sustainability

MDPI

Article

Unveiling Surface Water Quality and Key Influencing Factors in China Using a Machine Learning Approach

Yanli Li [1], Lei Liu [1], Lei Cheng [1] and Yahui Shan [2,*]

[1] State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, China; liyanli@hpu.edu.cn (Y.L.); wall2408zk@gmail.com (L.L.); lei.cheng@whu.edu.cn (L.C.)
[2] Wuhan Second Ship Design and Research Institute, Wuhan 430064, China
* Correspondence: shanyahui@hust.edu.cn

Abstract

Surface water quality assessment is critical for environmental protection and public health management, yet traditional methods are often time-consuming and costly, limiting their application for real-time monitoring. Machine learning (ML) approaches offer promising alternatives for automated water quality assessment and understanding of key influencing factors. This study employed six ML algorithms to predict water quality grades using comprehensive data from China's national surface water monitoring network. A dataset comprising 79,015 water quality measurements collected from 1 January to 14 February 2025 was processed with nine physicochemical parameters as input features. The XGBoost model demonstrated superior predictive performance with 99.04% accuracy. Feature importance analysis revealed that nutrient-related parameters (total phosphorus, permanganate index, ammonia nitrogen) consistently ranked as the most critical factors across all models. SHAP analysis provided interpretable explanations of model predictions, revealing grade-specific discrimination patterns where excellent quality waters are primarily distinguished by phosphorus limitation, while severely polluted waters require multi-parameter approaches. This study demonstrates the effectiveness of ML approaches for large-scale water quality assessment and provides a scientific foundation for optimizing monitoring strategies and environmental management decisions in China's surface water systems.

Keywords: machine learning; water quality assessment; SHAP analysis; surface water monitoring; environmental evaluation

MSC: 91B76; 68T99; 68-11

Academic Editor: Andreas N. Angelakis

Received: 1 September 2025
Revised: 3 October 2025
Accepted: 6 October 2025
Published: 17 October 2025

Citation: Li, Y.; Liu, L.; Cheng, L.; Shan, Y. Unveiling Surface Water Quality and Key Influencing Factors in China Using a Machine Learning Approach. *Sustainability* **2025**, *17*, 9205. https://doi.org/10.3390/su17209205

1. Introduction

Water quality assessment and monitoring are vital for environmental protection and public health governance, given the mounting pressures on freshwater resources from urbanization, industrial operations, and climate change [1,2]. Surface water bodies act as primary sources for drinking, irrigation, and ecosystem services, thus making their quality evaluation crucial for sustainable water resource management [3,4]. Although traditional water quality assessment methods are accurate, they are frequently time-consuming, labor-intensive, and expensive, which restricts their use in real-time monitoring and early warning systems [5,6].

The advent of artificial intelligence and machine learning technologies has transformed water quality prediction and assessment, providing promising alternatives to traditional

approaches [7,8]. Machine learning models have shown excellent ability to deal with complex, non-linear relationships among water quality parameters and can efficiently process large datasets [9,10]. Recent research has indicated that ensemble methods, especially tree-based algorithms like Random Forest, XGBoost, and gradient boosting models, consistently surpass traditional statistical methods in water quality prediction tasks [11–13].

A variety of machine learning algorithms have been successfully used for water quality prediction in different aquatic systems. Support vector machines and neural networks have been effective in predicting specific water quality parameters [14,15], while ensemble methods have shown superior performance in multi-parameter prediction situations [16,17]. Deep learning approaches, including long short-term memory networks and hybrid models, have been particularly valuable for time-series water quality prediction [18–20]. The integration of advanced optimization techniques, such as Bayesian optimization and particle swarm optimization, has further improved model performance by systematically adjusting hyperparameters [21,22]. Moreover, the inclusion of explainable artificial intelligence techniques, especially Shapley additive explanations (SHAP), has tackled the "black box" characteristic of machine learning models, offering interpretable insights into feature importance and model decision-making processes [13,23,24].

Recent progress in water quality prediction has highlighted the significance of comprehensive feature selection and data preprocessing techniques [25,26]. Studies have shown that proper data denoising, outlier detection, and feature engineering greatly enhance model accuracy and robustness [27,28]. In addition, the use of remote sensing data and satellite imagery has expanded the spatial coverage and temporal resolution of water quality monitoring systems [14,29]. Despite substantial progress in machine learning-based water quality prediction, several challenges persist. These include dealing with imbalanced datasets, managing high-dimensional data with limited samples, and ensuring model transferability across different geographical areas and water systems [28,30]. Additionally, the requirement for real-time prediction capabilities and integration with existing monitoring infrastructure poses ongoing technical challenges [31,32].

China's national surface water monitoring network provides an unparalleled opportunity to develop and validate machine learning models for large-scale water quality assessment. The extensive coverage of monitoring stations across various geographical and environmental conditions offers a robust dataset for training and testing advanced machine learning algorithms [2,9]. Applying machine learning techniques to China's water quality data has the potential to significantly improve environmental monitoring capabilities and support evidence-based water resource management decisions.

This study intends to develop and assess machine learning models for automated water quality assessment using comprehensive data from China's national surface water monitoring network. The specific objectives are as follows: (1) developing and comparing six machine learning algorithms for multi-grade water quality assessment; (2) implementing Bayesian optimization for hyperparameter tuning to enhance model performance; (3) conducting comprehensive feature importance analysis to identify key water quality parameters; (4) applying SHAP analysis to provide interpretable explanations of model predictions; and (5) evaluating model performance across different water quality grades and geographical regions. The results of this research will contribute to the advancement of intelligent water quality monitoring systems and provide valuable insights for environmental management and policy development.

2. Materials and Methods

2.1. Data Collection and Preprocessing

Water quality data for this study were obtained from the National Surface Water Quality Automatic Monitoring Real-time Data Release System, which provides standardized real-time measurements from automated stations across China's major water bodies [2,9]. The target variable for prediction is water quality grade, defined per the Environmental Quality Standards for Surface Water (GB 3838-2002) [33]. Nine key physicochemical parameters were included: temperature (°C), pH, dissolved oxygen (mg/L), conductivity (μS/cm), turbidity (NTU), permanganate index (mg/L), ammonia nitrogen (mg/L), total phosphorus (mg/L), and total nitrogen (mg/L). These parameters characterize comprehensive water quality through physical, chemical, and biological indicators.

The water quality classification system follows the Chinese national standard GB 3838-2002, which defines five quality classes (I to V) based on parameter threshold values. Table 1 presents the standard limit values for the key parameters used in this study. Class I represents excellent water quality suitable for source water and national nature reserves, while Class V indicates heavily polluted water with limited use potential.

Data spanning 1 January to 14 February 2025 initially contained 79,015 records.

The distribution of water quality classes in the dataset is illustrated in Figure 1, showing that Class II waters constitute the largest proportion (52.0%) of the dataset, followed by Class III (21.7%) and Class I (18.5%). Classes IV and V represent smaller proportions at 5.9% and 1.8%, respectively, indicating that the majority of monitored water bodies maintain acceptable quality levels, though the dataset includes sufficient samples across all quality categories for robust model training.

Table 1. Surface water quality standards for key parameters used in this study (GB 3838-2002).

Parameter	Unit	Class I	Class II	Class III	Class IV	Class V
Temperature	(°C)	Weekly average temperature change ≤ ±2 °C				
pH	(-)	6–9				
DO	(mg/L)	≥7.5	≥6	≥5	≥3	≥2
COD_{Mn}	(mg/L)	≤2	≤4	≤6	≤10	≤15
NH_3-N	(mg/L)	≤0.15	≤0.5	≤1.0	≤1.5	≤2.0
TP	(mg/L)	≤0.02	≤0.1	≤0.2	≤0.3	≤0.4
TN	(mg/L)	≤0.2	≤0.5	≤1.0	≤1.5	≤2.0

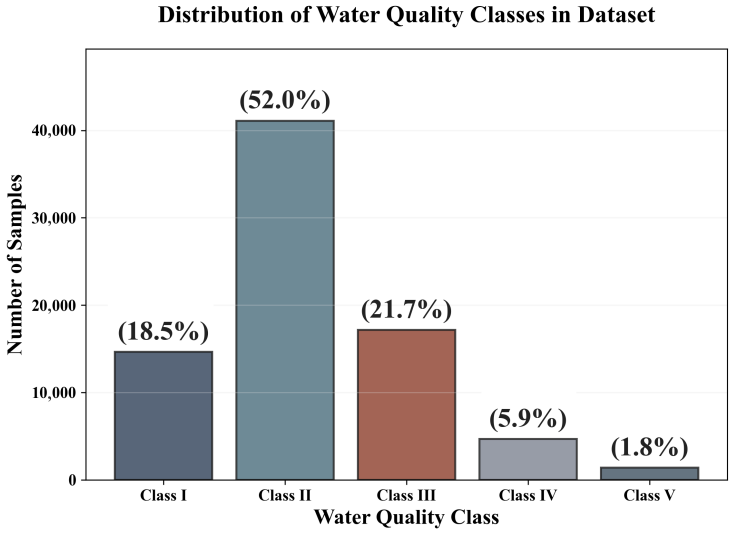

Figure 1. Distribution of water quality classes in the dataset. The bar chart shows the number of samples and percentage distribution across five quality categories (Class I to Class V) according to GB 3838-2002 standards. Class II waters dominate the dataset (52.0%), while severely polluted waters (Class V) represent the smallest proportion (1.8%).

Missing values were addressed via median imputation [25,26] to preserve data integrity. Preprocessing included type conversion, decimal standardization, and outlier treatment using a modified interquartile range (IQR) approach with 10th and 90th percentiles (P10 and P90). Values below P10 − 1.5 × (P90 − P10) were clipped to the lower boundary, while values above P90 + 1.5 × (P90 − P10) were clipped to the upper boundary, effectively replacing extreme values with boundary limits rather than removing them entirely.

Figure 2 outlines the study's methodology, encompassing three stages: (1) data acquisition (including the nine parameters noted) and preprocessing (type conversion, missing value handling, outlier treatment, target encoding, and feature standardization); (2) model development involving 70:30 train–test splitting, hyperparameter optimization via Bayesian optimization with 5-fold cross-validation, and model evaluation; and (3) model interpretability analysis using feature importance and SHAP methods to clarify parameter contributions. All analyses were implemented in Python 3.8.

