ISSN 0097-8078, Water Resources, 2025, Vol. 52, No. 6, pp. 1340–1355. © Pleiades Publishing, Ltd., 2025.

HYDROCHEMISTRY, HYDROBIOLOGY: ENVIRONMENTAL ASPECTS

# Enhanced Water Quality Index Prediction through Machine Learning and IoT with Hybrid Voting Methods

**Md Nadim Kaysar[a], *, Mahjabin Namira Mazumder[b], Wahidur Rahman[c], Al Shahriar Rubel[d], and Muhammad Suleman[a]**

[a] *Nagoya Institute of Technology, Nagoya, 8555 Japan*

[b] *Bangladesh University of Business and Technology, Dhaka, 1216 Bangladesh*

[c] *Uttara University, Dhaka, 1230 Bangladesh*

[d] *New Jersey Institute of Technology, Newark, 07102 USA*

**e-mail: m.kaysar.655@stn.nitech.ac.jp*

Received January 13, 2025; revised May 18, 2025; accepted May 20, 2025

**Abstract**—Water is a vital resource for our daily existence and plays a crucial role. Therefore, utilizing artificial intelligence and Internet of Things (IoT) devices can be imperative to monitor water quality efficiently. To address water quality, this study presents the development of a hybrid voting method for detecting water quality. The objectives of this study: 1) develop a hybrid voting model; 2) Analyze the ML model's memory and time complexity; 3) Integrate the best model with an IoT to monitor water quality. However, the voting method, consisting of LR, SVM, and XGBoost models, achieved an accuracy of 96%. Furthermore, we extensively examined diverse machine learning models, encompassing both instances with and without feature reduction. The analysis reveals that the models, namely XGBoost, SVM, LR, KNN, RF, and DT, exhibited accuracy rates of 94, 95, 85, 92, and 90%, respectively, without employing any feature reduction techniques. In contrast, when utilizing the PCA feature reduction technique, the accuracy percentages for the same models were 89, 92, 91, 84, 86 and 78%, respectively. Ultimately, we thoroughly examined our suggested model, considering variables such as memory complexity, time complexity, statistical error analysis, measurements based on the Internet of Things (IoT), and a comparative investigation with other research discoveries.

**Keywords:** water quality, environmental monitoring, voting method, machine learning

**DOI:** 10.1134/S0097807825700307

## INTRODUCTION

Surface water serves a vital role in all life. Water is the most significant resource for our daily lives and is essential for maintaining the existence of most living things, including humans. The water of sufficient purity is necessary for living organisms' survival. Wer has become most important for the ecology, social well-being, agriculture, and industry [16]. In some reports of the United Nations Environment Program (UNEP), In some nations, surface water provides up to 90% of the population's primary resources of drinking water. The UNEP included three significant findings in their report [16]: (1) approximately one-third of the rivers in Latin America, Africa, and Asia have been contaminated by severe pathogen pollution; (2) one-seventh of the rivers on the same continents have severe organic contamination and (3) nearly one-tenth of the rivers reported moderate to severe salinity contamination. Water is significantly crucial in the human body. Therefore, depending on the human organism, water can make up to 90% of its body weight. An adult human body typically contains around 60% water. According to Mitchell and others (1945), our brains and hearts comprise 73% water, and the lungs are about 83% water [28]. On the other hand, river water quality is an essential issue, and there are several reasons for river water pollution. It depends on location, circumstances, and other issues [16]. For instance, industrial wastewater, chemicals, and heavy metals are responsible for water pollution. Moreover, pesticides and fertilizers play a vital role in river water pollution.

Traditional quality assessment methods depend on the physical, chemical, and biological processes-based theoretical models used to detect water quality [31]. By creating mathematical equations, these models describe and forecast changes in characteristics related to water quality. These models provide predictive modeling and analysis of changes in water quality, facilitating evaluations of the state of the aquatic environment and the development of policies to enhance water quality [31]. However, traditional mechanism models have some limitations. Firstly, they required many input data and parameters, such as flow rate, rainfall, sediment properties, etc., necessitating intricate data collection and processing [18]. Secondly, developing and calibrating the model requires high practical expertise, in-depth professional knowledge, and significant data [30].

