorgensen, A.L.; Toh, C.H.; Nicholson, T.; Kesteven, P.; Christersson, C.; Wahlström, B.; Stafberg, C. A randomized trial of genotype-guided dosing of warfarin. *N. Engl. J. Med.* **2013**, *369*, 2294–2303. [CrossRef]
60. Sharafati, A.; Yasa, R.; Azamathulla, H.M. Assessment of stochastic approaches in prediction of wave-induced pipeline scour depth. *J. Pipeline Syst. Eng. Pract.* **2018**, *9*, 04018024. [CrossRef]
61. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobot.* **2013**, *7*, 21. [CrossRef]
62. Boulesteix, A.L.; Janitza, S.; Kruppa, J.; König, I.R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2012**, *2*, 493–507. [CrossRef]

63. Fan, J.; Wang, X.; Wu, L.; Zhou, H.; Zhang, F.; Yu, X.; Lu, X.; Xiang, Y. Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. *Energy Convers. Manag.* **2018**, *164*, 102–111. [CrossRef]
64. Jadhav, S.D.; Channe, H. Comparative study of K-NN, naive Bayes and decision tree classification techniques. *Int. J. Sci. Res. IJSR* **2016**, *5*, 1842–1845.
65. Sheldon, M.R.; Fillyaw, M.J.; Thompson, W.D. The use and interpretation of the Friedman test in the analysis of ordinal-scale data in repeated measures designs. *Physiother. Res. Int.* **1996**, *1*, 221–228. [CrossRef] [PubMed]
66. Pereira, D.G.; Afonso, A.; Medeiros, F.M. Overview of Friedman's test and post-hoc analysis. *Commun. Stat.-Simul. Comput.* **2015**, *44*, 2636–2653. [CrossRef]
67. Pohlert, T. The pairwise multiple comparison of mean ranks package (PMCMR). *R Package* **2014**, *27*, 9.
68. Garcia, S.; Herrera, F. An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. *J. Mach. Learn. Res.* **2008**, *9*, 2677–2694.
69. Townsend, J.T. Theoretical analysis of an alphabetic confusion matrix. *Percept. Psychophys.* **1971**, *9*, 40–50. [CrossRef]
70. Zeng, D.; Song, Y.; Wang, M.; Lu, Y.; Chen, Z.; Xiao, R. A machine learning approach for predicting the performance of oxygen carriers in chemical looping oxidative coupling of methane. *Sustain. Energy Fuels* **2023**, *7*, 3464–3470. [CrossRef]
71. Tran, H.D.; Li, H. Sound event recognition with probabilistic distance SVMs. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *19*, 1556–1568. [CrossRef]
72. Sun, J.; Yang, Y.; Wang, Y.; Wang, L.; Song, X.; Zhao, X. Survival risk prediction of esophageal cancer based on self-organizing maps clustering and support vector machine ensembles. *IEEE Access* **2020**, *8*, 131449–131460. [CrossRef]
73. Zhang, L.; Zhu, T.; Zhang, H.; Xiong, P.; Zhou, W. Fedrecovery: Differentially private machine unlearning for federated learning frameworks. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 4732–4746. [CrossRef]
74. Wang, W.; Liu, X. Intuitionistic fuzzy information aggregation using Einstein operations. *IEEE Trans. Fuzzy Syst.* **2012**, *20*, 923–938. [CrossRef]