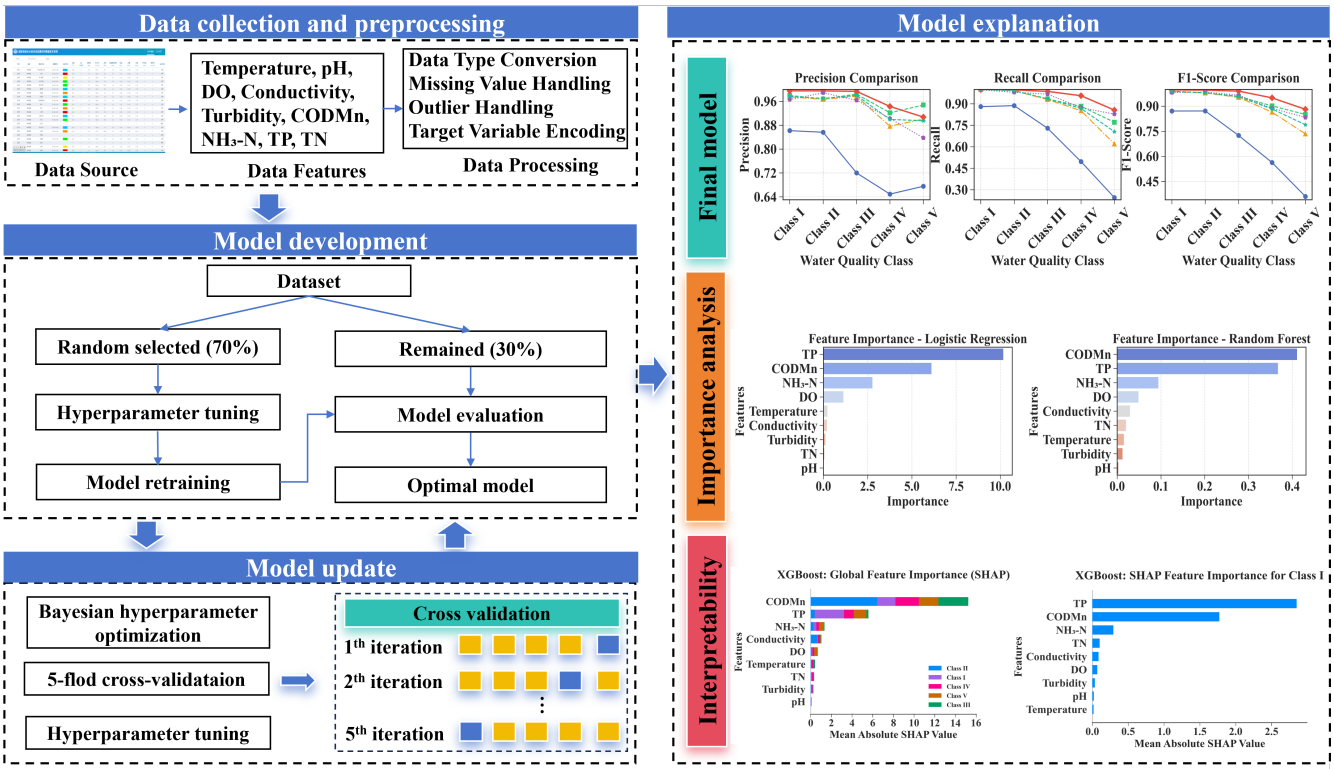

Figure 2. Schematic illustration of the workflow for this study. The workflow consists of three main components: (1) data collection and preprocessing, including data source acquisition, feature extraction, and comprehensive data processing procedures; (2) model development, encompassing dataset splitting, hyperparameter optimization through Bayesian optimization and cross-validation, model training and evaluation to obtain the optimal model; (3) model explanation, featuring feature importance analysis and SHAP interpretability analysis to understand parameter contributions to water quality assessment.

2.2. Machine Learning Modeling

To predict water quality grades, six distinct machine learning models were utilized in this study: Logistic Regression, Random Forest, CatBoost, XGBoost, Multi-Layer Perceptron (MLP), and Gradient Boosting Decision Tree (GBDT) [9,11,12]. These algorithms span a wide range of computational paradigms from linear methodologies to ensemble strategies and neural network architectures [13,15].

For machine learning modeling, dataset partitioning into training and testing subsets is essential. Models derive patterns from training samples, capturing input–output feature

relationships to build a predictive model for water quality assessment. The testing subset then evaluates the model's performance, gauging its generalization ability.

Hyperparameter tuning was performed to find the best hyperparameters for the six models [21,22]. Bayesian optimization with Gaussian process regression was used for this, and 5-fold cross-validation was incorporated into the search to prevent overfitting [17]. The hyperparameter search ranges and optimal values for each model are detailed in Table 2.

Table 2. Hyperparameter search ranges and optimal values for machine learning models.

Model	Hyperparameter	Search Range	Optimal Value
Logistic Regression	C	[0.01, 100]	1.0
	max_iter	[100, 2000]	1000
	solver	[liblinear, lbfgs]	lbfgs
Random Forest	n_estimators	[10, 200]	40
	max_depth	[3, 20]	10
	min_samples_split	[2, 20]	2
	min_samples_leaf	[1, 10]	1
CatBoost	n_estimators	[50, 300]	50
	learning_rate	[0.01, 0.3]	0.1
	depth	[3, 10]	6
	l2_leaf_reg	[1, 10]	3
XGBoost	n_estimators	[50, 300]	100
	learning_rate	[0.01, 0.3]	0.1
	max_depth	[3, 10]	6
	subsample	[0.6, 1.0]	0.8
MLP	hidden_layer_sizes	[(50), (100,50)]	(100, 50)
	learning_rate_init	[0.001, 0.1]	0.001
	alpha	[0.0001, 0.01]	0.0001
	max_iter	[200, 1000]	500
GBDT	n_estimators	[50, 200]	50
	learning_rate	[0.01, 0.3]	0.1
	max_depth	[3, 10]	3

2.3. Model Evaluation

To evaluate the performance of the developed machine learning models, five widely used metrics were adopted: accuracy, precision, recall, and F1-score. The mathematical formulations for these metrics are given as follows.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Here, TP denotes true positive, TN true negative, FP false positive, and FN false negative.

2.4. Feature Importance Analysis

Understanding which water quality parameters contribute most significantly to assessment decisions is crucial for both model interpretation and practical water management applications. Two complementary approaches were employed to analyze feature importance: model-specific importance measures and unified SHAP analysis.

2.4.1. Model-Specific Feature Importance

Feature importance was extracted using model-specific methods tailored to each algorithm's internal structure. Tree-based models (Random Forest, CatBoost, XGBoost, GBDT) provide intrinsic importance measures based on the frequency and effectiveness of feature splits across all trees in the ensemble. For Logistic Regression, absolute coefficient values represent the magnitude of each feature's contribution to the log-odds of grade prediction. Multi-Layer Perceptron importance was approximated using the mean absolute values of first-layer connection weights, indicating the relative influence of input features on the neural network's initial transformations.

2.4.2. SHAP Analysis for Model Interpretability

To provide unified and theoretically grounded feature importance analysis, SHAP was applied to the best-performing XGBoost model [13,23,24]. SHAP values are based on cooperative game theory, specifically the Shapley value concept from coalition games, which fairly distributes the contribution of each feature to individual predictions [1,34].

The SHAP framework decomposes each prediction into additive feature contributions:

$$f(x) = \phi_0 + \sum_{i=1}^{n} \phi_i \tag{5}$$

where $f(x)$ is the model prediction, ϕ_0 is the base value (expected model output), and ϕ_i represents the SHAP value (contribution) of feature i. The SHAP values satisfy three fundamental axioms: efficiency (contributions sum to the difference between prediction and base value), symmetry (features with identical contributions receive equal SHAP values), and dummy (irrelevant features receive zero contribution).

For tree-based models like XGBoost, SHAP values are computed using TreeExplainer, which efficiently calculates exact Shapley values by leveraging the tree structure. The algorithm traces all possible paths through the decision trees and computes the marginal contribution of each feature across all coalitions.

3. Results and Discussion

3.1. Data Statistical Results and Visualization

The statistical distribution of nine input features and water quality grades from 79,015 measurements collected during 1 January to 14 February 2025, is illustrated in Table 3. The descriptive statistics can effectively reveal the central tendencies, distribution variations, and highlight the characteristics of the data [4,29]. In the current study, the average temperature value is 9.40 °C, which is mainly influenced by seasonal variations and geographical differences across China's monitoring network [2]. The pH values are predominantly distributed around neutral conditions (mean = 7.93), with most measurements clustering between 7.0 and 8.5. The dissolved oxygen content exhibits significant variation, primarily concentrated around 10.96 mg/L. The mean values of conductivity and turbidity are 677.41 μS/cm and 16.47 NTU, respectively, with some water bodies showing extremely high conductivity values reaching 3146.04 μS/cm. The permanganate index is distributed between 0.20 and 11.30 mg/L, and the nutrient parameters (NH_3-N, total phosphorus, total nitrogen) show considerable dispersion, posing challenges for water quality assessment.

Figure 3 presents the Pearson correlation coefficients across all features. It is evident that the absolute values of the Pearson correlation coefficients between each input feature and water quality assessment mostly fall below 0.5. This points to a non-linear association between the classification and these input features, thereby underscoring the necessity and appropriateness of using machine learning approaches to predict water quality assessment.

Table 3. Descriptive statistics of water quality parameters in the dataset.

Statistic	Temperature (°C)	pH (-)	DO (mg/L)	Conductivity (μS/cm)	Turbidity (NTU)	COD_{Mn} (mg/L)	NH_3-N (mg/L)	TP (mg/L)	TN (mg/L)
Central Tendency	9.40 (9.10)	7.93 (8.00)	10.96 (10.80)	677.41 (465.10)	16.47 (8.80)	2.67 (2.20)	0.138 (0.040)	0.050 (0.039)	3.30 (2.21)
Variability	5.25 [5.30–12.60]	0.55 [8.00–8.00]	2.46 [9.40–12.30]	655.00 [292.00–806.35]	20.80 [4.40–18.80]	1.84 [1.30–3.70]	0.207 [0.020–0.150]	0.042 [0.020–0.066]	3.06 [1.37–4.04]
Range	0.01 32.70	4.00 9.00	0.10 22.20	0.0002 3146.04	0.01 95.45	0.20 11.30	0.020 0.970	0.005 0.244	0.05 17.21

Values shown as mean (median) for central tendency; standard deviation [Q1–Q3] for variability; min/max for range. DO: dissolved oxygen; COD_{Mn}: permanganate index; NH_3-N: ammonia nitrogen; TP: total phosphorus; TN: total nitrogen.