To overcome the limitations of the traditional methods, some Machine learning (ML) models have been increasingly applied to solve numerous environmental engineering problems, including river water quality modeling to address these issues. The most popular machine learning algorithms are the gradient Boosting Algorithm [6, 10, 23], support vector machine [12, 27, 36], Random Forest Algorithm [24, 25, 45], and others. Although machine learning models are frequently used for water quality analysis, they also have several drawbacks, such as a heavy reliance on training and testing data and poor forecast accuracy [26]. Moreover, some researchers come up with a deep learning-based model (DNN, LSTM CNN) used for water quality prediction [5, 9, 19, 34].

Wang et al. [32] suggest a thorough weight-based method to identify critical elements in water quality prediction by combining the Pearson correlation coefficient and entropy weighting. This method takes information content and feature correlation into account in an efficient manner. In their work, Hassan, Mehedi, et al. [8] classified a water quality dataset in various locations across India, using machine learning algorithms including RF, NN, MLR, SVM, and BTM. In their study, Kouadri, Saber, et al. [11] developed two scenarios: one using 8 models with all parameters for efficiency and another using sensitivity analysis to identify key parameters for predicting water quality index in critical cases. Neural network-based approaches such as Nonlinear autoregressive neural network (NARNET) and long short-term memory (LSTM), Multi-Layer Perceptron Neural Networks (MLP-ANN), Radial Basis Function Neural Networks (RBF-ANN), and Adaptive Neuro-Fuzzy Inference System (ANFIS) [17] is also used for better water quality prediction. They demonstrate how this research proposes to combine recurrent neural networks (RNNs) with enhanced Dempster/Shafer (D-S) evidence theory (RNNs-DS), and this approach offers strong assistance for early risk assessment and mitigation in aquatic environments [13]. Future perspectives on the viability of different water quality input variables as novel discriminant features are explored in this work. The new discriminant features make adaptive water quality criteria based on the land use activities along the river possible [33]. Some regression-based machine learning models also detect the water quality index.

Uddin, Md Galal, et al. [29] presents four machine learning-based models, including support vector machines (SVM), Naïve Bayes (NB), random forest (RF), k-nearest neighbor (KNN), and gradient boosting (XGBoost). When it came to successfully forecasting the water quality for seven WQI models, the XGBoost algorithm (99.9% correct and 0.1% incorrect) outperformed the KNN algorithm (100% correct and 0% incorrect). In this study [6], machine learning models were developed to predict Escherichia coli and enterococci levels based on environmental parameters and examined their dynamics and relationships with environmental stressors. The data from routine monitoring of these bacteria were collected from 15 public beaches in the Croatian city of Rijeka.

In this research, they propose a multi-source transfer learning-based water quality prediction approach for an IoT water environmental system so that the water quality data from neighboring monitoring stations can be efficiently utilized to increase the prediction accuracy [37]. The evaluation predicts the concentration of fecal indicator microorganisms using physico-chemical data from sensors. As the findings demonstrate, water quality monitoring in smart cities may be improved by utilizing transfer learning [3].

Despite having the above existing literature, our observation reveals a gap in research concerning nonlinear complex relationships in the ML model. Some studies exhibit nearly perfect accuracy, potentially indicating overfitting, while others identify underfitting in model training and testing. Most of the above research studies do not consider the models' memory usage, time complexity, and model utilization in real-world applications. Therefore, this research aims to fill the gap using different machine learning models with adaptive methodological architecture, hyperparameter tuning, and a hybrid voting method to asses the best model for performance. Analyze the ML algorithm's memory and time complexity. In addition, this study's aim is the best model integrated with the IoT approach for real-time water monitoring.

In this research study, we have presented a new approach to predicting water quality index using a hybrid machine learning technique with the voting approach. Our study, which included IoT-based data collection with Arduino, has the potential to impact the field of environmental science significantly. Additionally, we use the dimension reduction technique to examine the effectiveness of the machine learning model. We also thoroughly investigated the memory and time complexity of the machine-learning algorithms to make sure it was efficient, scalable, and easy to use. Our comprehensive experimental evaluations demonstrated that our proposed hybrid voting classifier strategy offers improved prediction accuracy and performs considerably better than current techniques, inspiring further research and development in this area.

However, this study has some specific objectives, which are listed below:

to develop and implement Logistic Regression, SVM, and XGBoost-based voting methods to reduce overfitting, underfitting, and increase the reliability of model performance;

to investigate various machine learning models with hyperparameter tuning, evaluating their performance with and without feature reduction techniques such as PCA and LDA;

to comprehensively evaluate our proposed model, considering various factors, including memory complexity, prediction time complexity, statistical error analysis, and measurements based on the Internet of Things (IoT);

to integrate the best-performance machine learning model with IoT for water quality monitoring.

The research study is structured into four interconnected sections. Section two presents the entire study materials and method. Section three analyzes the study performance. Finally, section four concludes the study.

## MATERIALS AND METHOD

Our proposed research has two key components. In the first phase, we developed a robust machine-learning model to predict water quality utilizing data collected through an Arduino device. In the second phase, we collected data from water resources through IoT devices. The proposed working architecture is shown below.

### *Machine Learning Model*

This second section dissects the machine learning model build architecture into several segments. First, we delve into data collection and preprocessing. Second, we meticulously examine data analysis and feature engineering, ensuring a comprehensive understanding. Additionally, in the classifier section, we explore different algorithms for this study using a hyperparameter tuning approach. Finally, the most accurate algorithms are used for the voting approach in this study.

We collected the dataset from here [8] for this research study. In the pre-processing, we used numpy and Pandas libraries to remove unnecessary columns from the dataset. Then, we checked if there were any duplicate values in the columns. After that, we checked for any missing values in every feature column. If it finds so, we convert the missing values with NumPy-none-value and fill the missing values using the median. Then, we saved the dataset correctly, and the z-score normalization method was used for outlier detection. Subsequently, we built a water quality index label column with the water quality index equations [8].

**Table 1.** Dataset feature description

| Features | Descriptions |
|---|---|
| Temp | Temperature of water |
| DO | Amount of oxygen |
| PH | Water acidity and basicity are measured |
| Conductivity | Conductivity is a measure of the ability of water to conduct an electric current |
| BOD | BOD is an important indicator of the level of organic pollution in water bodies |
| NI | Excess nitrates, but essential plant nutrition, in combination with phosphorus, can hasten eutrophication |
| Fec_Col | Fecal coliforms are a group of bacteria that are used as indicators of fecal contamination in water |
| Tot_Col | Total coliforms are a group of bacteria that are commonly used as indicators of the microbial quality of water |

**Table 2.** Water Quality Index Wise Label

| Water Quality Index (WQI) | Label |
|---|---|
| 0–25 | Excellent (3) |
| 26–50 | Good (2) |
| 51–100 | Poor (1) |
| WQI > 100 | Unsuitable for Drink (0) |

$$\mathrm{WQI} = \frac{\sum_{i=1}^{N} q_i w_i}{\sum_{i=1}^{N} w_i}. \quad (1)$$

After building the water quality index, we categorized the label based on the water quality index value (Tables 1, 2).

In this part, we have analyzed data to explore the knowledge and understand the data pattern. For this analysis, understanding the statistical distribution is needed, as described in Table 3. On the other hand, correlation analysis is done for feature importance detection. This process involves understanding the data for analysis. We have shown the correlation graph with the water quality index and WQI classifier label. In addition, only eight features are used for this study that are impactful for accurate results. Finally, we have used the StandardScaler method for feature scaling.