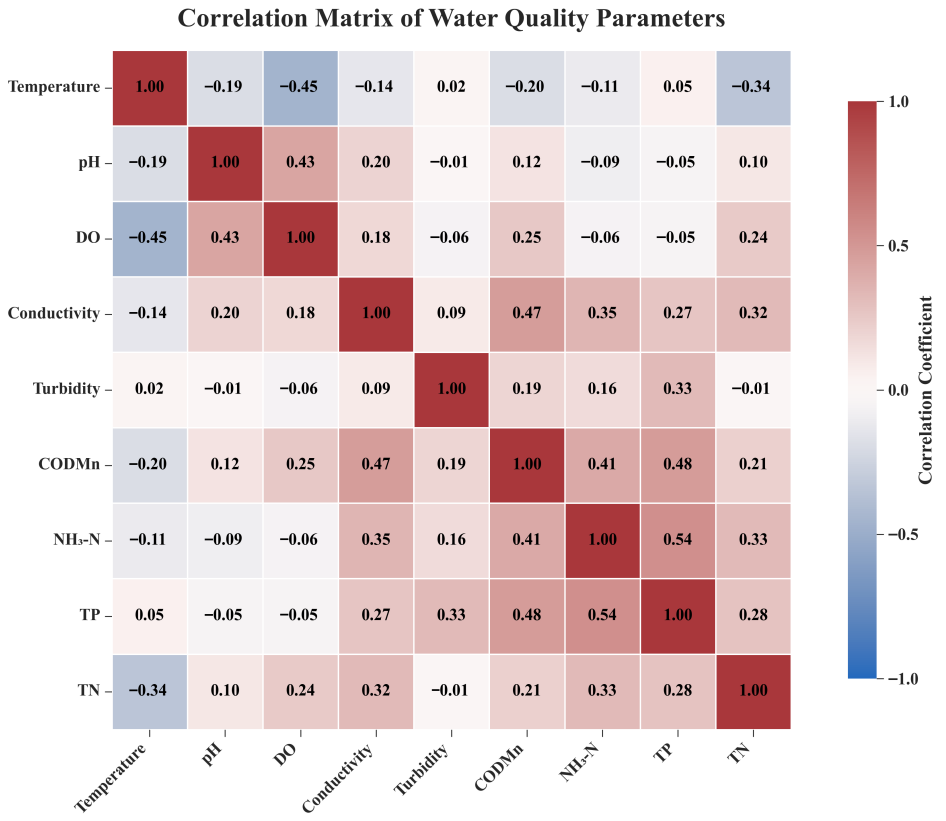

Figure 3. Correlation matrix of water quality parameters showing Pearson correlation coefficients between nine physicochemical variables. The color scale represents correlation strength, with red indicating positive correlations, blue indicating negative correlations, and color intensity reflecting correlation magnitude.

The distribution characteristics of individual water quality parameters are further illustrated in Figure 4, which presents histograms for all nine physicochemical indicators across the entire dataset. The distributions reveal distinct patterns that provide insights into the environmental conditions and pollution sources affecting China's surface waters. Temperature shows a normal distribution centered around 10 °C, reflecting the winter sampling period. pH exhibits a narrow distribution concentrated between 7.5 and 8.5, indicating predominantly alkaline conditions typical of natural surface waters. Dissolved

oxygen displays a right-skewed distribution with most values between 8 and 12 mg/L, suggesting generally healthy oxygenation levels. Conductivity shows a highly right-skewed distribution with most measurements below 1000 µS/cm but extending to much higher values, indicating varying degrees of mineralization and potential pollution. Turbidity demonstrates an exponential decay pattern, with most waters having low turbidity but some showing extreme values up to 100 NTU. The nutrient parameters (COD_{Mn}, NH_3-N, TP, TN) all exhibit right-skewed distributions with long tails, characteristic of environmental data where most samples have low concentrations but pollution hotspots create extreme values. Figure 5 presents the count of surface water quality measurement times by province across China. The results show variation in the number of measurements collected from different provinces during the study period.

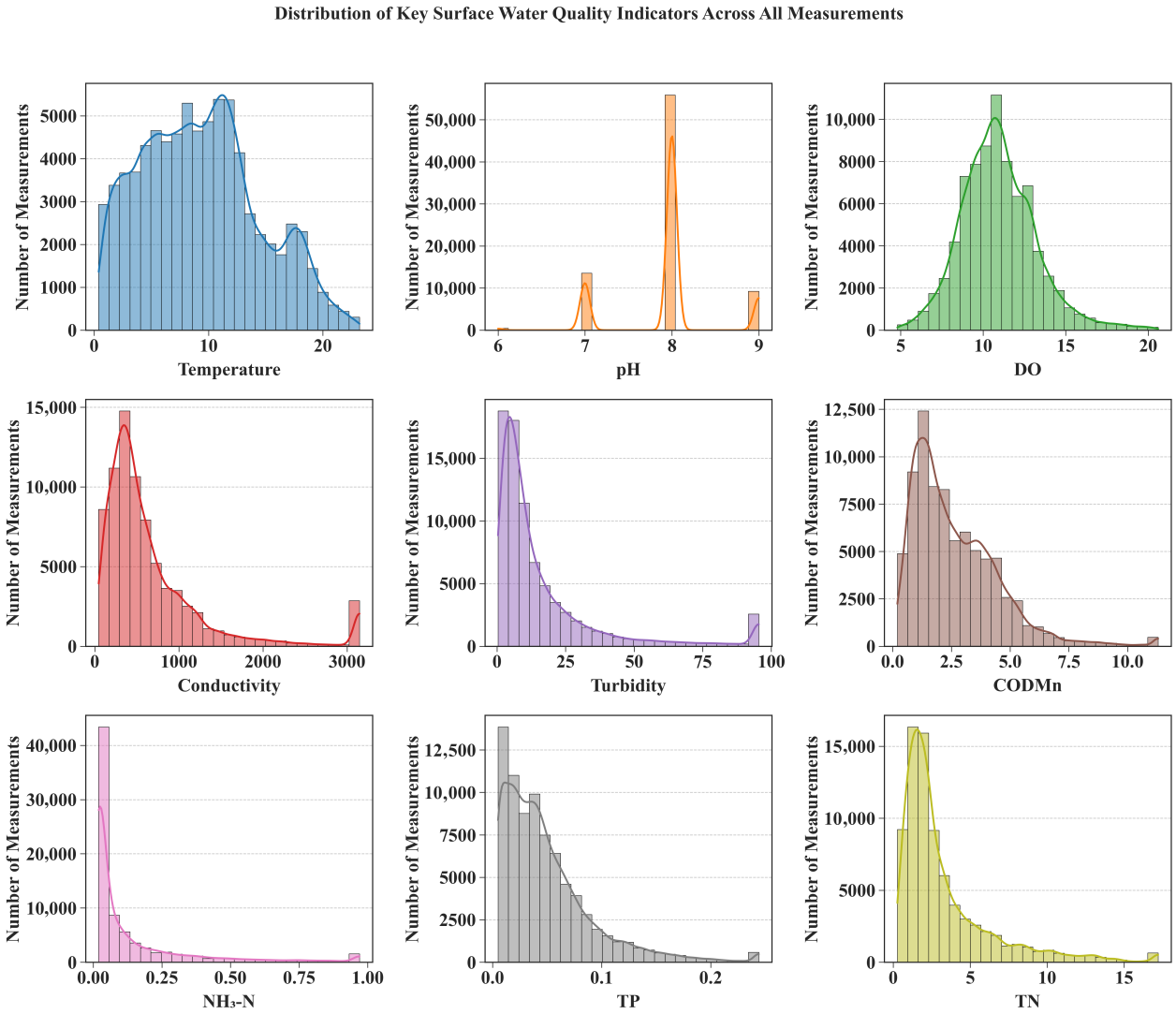

Figure 4. Distribution of key surface water quality indicators across all measurements in the dataset. The nine subplots show histograms with kernel density estimation curves for each physicochemical parameter: temperature (°C), pH, dissolved oxygen (DO, mg/L), conductivity (µS/cm), turbidity (NTU), permanganate index (COD_{Mn}, mg/L), ammonia nitrogen (NH_3-N, mg/L), total phosphorus (TP, mg/L), and total nitrogen (TN, mg/L). The distributions reveal varying patterns from normal (temperature) to highly right-skewed (nutrient parameters), reflecting the diverse environmental conditions across China's surface water monitoring network.

Figure 6 presents the distribution characteristics of water quality parameters across the five classification levels, revealing distinct patterns associated with pollution gradients. Class I waters (excellent quality) consistently demonstrate optimal parameter ranges: low turbidity (median < 5 NTU), high dissolved oxygen (median > 10 mg/L), minimal nutrient concentrations (NH_3-N < 0.05 mg/L, total phosphorus < 0.02 mg/L), and low organic pollution indicators (COD_{Mn} < 2 mg/L). The progressive deterioration from Class I to Class V is evident across all parameters, with Class V waters exhibiting elevated turbidity, reduced dissolved oxygen, and substantially increased concentrations of NH_3-N, total phosphorus, and total nitrogen.

Particularly notable is the dramatic increase in nutrient concentrations with declining water quality classes. NH_3-N concentrations increase exponentially from Class I (median ≈ 0.02 mg/L) to Class V (median > 0.5 mg/L), indicating progressive eutrophication and organic pollution. Total phosphorus follows a similar pattern, with Class V waters showing concentrations exceeding 0.15 mg/L, well above the threshold for eutrophic conditions. The temperature and pH parameters show relatively smaller variations across quality classes, suggesting that thermal and acid–base conditions are less discriminative for water quality assessment compared to chemical and biological indicators.

These statistical patterns provide crucial insights for machine learning model development, indicating that nutrient parameters and dissolved oxygen are likely to emerge as the most important features for water quality assessment, while physical parameters may serve as supporting indicators for specific environmental conditions.

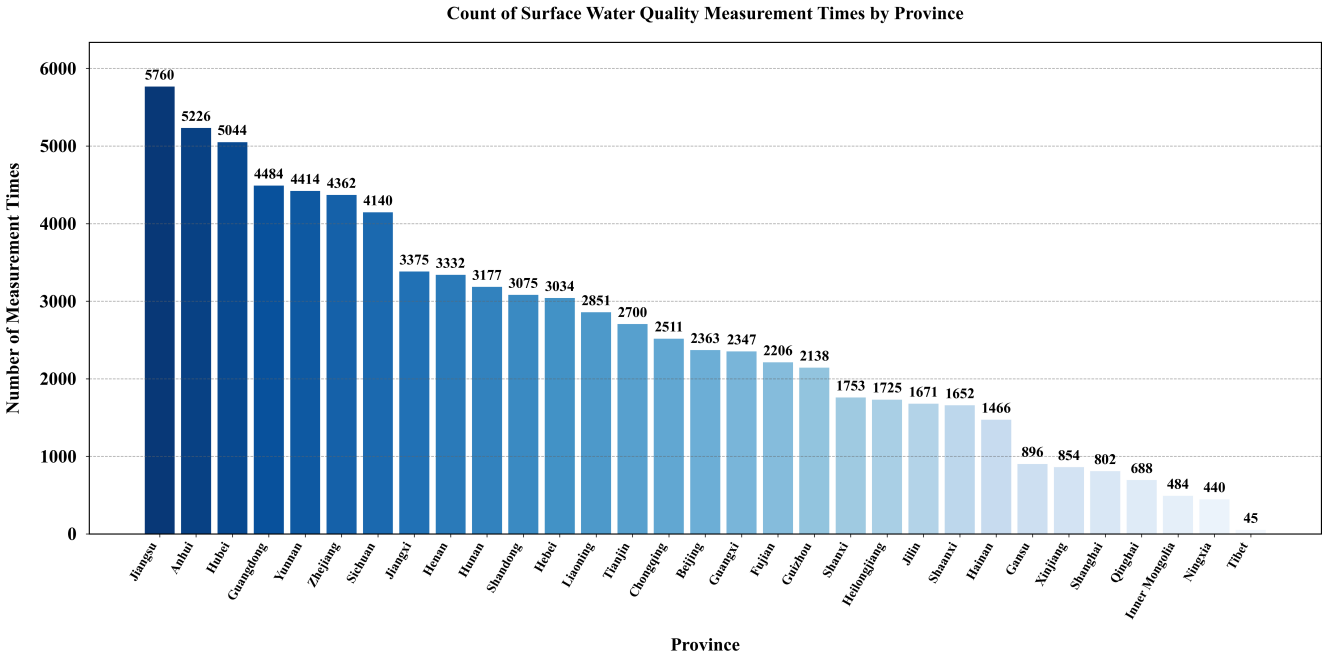

Figure 5. Count of surface water quality measurement times by province in China. The bar chart displays the total number of water quality measurements collected from each province during the study period, illustrating the geographical distribution of data collection intensity across different administrative regions.

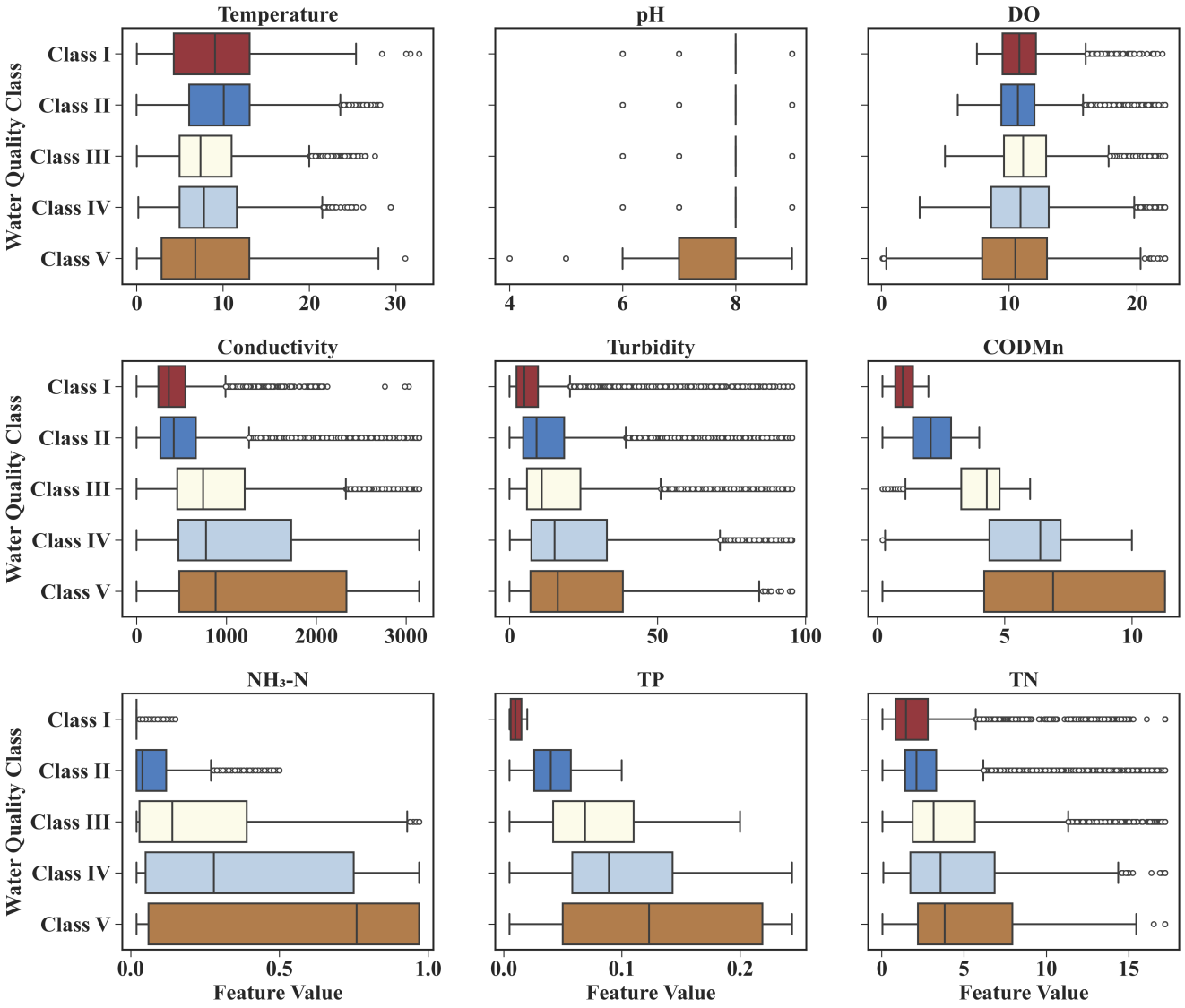

Figure 6. Box plots showing the distribution of water quality parameters across five classification levels (Class I to Class V), revealing pollution gradients and parameter discrimination capabilities.

3.2. Comparison of Model Performance

Six machine learning models were trained and optimized using the identical training set, with the optimal hyperparameters for each model detailed in Table 2. Subsequently, these optimal hyperparameters were utilized in their respective models to predict water quality on the testing set. The comparison between the predicted and actual values for the six models on both the training and testing sets is shown in Table 4.

It is evident that the XGBoost model demonstrated the best predictive performance and generalization ability among all tested algorithms, consistent with findings from previous water quality prediction studies [19,35,36]. XGBoost achieved the highest accuracy of 99.04% on the testing set, significantly outperforming other models. The superior performance of XGBoost can be attributed to its advanced gradient boosting algorithm and effective regularization techniques [16,27]. MLP model showed competitive performance with 97.14% accuracy, followed by Random Forest (97.07%) and GBDT (96.82%). CatBoost achieved 96.45% accuracy, while Logistic Regression showed the lowest performance with 81.69% accuracy, highlighting the effectiveness of ensemble methods over linear approaches in complex environmental classification tasks [6,11].

Table 4. Comprehensive performance comparison of machine learning models for water quality assessment.

Model	Metric	Class-Specific Performance					Average Performance	
		Class I	Class II	Class III	Class IV	Class V	Macro Avg	Weighted Avg
XGBoost	Precision	**0.9955**	**0.9953**	**0.9937**	**0.9431**	**0.9075**	**0.9670**	**0.9903**
	Recall	**1.0000**	**0.9970**	**0.9866**	**0.9565**	**0.8561**	**0.9593**	**0.9904**
	F1-Score	**0.9977**	**0.9962**	**0.9902**	**0.9498**	**0.8811**	**0.9630**	**0.9903**
	Accuracy				**0.9904**			
Random Forest	Precision	0.9799	0.9680	0.9844	0.9218	0.9478	0.9604	0.9707
	Recall	1.0000	0.9927	0.9336	0.8824	0.7712	0.9160	0.9707
	F1-Score	0.9899	0.9802	0.9583	0.9017	0.8505	0.9361	0.9703
	Accuracy				0.9707			
CatBoost	Precision	0.9752	0.9667	0.9771	0.8767	0.9007	0.9393	0.9640
	Recall	1.0000	0.9909	0.9299	0.8517	0.6203	0.8786	0.9645
	F1-Score	0.9874	0.9787	0.9529	0.8641	0.7346	0.9035	0.9635
	Accuracy				0.9645			
MLP	Precision	0.9661	0.9888	0.9640	0.9027	0.8377	0.9319	0.9714
	Recall	0.9977	0.9799	0.9676	0.8724	0.8278	0.9291	0.9714
	F1-Score	0.9817	0.9843	0.9658	0.8873	0.8327	0.9304	0.9713
	Accuracy				0.9714			
GBDT	Precision	0.9767	0.9699	0.9801	0.8995	0.8952	0.9443	0.9679
	Recall	1.0000	0.9913	0.9350	0.8674	0.7052	0.8998	0.9682
	F1-Score	0.9882	0.9805	0.9570	0.8832	0.7889	0.9196	0.9676
	Accuracy				0.9682			
Logistic Regression	Precision	0.8619	0.8564	0.7201	0.6490	0.6753	0.7525	0.8123
	Recall	0.8801	0.8870	0.7293	0.4968	0.2453	0.6477	0.8169
	F1-Score	0.8709	0.8714	0.7247	0.5628	0.3599	0.6779	0.8120
	Accuracy				0.8169			

Bold values indicate the best performance for each metric. Macro Avg: unweighted mean across all classes; Weighted Avg: weighted by class support.

Figure 7 displays the confusion matrices for all six models on the testing set. A greater concentration of data points along the diagonal indicates superior model performance. XGBoost shows the most concentrated diagonal pattern with minimal misclassification errors. All tree-based models (XGBoost, Random Forest, CatBoost, GBDT) demonstrate excellent performance for Classes I and II, with classification accuracies exceeding 95%. However, performance degradation becomes evident for Classes IV and V, particularly in the Logistic Regression model.