**Support Vector Machine.** It was the identification of the ideal hyperplane that effectively distinguishes between the classes. While it is simple to have a hyperplane in the SVM classifier among these classes [8, 29], the best hyperparameter was identified for the

**Table 3.** Statistical Information of Feature Column

| | Temp | DO | PH | Con | BOD | NI | Fec_Col | Tot_Col |
|---|---|---|---|---|---|---|---|---|
| Mean | 26.24 | 6.43 | 7.22 | 1006 | 3.96 | 1.11 | 2264.42 | 7242.59 |
| Std | 3.23 | 1.25 | 0.58 | 2764 | 7.13 | 1.62 | 8259.47 | 40230.19 |
| Min | 10 | 0.0 | 2.90 | 11 | 0.10 | 0 | 0 | 0 |
| 25% | 25 | 6 | 6.90 | 83 | 1.10 | 0.26 | 46 | 108.75 |
| 50% | 27 | 6.7 | 7.22 | 183 | 1.80 | 0.52 | 233 | 465.0 |
| 75% | 28.20 | 7.2 | 7.60 | 489.25 | 3.40 | 1.10 | 672.5 | 1650.0 |
| Max | 30 | 10 | 9.01 | 18569 | 88 | 13.20 | 150250 | 967500 |

training using GridSearchCV before the implementation of SVM. This process helps us build robust and accurate models, balancing bias and variance and preventing the model from overfitting or underfitting. The RBF (Radial Basis Function) kernel is better at achieving comparative accuracy with another kernel. We used $C = 1000$ for a smaller margin, which helped avoid misclassifying each training sample. The $C$ value was selected for SVM after searching within the range of [400–1000] because it provided the best accuracy on the validation set. The GridSearchCV was used with a 5-fold cross-validation strategy to systematically explore the hyperparameter space. In addition, Gamma = 0.01 has been used for the closest point decision boundary with hyperplan in SVM algorithm. Finally, the other parameters are used as a default value, following standard practices in SVM implementation. The all hyperparameter values are stated in Table 4.

**XGBoost.** XGBoost is a distributed gradient boosting library optimized for efficiency and scalability in machine learning model training. This ensemble learning technique generates more robust predictions by combining the predictions of several weak models. It can also be applied to address classification and regression difficulties. Researchers want XGBoost because it executes quickly out of core computation [29]. The XGBoost works as follows: Let DS be a dataset of m features and $n$ samples.

$$\mathrm{DS} = \{(x_i, y_i);\ i = 1\ldots n,\quad x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}. \tag{3}$$

Let $y_i$ represent the expected result of an ensemble tree model obtained using Eq. (4).

**Table 4.** Hyperparameter for SVM Model

| Parameter | Value |
|---|---|
| C | 1000 |
| Gamma | 0.01 |
| Kernel | rbf |
| Others Parameter | Default |

$$A_i = \phi(x_i) = \sum_k^K f_k(x_i). \tag{4}$$

To answer the above Equation, we must identify the optimal set of functions by minimizing the loss and regulation, where $k$ is the number of trees in the model and $f_k$ is the ($k$th) tree. On the other hand, we have used specific parameters for tuning the model, such as booster, device, verbosity, and others. The booster parameter value = gbtree for XGBoost was selected based on its superior performance on the validation set compared to other booster types. This study set validate_parameters = false to allow flexibility during experimentation. The nthread parameter was left unset, allowing XGBoost to utilize the maximum number of available threads. The disable_default_eval_metric was set to false so the default evaluation metric would be used. All other parameters were kept at their default values. All the hyperparameters are stated in Table 5.

**K-Nearest Neighbor.** This KNN algorithm measures the distances between a given data point in the set and every other K number of data points in the dataset close to the initial point. Different distances are typically used to calculate distances. Therefore, the final model is just the labeled data arranged in a certain way [29]. KNN reduces overfitting, and selecting the optimal value for K is crucial to ensure its effectiveness. The n_neighbors parameter value = 5 for KNN was selected after searching within the range [2–6], as it provided the best accuracy on the testing set, along with weights = ['uniform','distance'], metric = ['minkowski','euclidean','manhattan'] and 3 cross validation. All the hyperparameters are stated in Table 6.

**Logistic Regression.** The relationship between a set of independent (explanatory) factors and a categorical dependent variable is known as the logistic regression model, and it is used to calculate the likelihood that an event will occur. The mean of a continuous dependent variable in multiple regression is predicted using a mathematical model of a group of explanatory factors. Equation (5) demonstrates that the logit transformation is expressed mathematically, where p is the probability and the associated odds. The value of

C = np.logspace(−4, 4, 20) for Logistic Regression was selected along with penalty = ['l1', 'l2', 'elasticnet', 'none'], solver = ['lbfgs','newton-cg','liblinear','sag','saga'], and Max_iter = [100, 1000, 2500, 5000] as this combination provided the best accuracy on the testing set. Other parameters were kept at their default values. Nonetheless, Table 7 displays the LR values utilized in the code for fine-tuning.

$$l = \mathrm{logit}\left(p\right) = \ln\left(\frac{p}{1-p}\right). \quad (5)$$

**Decision Tree.** The decision tree (DT) algorithm, a top choice in data mining, is characterized by its top-down divide-and-conquer approach. At its core is the Greedy algorithm, which drives the recursive top-down division. To enhance its decision tree algorithm performance, we utilized the Table 8 parameter.

**Random Forest.** The random forest (RF) technique builds an ensemble of decision trees in training by switching and altering the covariates to enhance prediction performance [29]. We fine-tuned the model using Table 9 parameters to enhance the performance of this algorithm. To enhance the performance, the n_estimators parameter value range = [5, 10, 25, 50, 100], Max_feature = ['sqrt','log2'], Max_depth = [50, 100], Max_leaf_nodes = [50, 100], and Others Parameter = default. The RF prediction process also involves storing the expected result, or target, and then analyzing the votes for each possible target.

**Hybrid Voting Method.** In this method, we have used the top three voting models: support vector machines, logistic regression, and the XGBoost algorithm. The findings of the SVM, LR, and XGBoost models are inputted into our hybrid voting model and predicted based on the average probability given to that class. For this voting approach, the soft voting method is used in this study. The voting method gained experience for classification from SVM, LR, and XGBoost base models. On the other hand, the final determination used the majority vote of the base model.

### *IoT-based Data Collection*

This part is responsible for IoT-based data collection for water quality prediction. The core part of this scheme is Arduino, which acts as the center of the whole IoT orientation. A set of water-related sensors connects to the microcontroller, collecting raw data in real time and relaying it back to the microcontroller. Then, the microcontroller moves on to the data transmission via the Global System for Mobile Communications (GSM) module to build a remote connection between the microcontroller and the cloud server. The server is responsible for real-time water monitoring and efficient data storage. In this server, we have deployed our voting model for water quality prediction and monitoring.

**Table 5.** Hyperparameter for XGBoost Model

| Parameter | Value |
|---|---|
| Booster | gptree |
| Device | cpu |
| Verbosity | 1 |
| Validate_parameters | False |
| nthread | Default to maximum number of threads available if not set |
| Disable_default_eval_metric | False |
| Others Parameter | Default |

**Table 6.** Hyperparameter for KNN Model

| Parameter | Value |
|---|---|
| n_neighbors | [2–6] |
| Weights | ['uniform','distance'] |
| metrics | ['minkowski','euclidean','manhattan'] |
| Verbose | 1 |
| cv | 3 |
| n_jobs | –1 |

**Table 7.** Hyperparameter for LR Model

| Parameter | Value |
|---|---|
| Penalty | ['l1', 'l2', 'elasticnet', 'none'] |
| C | np.logspace(–4, 4, 20) |
| Solver | ['lbfgs','newton-cg','liblinear','sag','saga'] |
| Max_iter | [100, 1000, 2500, 5000] |
| Others Parameter | default |

**Table 8.** Hyperparameter for DT Model

| Parameter | Value |
|---|---|
| Max_depth | [2, 3, 5, 10, 20] |
| min_samples_Leaf | [5, 10, 20, 50, 100] |
| Criterion | ['gini', 'entropy'] |
| cv | 4 |
| Verbose | 1 |
| n_jobs | –1 |
| Others Parameter | default |