Figure 8 illustrates the class-specific performance comparison across all models. The results show that Classes I and II are consistently well-classified by all models, while Classes IV and V present greater challenges, which is consistent with previous studies on imbalanced water quality datasets [28,37]. XGBoost maintains the most stable performance across all water quality classes. The performance differences between models highlight the importance of algorithm selection for multi-class environmental classification tasks [38,39].

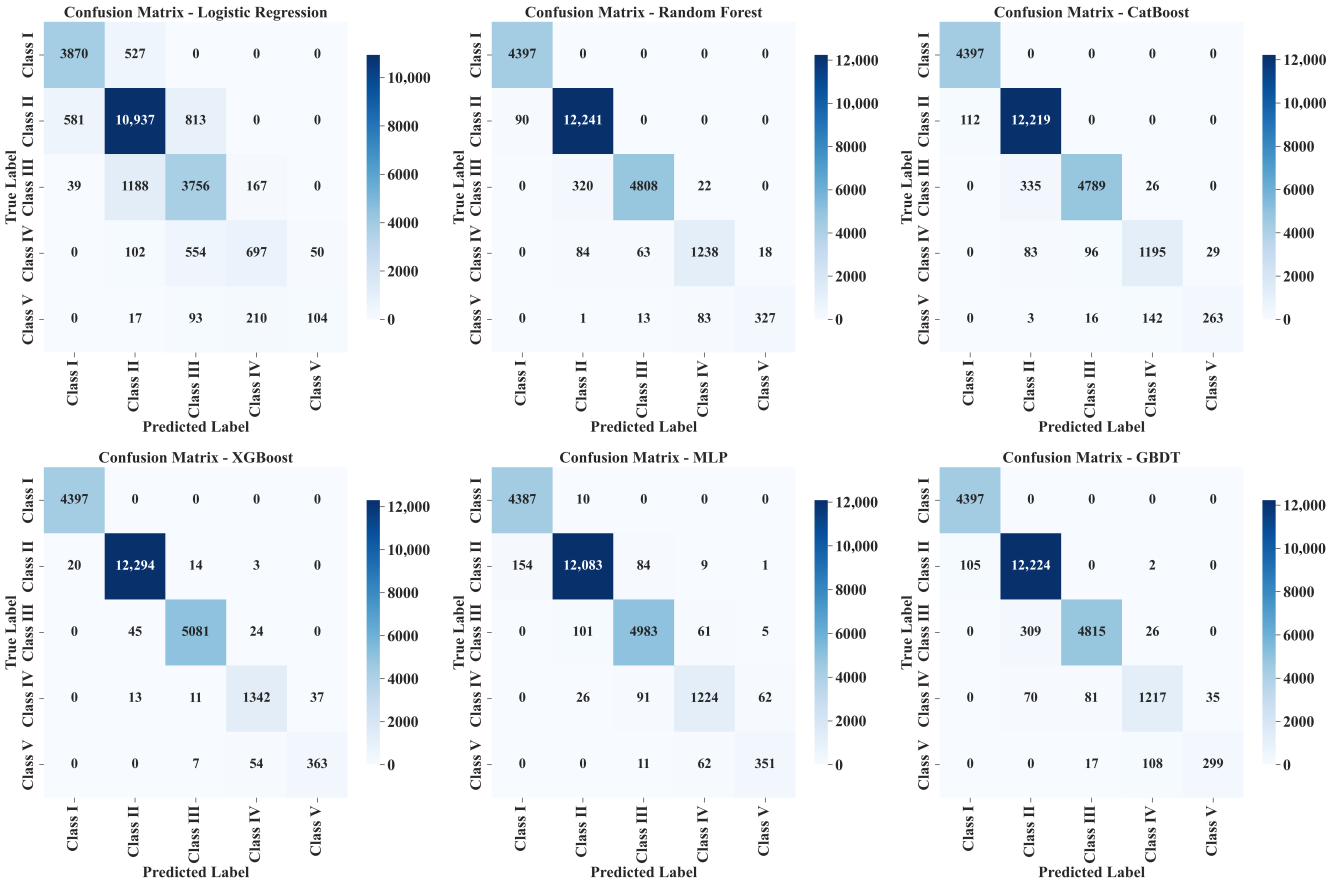

Figure 7. Confusion matrices for all six machine learning models, showing classification accuracy patterns and misclassification distributions across water quality classes.

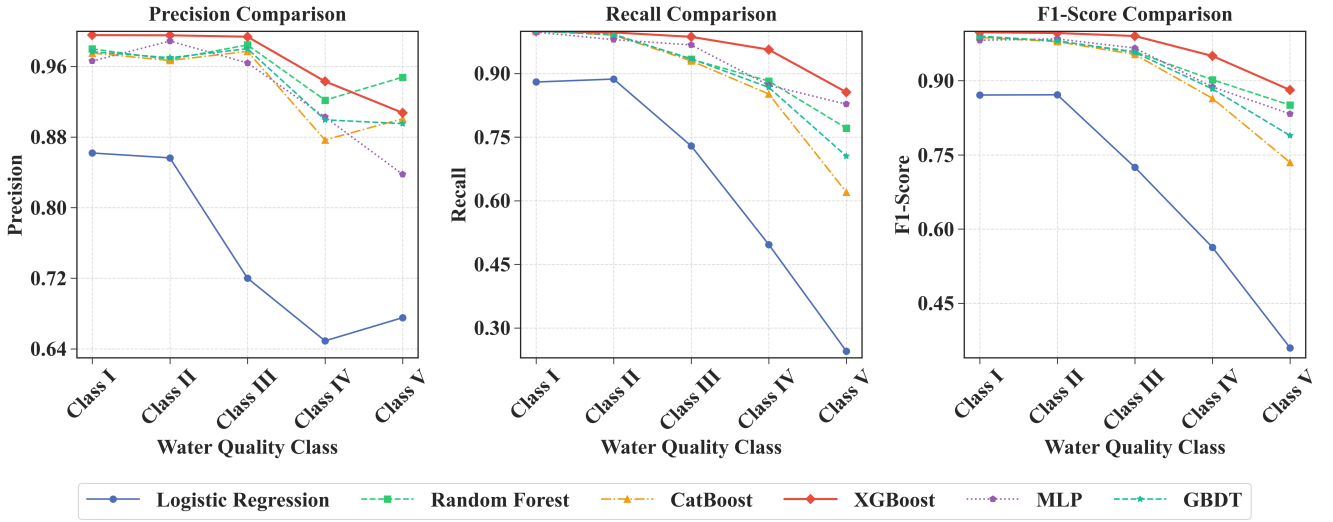

Figure 8. Class-specific performance comparison across all models, displaying precision, recall, and F1-score trends for each water quality class.

3.3. Feature Importance

Understanding the relative contribution of water quality parameters to classification decisions provides critical insights for both model interpretability and environmental management applications [9,24]. Feature importance analysis reveals which physicochemical parameters serve as the most discriminative indicators for water quality assessment across different machine learning algorithms [13,32].

Figure 9 presents the feature importance rankings for all six machine learning models, demonstrating both consistency and algorithmic specificity in parameter prioritization. Across all models, nutrient-related parameters emerge as the most critical discriminators for water quality assessment. Total phosphorus (TP) consistently ranks among the top three most important features across five of the six models, reflecting its fundamental role as a limiting nutrient and key indicator of eutrophication potential in aquatic ecosystems.

The permanganate index (COD_{Mn}), representing organic pollution load, demonstrates exceptional importance across tree-based models, ranking first in XGBoost, Random Forest, CatBoost, and GBDT algorithms. This consistent prioritization underscores the significance of organic matter content in distinguishing water quality classes, as elevated organic pollution directly correlates with deteriorating water conditions and ecosystem health.

NH_3-N exhibits moderate to high importance across all models, particularly in linear regression, where it ranks third. This parameter serves as a critical indicator of recent organic pollution and potential toxicity to aquatic organisms, explaining its consistent relevance in classification algorithms. The nitrogen–phosphorus nutrient complex (TP, TN, NH_3-N) collectively dominates the feature importance rankings, confirming the central role of nutrient pollution in water quality degradation.

Physical parameters demonstrate varied importance across different algorithmic approaches. Dissolved oxygen (DO) shows relatively lower importance in most models, likely due to its high temporal variability and complex relationships with temperature, biological activity, and pollution loads. However, its inclusion remains valuable for capturing oxygen depletion scenarios characteristic of severely polluted waters.

Temperature and pH exhibit consistently low importance across all models, suggesting that thermal and acid–base conditions provide limited discriminative power for water quality assessment compared to chemical pollution indicators. This finding aligns with the relatively narrow pH distribution observed in the dataset (mean = 7.93 ± 0.55) and the standardization procedures applied during preprocessing.

Electrical conductivity demonstrates variable importance across models, ranking moderately in MLP and Logistic Regression but showing lower significance in tree-based algorithms. This parameter reflects the total ionic content and can indicate both natural mineral content and pollution-derived ions, contributing to its moderate discriminative value.

The algorithmic differences in feature importance rankings reveal complementary perspectives on parameter significance. Linear models (Logistic Regression) emphasize individual parameter contributions through coefficient magnitudes, while tree-based models capture complex interactions and non-linear relationships between parameters. Neural networks (MLP) demonstrate unique prioritization patterns, potentially reflecting their ability to identify subtle parameter combinations not readily apparent in other approaches.

These feature importance patterns provide valuable guidance for water quality monitoring optimization. The dominance of nutrient parameters (TP, COD_{Mn}, NH_3-N) suggests that monitoring programs should prioritize accurate measurement of these indicators to maximize classification accuracy. Additionally, the consistent low importance of temperature and pH indicates that resources might be more effectively allocated to enhanced nutrient monitoring rather than expanding thermal or acid–base measurements.

From a water management perspective, the feature importance analysis confirms that nutrient pollution control represents the most critical intervention point for water quality improvement. The high importance of phosphorus and nitrogen compounds aligns with established eutrophication management strategies and supports targeted pollution control measures focusing on agricultural runoff, wastewater treatment optimization, and point source discharge regulation.