## PERFORMANCE ANALYSIS

We organize the performance analysis into three sections, each dedicated to machine learning performance metrics. In the first section, we analyze different error-based metrics, including means square error,

**Table 9.** Hyperparameter for RF Model

| Parameter | Value |
|---|---|
| n_estimators | [5, 10, 25, 50, 100] |
| Max_feature | ['sqrt','log2'] |
| Max_depth | [50, 100] |
| Max_leaf_nodes | [50, 100] |
| Others Parameter | default |

root means square error, and mean absolute error. The second section discusses the precision, recall, F1 scores, Confusion Matrix, and precision-recall curve. Again, the ROC-AUC curve will be discussed in the third section of this performance analysis part. In addition, the time and space complexity analysis is in the fourth section. Finally, we compare our study with other research studies for a comprehensive understanding. Through this analysis, we tried to measure the ML algorithm's capability for water index prediction for this dataset.

*Statistical Error Analysis*

In this section, we evaluate the models using some statistical indices, including root means square error (RMSE), means square error (MSE), and mean absolute error (MAE). We observed that the voting classifier model achieved the lowest error rate, as shown in Table 10. The RMSE represents the sample standard deviation of the discrepancies between computed and observed values. Additionally, the MAE provides another measurement of the discrepancy between computed and observed values.

*Precision, Recall, F1 Scores and Confusion Matrix*

**Precision** is the ratio of the total number of expected positive observations to the accurately anticipated positive observations. The low false-positive rate is achieved through excellent accuracy. Equation (6) illustrates accuracy:

$$\text{Precision} = \frac{\text{True Positives}}{(\text{True positives} + \text{False Positives})} = \frac{TP}{(TP + FP)}. \quad (6)$$

**Recall**, also known as the true positive rate (TPR), represents the percentage of data samples correctly classified as belonging to the positive class out of all the samples for that class.

Modifying different parameters or hyperparameters can affect the recall value of a machine learning model, either raising or lowering it by adjusting these factors.

$$\text{Recall} = \frac{\text{True Positives}}{(\text{True Positives} + \text{False Negatives})} = \frac{TP}{(TP + FN)}. \quad (7)$$

**The F1 score** provides an overall assessment of a model's accuracy by combining precision and recall, similar to how addition and multiplication combine two ingredients to create a whole new dish. In other words, having a low false positive and false negative score indicates correctly identifying serious threats and not being bothered by false alarms.

The precision, recall, and F1 score values are below in Tables 11–13. Table 11 shows the result without dimension reduction, Table 12 shows after dimension reduction using PCA, and Table 13 shows LDA dimension reduction-based accuracy. In PCA implementation, we have achieved higher accuracy using 7 dimensions for model training and testing and achieved higher accuracy utilizing 4 dimensions used for the LDA dataset. Finally, we found that LDA and PCA-based feature reduction models demonstrated limited effectiveness in the context of this research study. The hybrid voting model was built using three effective models: SVM, Logistic Regression, and XGBoost.

In Fig. 1, we present the confusion matrices for the top performance six models in Fig. 1. This confusion matrix is used for error analysis of ML algorithms. Additionally, Fig. 2 displays the precision-recall

**Table 10.** Statistical Error Evaluation

| Without Dimension Reduction | | | | With PCA | | | With LDA | | |
|---|---|---|---|---|---|---|---|---|---|
| models | MSE | RMSE | MAE | MSE | RMSE | MAE | MSE | RMSE | MAE |
| XGBoost | 0.07 | 0.26 | 0.06 | 0.10 | 0.32 | 0.10 | 0.31 | 0.55 | 0.27 |
| SVM | 0.05 | 0.22 | 0.05 | 0.08 | 0.28 | 0.08 | 0.89 | 0.94 | 0.46 |
| KNN | 0.16 | 0.40 | 0.15 | 0.17 | 0.41 | 0.16 | 0.37 | 0.61 | 0.33 |
| Random Forest | 0.11 | 0.33 | 0.09 | 0.15 | 0.39 | 0.14 | 0.33 | 0.57 | 0.29 |
| Decision Tree | 0.14 | 0.38 | 0.11 | 0.22 | 0.47 | 0.22 | 0.41 | 0.64 | 0.31 |
| Logistic Regression | 0.04 | 0.21 | 0.04 | 0.10 | 0.31 | 0.09 | 1.18 | 1.08 | 0.57 |
| Voting Classifier | 0.04 | 0.20 | 0.04 | 0.08 | 0.29 | 0.07 | 0.28 | 0.53 | 0.24 |

**Table 11.** Models Precision, Recall, F1 Scores

| Models | Label | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|---|
| XGBoost | Excellent (3) | 1.00 | 1.00 | 1.00 | 0.94 |
| | Good (2) | 0.96 | 0.92 | 0.94 | |
| | Poor (1) | 0.94 | 0.97 | 0.95 | |
| | Unsuitable (0) | 0.93 | 0.88 | 0.90 | |
| SVM | Excellent (3) | 1.00 | 1.00 | 1.00 | 0.95 |
| | Good (2) | 1.00 | 0.96 | 0.98 | |
| | Poor (1) | 0.91 | 1.00 | 0.95 | |
| | Unsuitable (0) | 1.00 | 0.81 | 0.90 | |
| KNN | Excellent (3) | 1.00 | 1.00 | 1.00 | 0.85 |
| | Good (2) | 0.76 | 0.50 | 0.60 | |
| | Poor (1) | 0.82 | 0.94 | 0.88 | |
| | Unsuitable (0) | 0.93 | 0.79 | 0.85 | |
| Random Forest | Excellent (3) | 1.00 | 1.00 | 0.90 | 0.92 |
| | Good (2) | 0.77 | 0.89 | 0.94 | |
| | Poor (1) | 0.99 | 0.87 | 0.82 | |
| | Unsuitable (0) | 0.81 | 1.00 | 1.00 | |
| Decision Tree | Excellent (3) | 1.00 | 1.00 | 1.00 | 0.90 |
| | Good (2) | 0.85 | 0.69 | 0.76 | |
| | Poor (1) | 0.92 | 0.92 | 0.92 | |
| | Unsuitable (0) | 0.85 | 0.95 | 0.90 | |
| Logistic Regression | Excellent (3) | 1.00 | 0.94 | 0.97 | 0.95 |
| | Good (2) | 0.96 | 1.00 | 0.98 | |
| | Poor (1) | 0.99 | 0.94 | 0.96 | |
| | Unsuitable (0) | 0.85 | 0.98 | 0.91 | |
| Voting Classifier | Excellent (3) | 1.00 | 1.00 | 1.00 | 0.96 |
| | Good (2) | 1.00 | 0.96 | 0.98 | |
| | Poor (1) | 0.93 | 1.00 | 0.96 | |
| | Unsuitable (0) | 1.00 | 0.85 | 0.92 | |

curves for these models. Moreover, it is important to note that all the confusion matrix and precision-recall curves depicted in the figures are without feature reduction.

### *ROC-AUC Curve*

In machine learning, the prediction strength of a classification model is evaluated using the ROC-AUC curve, especially in scenarios where the class distribution is imbalanced or where there is a need to fine-tune the model's sensitivity and specificity. To investigate it, Fig. 3 shows the ROC-AUC curve for the Voting Classifier, Logistic Regression, SVM, XGBoost, Random Forest, and Decision Tree. We obtained comparatively better performance in the Voting classifier, SVM, and Logistic regression models than in others without using dimension reduction. The AUC values for unsuitable class water are 0.97 in the Voting and SVM models, while the Logistic Regression model achieves an AUC of 0.95 for the same class. The models, including Voting, SVM, and Logistic Regression, for poor, good, and excellent water classes achieved AUC values of 1.00, which shows the model's capability for water quality prediction.