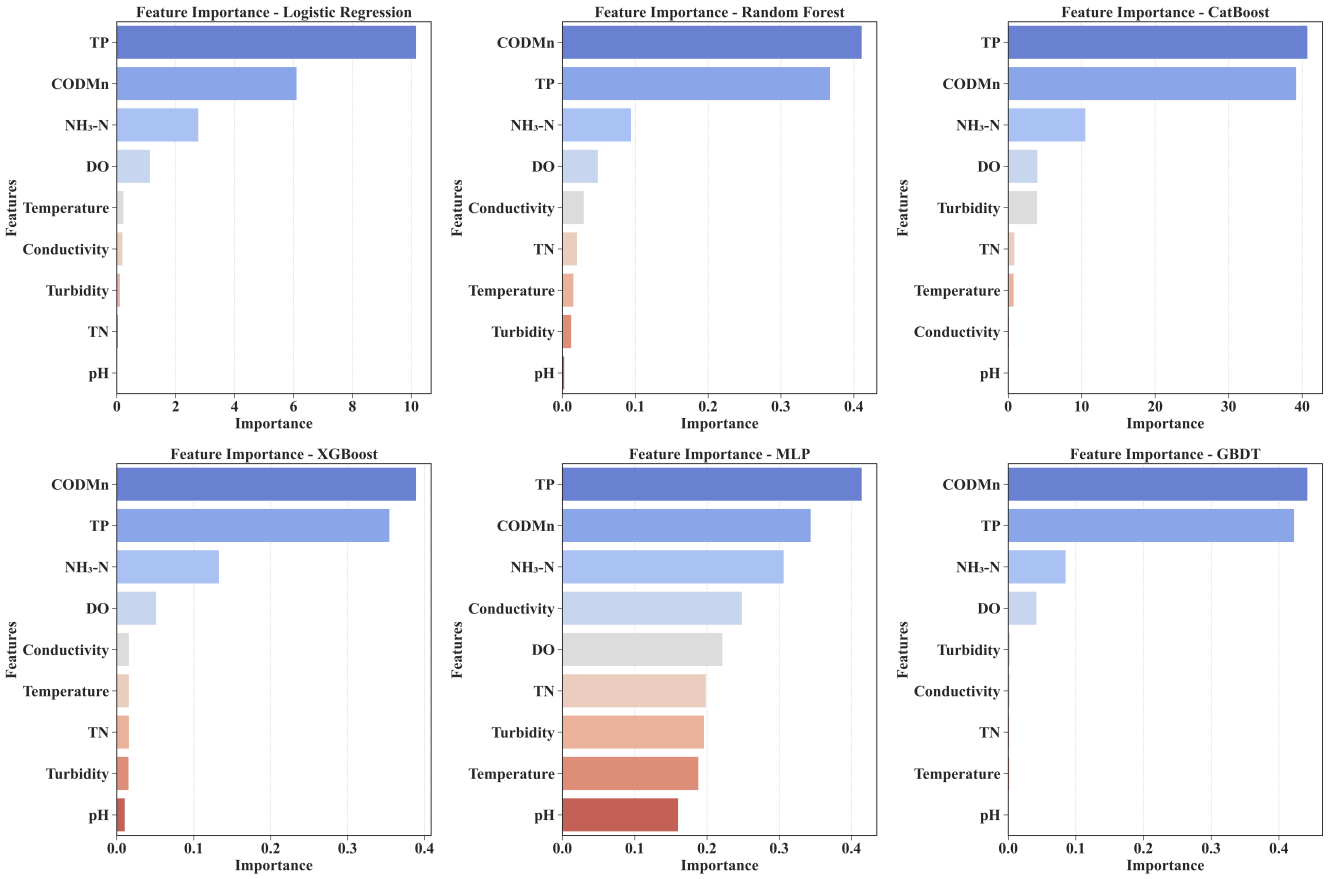

Figure 9. Feature importance comparison across six machine learning models for water quality assessment. Each subplot displays the relative importance of nine physicochemical parameters, revealing consistent prioritization of nutrient-related indicators (TP, COD_{Mn}, NH_3-N) across different algorithmic approaches.

3.4. Shapley Additive Explanations (SHAP)

To provide deeper insights into the decision-making mechanisms of the best-performing XGBoost model, Shapley Additive Explanations (SHAP) analysis was conducted to quantify individual feature contributions to specific predictions and reveal class-specific importance patterns [13,23]. SHAP analysis offers a unified framework for model interpretability by decomposing each prediction into additive contributions from individual features, providing both global and local explanations of model behavior [34,40].

Figure 10 presents the global SHAP feature importance analysis for XGBoost across all water quality classes, revealing both overall parameter significance and class-specific discrimination patterns. The global importance ranking confirms the dominance of chemical pollution indicators, with permanganate index (COD_{Mn}) demonstrating the highest overall importance, followed by total phosphorus (TP) and ammonia nitrogen (NH_3-N). This global perspective validates the conventional feature importance analysis while providing additional insights into the directional influence of parameters on classification decisions.

The class-specific SHAP importance analysis reveals distinct discrimination patterns across water quality categories. For Class I (excellent water quality), total phosphorus emerges as the most critical discriminator, with SHAP values indicating that lower TP concentrations strongly contribute to excellent quality classification. This finding underscores the fundamental role of phosphorus limitation in maintaining pristine aquatic conditions and aligns with established limnological principles regarding nutrient thresholds for oligotrophic systems.

Class II and Class III classifications demonstrate the increasing importance of organic pollution indicators, with COD_{Mn} showing progressively higher SHAP importance values. The transition from nutrient-limited (Class I) to organically influenced (Classes II–III) discrimination patterns reflects the hierarchical nature of water quality degradation, where initial pollution manifests through increased organic matter content before severe nutrient enrichment occurs.

For severely polluted waters (Classes IV and V), the SHAP analysis reveals a shift toward multi-parameter discrimination, with both COD_{Mn} and total phosphorus maintaining high importance while NH_3-N gains significant relevance. Class V discrimination shows particularly high SHAP values for COD_{Mn}, indicating that extreme organic pollution serves as the primary indicator for heavily polluted water classification. The elevated importance of NH_3-N in Class V reflects the toxic effects of ammonia accumulation in severely degraded aquatic systems.

Figure 11 presents waterfall plots for representative samples from each water quality class, illustrating how individual parameter values contribute to specific classification decisions. These local explanations demonstrate the additive nature of SHAP contributions, where each parameter either increases or decreases the probability of a particular class assignment relative to the model's baseline prediction.

The waterfall analysis for Class I samples reveals that low concentrations of pollution indicators (negative SHAP values for COD_{Mn}, TP, NH_3-N) collectively drive the model toward excellent quality classification. The base value f(x) represents the model's average prediction across all samples, while individual parameter contributions push the final prediction toward the correct class. For Class I, multiple parameters contribute negative SHAP values, indicating that their low concentrations collectively support excellent water quality assessment.

Class II waterfall plots demonstrate a more balanced contribution pattern, with some parameters contributing positively and others negatively to the classification decision. The moderate SHAP values reflect the intermediate nature of good water quality, where parameter concentrations fall between pristine and polluted thresholds. Conductivity often shows positive SHAP contributions in Class II samples, suggesting that moderate ionic content serves as a distinguishing characteristic of good-quality waters.

For moderately and heavily polluted samples (Classes IV and V), the waterfall plots reveal dominant positive SHAP contributions from pollution indicators, particularly COD_{Mn} and TP. Class V samples consistently show large positive SHAP values for organic pollution parameters, with f(x) values indicating strong model confidence in heavily polluted classification. The magnitude of individual parameter contributions in Class V samples often exceeds those in higher quality classes, reflecting the more extreme parameter values characteristic of severely degraded waters.

The SHAP analysis provides several critical insights for water quality management and monitoring optimization. First, the class-specific importance patterns suggest that monitoring strategies should be tailored to expected pollution levels, with enhanced nutrient monitoring for pristine systems and comprehensive organic pollution assessment for degraded waters. Second, the waterfall plots demonstrate that water quality assessment relies on parameter combinations rather than single indicators, supporting the multi-parameter monitoring approaches commonly employed in environmental assessment programs.

From a model interpretability perspective, the SHAP analysis confirms that XGBoost's superior performance stems from its ability to capture complex parameter interactions and non-linear relationships between water quality indicators. The varying importance patterns across classes indicate that the model has learned class-specific discrimination

rules that align with established environmental science principles, providing confidence in its practical applicability for automated water quality assessment.

The threshold effects revealed through SHAP analysis also provide guidance for water quality standards development and revision. The sharp transitions in parameter importance between classes suggest natural breakpoints in the parameter-quality relationship continuum, supporting the discrete classification approach employed in water quality standards such as GB 3838-2002.

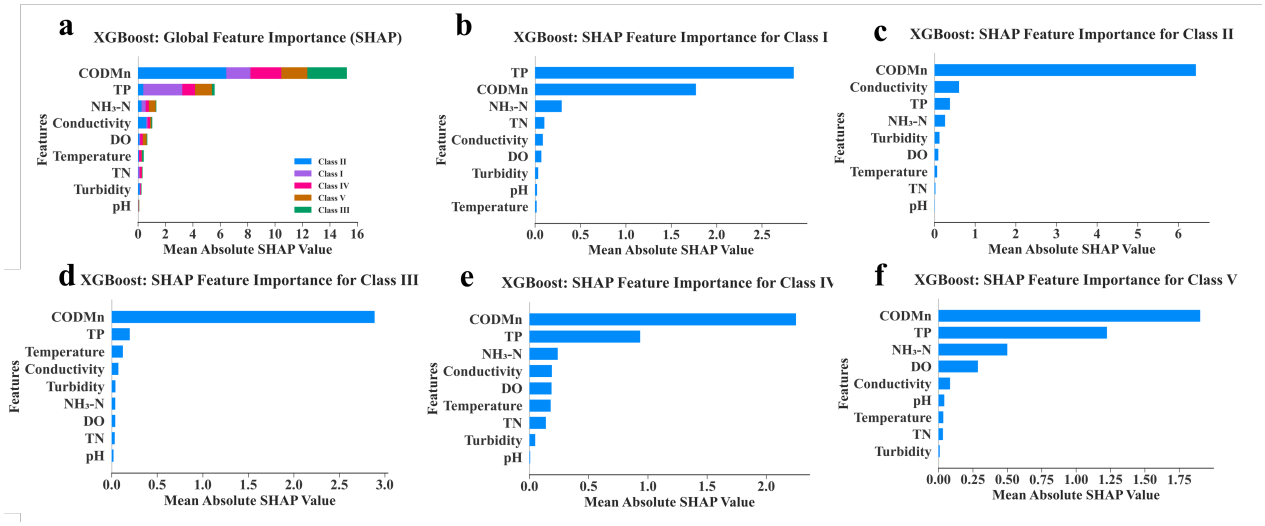

Figure 10. SHAP global and class-specific feature importance analysis for XGBoost model. (**a**) Global feature importance showing overall parameter significance across all classes with color-coded class contributions. (**b**–**f**) Class-specific SHAP importance, revealing distinct discrimination patterns for each water quality category and demonstrating the evolution from nutrient-dominated (Class I) to multi-parameter (Classes IV–V) classification mechanisms.

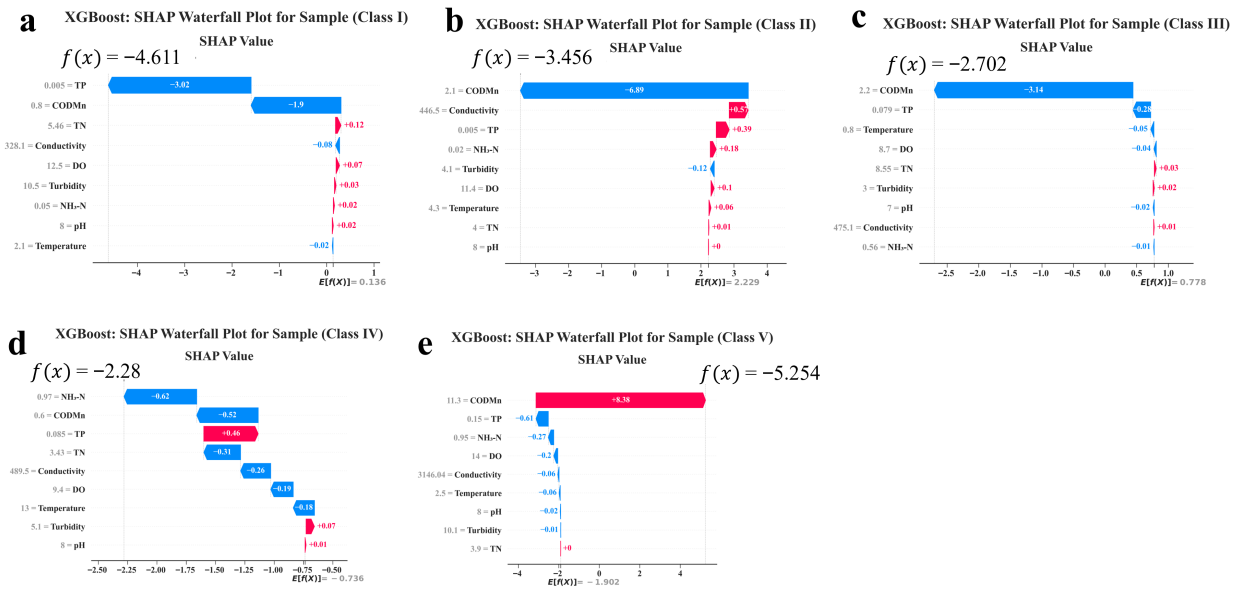

Figure 11. SHAP waterfall plots for XGBoost predictions of samples from water quality Classes I–V: (**a**) Class I sample, with TP and CODMn as major negative contributors; (**b**) Class II sample, where CODMn and Conductivity drive negative contributions; (**c**) Class III sample, with CODMn exerting negative influence; (**d**) Class IV sample, where NH_3-N and CODMn are key negative contributors; (**e**) Class V sample, with CODMn as a prominent positive contributor.

4. Environmental Implications

Surface water contamination represents one of the most pressing environmental challenges in China, with nutrient pollution and organic contamination posing significant threats to aquatic ecosystems and human health [2,21]. Traditional water quality assessment methods often rely on expert judgment and simple statistical approaches, which may not capture the complex interactions between physicochemical parameters and water quality status [10,41]. The machine learning approach developed in this study provides a robust framework for understanding these relationships and optimizing water management strategies [7,8].

The consistent dominance of nutrient-related parameters (total phosphorus, permanganate index, ammonia nitrogen) across all machine learning models indicates that eutrophication control should be prioritized in water quality management programs. Current monitoring systems often allocate resources equally across multiple parameters, but the feature importance analysis suggests that enhanced precision in nutrient measurements could significantly improve classification accuracy while reducing monitoring costs. The relatively low importance of temperature and pH parameters indicates that these measurements, while environmentally relevant, provide limited discriminative power for water quality assessment.

The class-specific SHAP importance patterns reveal that water quality management strategies should be tailored to the current pollution status of water bodies. For pristine systems (Class I), phosphorus monitoring provides the greatest discriminative value, supporting early detection of eutrophication risks before irreversible ecosystem changes occur. In moderately polluted systems (Classes II–III), organic pollution indicators become increasingly critical, requiring robust permanganate index monitoring capabilities. For heavily degraded systems (Classes IV–V), comprehensive multi-parameter monitoring remains essential, with particular emphasis on NH_3-N due to its direct toxicity implications for aquatic organisms.

The machine learning framework also enables the development of adaptive monitoring protocols that can respond dynamically to changing water quality conditions. Systems consistently classified as excellent quality could transition to reduced monitoring intensity, while waters showing classification uncertainty or degradation trends could trigger enhanced monitoring protocols and targeted intervention measures. This adaptive approach represents a significant advancement over current static monitoring programs and could substantially improve the efficiency of water quality management resources.

5. Conclusions

In this research, a comprehensive dataset with water quality measurements from China's national surface water monitoring network was utilized to develop machine learning models for automated water quality assessment [2,9]. Six machine learning algorithms—Logistic Regression, Random Forest, CatBoost, XGBoost, MLP, and GBDT—were trained and assessed for their capacity to classify water quality into five categories as per GB 3838-2002 standards [11,12]. The optimized XGBoost model showed superior predictive performance, achieving 99.04% accuracy, which aligns with recent progress in gradient boosting for water quality prediction [19,35]. Feature importance analysis in the machine learning models highlighted the significance of nine physicochemical parameters in water quality assessment [9,32]. Further analysis via SHAP unveiled the influence of each parameter on classification decisions, revealing the underlying discrimination mechanisms across different water quality classes [13,23]. The key conclusions are as follows:

(1) Through comprehensive evaluation of classification metrics and comparison of model performances on both training and testing datasets, XGBoost proved to have the best

predictive accuracy and generalization ability among all tested algorithms. The outstanding performance of tree-based ensemble methods stems from their ability to capture complex nonlinear relationships and parameter interactions that exist in environmental systems.

(2) Nutrient-related parameters were found to be the most crucial factors in water quality assessment, with total phosphorus, permanganate index, and NH_3-N consistently being among the top features across all models. This result offers solid scientific evidence for giving priority to nutrient pollution control in water quality management strategies and supports the implementation of targeted eutrophication control measures.

(3) SHAP analysis uncovered distinct class-specific discrimination patterns across water quality categories. Waters of excellent quality (Class I) are mainly differentiated by phosphorus limitation, while severely polluted waters (Classes IV–V) need multi-parameter discrimination approaches involving both nutrient and organic pollution indicators. These findings provide a scientific basis for developing customized monitoring and management strategies suitable for different water quality scenarios.

(4) The consistency of feature importance across multiple algorithmic approaches shows that the machine learning framework can reliably identify the most critical parameters for water quality assessment. Temperature and pH parameters had consistently low importance across all models, implying that monitoring resources could be more efficiently allocated to enhanced measurements of nutrient parameters.

For future studies, integrating temporal dynamics, spatial relationships, and climate change projections could improve the predictive ability of machine learning models for water quality assessment. The combination of remote sensing data and real-time monitoring technologies with advanced machine learning algorithms provides promising prospects for developing comprehensive adaptive water quality management systems.

6. Environmental Implication

Surface water quality degradation poses significant risks to aquatic ecosystems and human health, particularly in rapidly developing regions where industrial and agricultural activities intensify pollution pressures. The machine learning approach developed in this study provides an objective and scientifically robust method for water quality assessment, reducing reliance on subjective expert judgment and improving the consistency of classification decisions. The identification of nutrient parameters as primary drivers of water quality assessment supports the implementation of targeted pollution control measures focusing on phosphorus and nitrogen reduction. This research facilitates the optimization of water quality monitoring programs through strategic parameter prioritization, potentially improving monitoring efficiency while maintaining classification accuracy. The adaptive monitoring capabilities enabled by machine learning frameworks provide enhanced tools for environmental management agencies to respond effectively to changing water quality conditions and implement evidence-based protection strategies for aquatic ecosystem health.

Author Contributions: Conceptualization, Y.L. and Y.S.; Methodology, Y.S.; Software, Y.S.; Validation, Y.S. and Y.L.; Formal analysis, Y.S.; Investigation, Y.S.; Resources, Y.L.; Data curation, Y.L. and Y.S.; Writing—original draft preparation, Y.L., Y.S., L.C. and L.L.; Writing—review and editing, Y.L., Y.S., L.C. and L.L.; Visualization, Y.S.; Supervision, Y.L.; Project administration, Y.L.; Funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Open Fund of State Key Laboratory of Water Resources and Hydropower Engineering Science Foundation (grant number: No.2022WG01). Natural Science Foundation of Hubei Province of China (grant number: No.2022CFB935). Science and Technology Research Project of Henan Province (grant number: No.232102320141).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. Singha, C.; Bhattacharjee, I.; Sahoo, S.; Abdelrahman, K.; Uddin, M.G.; Fnais, M.S.; Govind, A.; Abioui, M. Prediction of urban surface water quality scenarios using hybrid stacking ensembles machine learning model in Howrah Municipal Corporation, West Bengal. *J. Environ. Manag.* **2024**, *370*, 122721. [CrossRef] [PubMed]
2. Islam, M.S.; Yin, H.; Rahman, M. Long-term trend prediction of surface water quality of two main river basins of China using Machine Learning Method. *Procedia Comput. Sci.* **2024**, *236*, 257–264. [CrossRef]
3. El Bilali, A.; Taleb, A. Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment. *J. Saudi Soc. Agric. Sci.* **2020**, *19*, 439–451. [CrossRef]
4. Berihun, M.L.; Bayabil, H.K.; Assegid, Y. Leveraging remote sensing–enabled machine learning for river water quality prediction in South Florida's hydrological systems. *Remote Sens. Appl. Soc. Environ.* **2025**, *38*, 101616. [CrossRef]
5. Jiang, S.; Cheng, X.; Shi, B.; Zhu, D.; Xie, J.; Zhou, Z. Optimal selection of machine learning algorithms for ciprofloxacin prediction based on conventional water quality indicators. *Ecotoxicol. Environ. Saf.* **2025**, *289*, 117628. [CrossRef] [PubMed]
6. Prasad, D.V.V.; Venkataramana, L.Y.; Kumar, P.S.; Prasannamedha, G.; Soumya, K.; Poornema, A. Prediction on water quality of a lake in Chennai, India using machine learning algorithms. *Desalin. Water Treat.* **2021**, *218*, 44–51. [CrossRef]
7. Kaur, A.; Goyal, S.; Batra, N.; Chhabra, K. Chapter 1—Artificial intelligence and machine learning based water quality monitoring, prediction, and analysis: A comprehensive review. In *Computational Automation for Water Security*; Dubey, A.K., Srivastav, A.L., Kumar, A., Garcia Marquez, F.P., Giannakoudakis, D.A., Eds.; Elsevier: Amsterdam, The Netherlands, 2025; pp. 1–10. [CrossRef]
8. Ewuzie, U.; Bolade, O.P.; Egbedina, A.O. Chapter 9—Application of deep learning and machine learning methods in water quality modeling and prediction: A review. In *Current Trends and Advances in Computer-Aided Intelligent Environmental Data Engineering*; Intelligent Data-Centric Systems; Marques, G., Ighalo, J.O., Eds.; Academic Press: Cambridge, MA, USA, 2022; pp. 185–218. [CrossRef]
9. Chen, K.; Chen, H.; Zhou, C.; Huang, Y.; Qi, X.; Shen, R.; Liu, F.; Zuo, M.; Zou, X.; Wang, J.; et al. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Res.* **2020**, *171*, 115454. [CrossRef]
10. Najah Ahmed, A.; Binti Othman, F.; Abdulmohsin Afan, H.; Khaleel Ibrahim, R.; Ming Fai, C.; Shabbir Hossain, M.; Ehteram, M.; Elshafie, A. Machine learning methods for better water quality prediction. *J. Hydrol.* **2019**, *578*, 124084. [CrossRef]
11. Bui, D.T.; Khosravi, K.; Tiefenbacher, J.; Nguyen, H.; Kazakis, N. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Sci. Total Environ.* **2020**, *721*, 137612. [CrossRef]
12. 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]
13. Makumbura, R.K.; Mampitiya, L.; Rathnayake, N.; Meddage, D.; Henna, S.; Dang, T.L.; Hoshino, Y.; Rathnayake, U. Advancing water quality assessment and prediction using machine learning models, coupled with explainable artificial intelligence (XAI) techniques like shapley additive explanations (SHAP) for interpreting the black-box nature. *Results Eng.* **2024**, *23*, 102831. [CrossRef]
14. Adusei, Y.Y.; Quaye-Ballard, J.; Adjaottor, A.A.; Mensah, A.A. Spatial prediction and mapping of water quality of Owabi reservoir from satellite imageries and machine learning models. *Egypt. J. Remote Sens. Space Sci.* **2021**, *24*, 825–833. [CrossRef]
15. Mohseni, U.; Pande, C.B.; Chandra Pal, S.; Alshehri, F. Prediction of weighted arithmetic water quality index for urban water quality using ensemble machine learning model. *Chemosphere* **2024**, *352*, 141393. [CrossRef]
16. del Castillo, A.F.; Garibay, M.V.; Díaz-Vázquez, D.; Yebra-Montes, C.; Brown, L.E.; Johnson, A.; Garcia-Gonzalez, A.; Gradilla-Hernández, M.S. Improving river water quality prediction with hybrid machine learning and temporal analysis. *Ecol. Inform.* **2024**, *82*, 102655. [CrossRef]
17. Poursaeid, M.; Poursaeed, A.H.; Shabanlou, S. Water quality fluctuations prediction and Debi estimation based on stochastic optimized weighted ensemble learning machine. *Process Saf. Environ. Prot.* **2024**, *188*, 1160–1174. [CrossRef]
18. Huan, S. A novel interval decomposition correlation particle swarm optimization-extreme learning machine model for short-term and long-term water quality prediction. *J. Hydrol.* **2023**, *625*, 130034. [CrossRef]

19. Zhang, K.; Wang, X.; Liu, T.; Wei, W.; Zhang, F.; Huang, M.; Liu, H. Enhancing water quality prediction with advanced machine learning techniques: An extreme gradient boosting model based on long short-term memory and autoencoder. *J. Hydrol.* **2024**, *644*, 132115. [CrossRef]
20. Nong, X.; He, Y.; Chen, L.; Wei, J. Machine learning-based evolution of water quality prediction model: An integrated robust framework for comparative application on periodic return and jitter data. *Environ. Pollut.* **2025**, *369*, 125834. [CrossRef]
21. Yan, T.; Zhou, A.; Shen, S.L. Prediction of long-term water quality using machine learning enhanced by Bayesian optimisation. *Environ. Pollut.* **2023**, *318*, 120870. [CrossRef] [PubMed]
22. Shah, M.I.; Javed, M.F.; Alqahtani, A.; Aldrees, A. Environmental assessment based surface water quality prediction using hyper-parameter optimized machine learning models based on consistent big data. *Process Saf. Environ. Prot.* **2021**, *151*, 324–340. [CrossRef]
23. Wang, S.; Peng, H.; Liang, S. Prediction of estuarine water quality using interpretable machine learning approach. *J. Hydrol.* **2022**, *605*, 127320. [CrossRef]
24. Nong, X.; Lai, C.; Chen, L.; Wei, J. A novel coupling interpretable machine learning framework for water quality prediction and environmental effect understanding in different flow discharge regulations of hydro-projects. *Sci. Total Environ.* **2024**, *950*, 175281. [CrossRef]
25. Huang, S.; Xia, J.; Wang, Y.; Lei, J.; Wang, G. Water quality prediction based on sparse dataset using enhanced machine learning. *Environ. Sci. Ecotechnology* **2024**, *20*, 100402. [CrossRef] [PubMed]
26. Zhong, H.; Yuan, Y.; Luo, L.; Ye, J.; Chen, M.; Zhong, C. Water quality prediction of MBR based on machine learning: A novel dataset contribution analysis method. *J. Water Process Eng.* **2022**, *50*, 103296. [CrossRef]
27. Lu, H.; Ma, X. Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere* **2020**, *249*, 126169. [CrossRef] [PubMed]
28. Mori, M.; Gonzalez Flores, R.; Suzuki, Y.; Nukazawa, K.; Hiraoka, T.; Nonaka, H. Prediction of Microcystis Occurrences and Analysis Using Machine Learning in High-Dimension, Low-Sample-Size and Imbalanced Water Quality Data. *Harmful Algae* **2022**, *117*, 102273. [CrossRef]
29. Anand, V.; Oinam, B.; Wieprecht, S. Machine learning approach for water quality predictions based on multispectral satellite imageries. *Ecol. Inform.* **2024**, *84*, 102868. [CrossRef]
30. Awaleh, M.O.; Boschetti, T.; Marlin, C.; Robleh, M.A.; Ahmed, M.M.; Al-Aghbary, M.; Vystavna, Y.; Waberi, M.M.; Dabar, O.A.; Rossi, M.; et al. Geochemical and isotopic studies of the Douda-Damerjogue aquifer (Republic of Djibouti): Origin of high nitrate and fluoride, spatial distribution, associated health risk assessment and prediction of water quality using machine learning. *Sci. Total Environ.* **2025**, *967*, 178789. [CrossRef]
31. M, G.J. Secure water quality prediction system using machine learning and blockchain technologies. *J. Environ. Manag.* **2024**, *350*, 119357. [CrossRef]
32. Saboe, D.; Ghasemi, H.; Gao, M.M.; Samardzic, M.; Hristovski, K.D.; Boscovic, D.; Burge, S.R.; Burge, R.G.; Hoffman, D.A. Real-time monitoring and prediction of water quality parameters and algae concentrations using microbial potentiometric sensor signals and machine learning tools. *Sci. Total Environ.* **2021**, *764*, 142876. [CrossRef]
33. *GB 3838-2002*; Environmental Quality Standards for Surface Water. Ministry of Ecology and Environment of the People's Republic of China: Beijing, China, 2002. Available online: https://www.mee.gov.cn/ywgz/fgbz/bz/bzwb/shjbh/shjzlbz/200206/t20020601_66497.shtml (accessed on 30 August 2025).
34. Grbčić, L.; Družeta, S.; Mauša, G.; Lipić, T.; Lušić, D.V.; Alvir, M.; Lučin, I.; Sikirica, A.; Davidović, D.; Travaš, V.; et al. Coastal water quality prediction based on machine learning with feature interpretation and spatio-temporal analysis. *Environ. Model. Softw.* **2022**, *155*, 105458. [CrossRef]
35. Jiang, Y.; Song, Y.; Liu, J.; Liu, H.; Zang, X.; Ji, Z. Machine learning assisted precise prediction of algae bloom in large-scale water diversion engineering. *Desalination* **2025**, *610*, 118880. [CrossRef]
36. Gao, Z.; Wang, G.; Chen, J.; Fang, L.; Ren, S.; Yinglan, A.; Ji, S.; Liu, R.; Wang, Q. Kalman filtering assimilated machine learning methods significantly improve the prediction performance of water quality parameters. *Ecol. Inform.* **2025**, *90*, 103337. [CrossRef]
37. Rahaman, M.H.; Sajjad, H.; Hussain, S.; Roshani; Masroor, M.; Sharma, A. Surface water quality prediction in the lower Thoubal river watershed, India: A hyper-tuned machine learning approach and DNN-based sensitivity analysis. *J. Environ. Chem. Eng.* **2024**, *12*, 112915. [CrossRef]
38. Koranga, M.; Pant, P.; Kumar, T.; Pant, D.; Bhatt, A.K.; Pant, R. Efficient water quality prediction models based on machine learning algorithms for Nainital Lake, Uttarakhand. *Mater. Today Proc.* **2022**, *57*, 1706–1712. [CrossRef]
39. Deng, T.; Chau, K.W.; Duan, H.F. Machine learning based marine water quality prediction for coastal hydro-environment management. *J. Environ. Manag.* **2021**, *284*, 112051. [CrossRef]

40. Chen, X.; Zhao, C.; Chen, J.; Jiang, H.; Li, D.; Zhang, J.; Han, B.; Chen, S.; Wang, C. Water quality parameters-based prediction of dissolved oxygen in estuaries using advanced explainable ensemble machine learning. *J. Environ. Manag.* **2025**, *380*, 125146. [CrossRef]
41. Imani, M.; Hasan, M.M.; Bittencourt, L.F.; McClymont, K.; Kapelan, Z. A novel machine learning application: Water quality resilience prediction Model. *Sci. Total Environ.* **2021**, *768*, 144459. [CrossRef]