### *Time and Memory Complexity Analysis*

This analysis assessed the memory and time complexity of the machine learning models for training and prediction. One thousand tabular augmented data were generated using tbGAN for the prediction time analysis. After that, we calculated memory using the trained model and recorded the prediction time for this new augmented dataset, as detailed in Table 14. Moreover, we found that the DT model required only

**Table 12.** Models Precision, Recall, F1 Scores After PCA Implementation

| PCA Implemented Accuracy | | | | | |
|---|---|---|---|---|---|
| models | label | precision | recall | F1 | accuracy |
| XGBoost | Excellent (3) | 1.00 | 1.00 | 1.00 | 0.89 |
| | Good (2) | 0.81 | 0.76 | 0.79 | |
| | Poor (1) | 0.92 | 0.89 | 0.91 | |
| | Unsuitable (0) | 0.83 | 0.95 | 0.87 | |
| SVM | Excellent (3) | 1.00 | 1.00 | 1.00 | 0.92 |
| | Good (2) | 0.89 | 0.80 | 0.84 | |
| | Poor (1) | 0.95 | 0.91 | 0.93 | |
| | Unsuitable (0) | 0.83 | 1.00 | 0.91 | |
| KNN | Excellent (3) | 1.00 | 1.00 | 1.00 | 0.84 |
| | Good (2) | 0.66 | 0.70 | 0.68 | |
| | Poor (1) | 0.83 | 0.89 | 0.86 | |
| | Unsuitable (0) | 0.97 | 0.75 | 0.85 | |
| Random Forest | Excellent (3) | 1.00 | 0.94 | 0.97 | 0.86 |
| | Good (2) | 0.63 | 0.74 | 0.68 | |
| | Poor (1) | 0.93 | 0.84 | 0.88 | |
| | Unsuitable (0) | 0.79 | 0.97 | 0.87 | |
| Decision Tree | Excellent (3) | 1.00 | 1.00 | 1.00 | 0.78 |
| | Good (2) | 0.60 | 0.69 | 0.64 | |
| | Poor (1) | 0.83 | 0.77 | 0.80 | |
| | Unsuitable (0) | 0.68 | 0.75 | 0.71 | |
| Logistic Regression | Excellent (3) | 1.00 | 1.00 | 1.00 | 0.91 |
| | Good (2) | 0.87 | 0.84 | 0.85 | |
| | Poor (1) | 0.95 | 0.90 | 0.93 | |
| | Unsuitable (0) | 0.82 | 0.97 | 0.89 | |
| Voting Classifier | Excellent (3) | 1.00 | 1.00 | 1.00 | 0.93 |
| | Good (2) | 0.96 | 0.83 | 0.89 | |
| | Poor (1) | 0.89 | 1.00 | 0.94 | |
| | Unsuitable (0) | 1.00 | 0.80 | 0.89 | |

9 KB, and for PCA and LDA, it needed 11 KB memory, which is lower than that of other algorithms. For the prediction time calculation, the SVM model performs more satisfactorily than other algorithms.

### *Comparative Analysis*

Table 15 compares different existing studies. The aim of this comparative analysis: validate the proposed method and compare it with the existing methods and their applicability. The waterbody column shows the water resource type, and the model's column states the machine learning and deep learning algorithms. Finally, the proposed voting classifier with XGBoost, SVM, and LR algorithm achieved 96% accuracy, surpassing the performance of most other studies. On the other hand, [2, 4] showed equal accuracy with our proposed voting method.

### *IoT-Based Analysis*

This subsection provides the experimental data analysis and calculation with the IoT-based orientation. We primarily focus on the data collection through the ten (0) samples in real-time. The most suitable features for our analysis are the Potential of Hydrogen (pH), Dissolved oxygen (DO), Chemical Oxygen Demand (COD), Nitrogen Peak values, Temperature, and Ammonia. We have collected these samples from different regions in Bangladesh, including various water resources like ponds, canals, drains, and rivers, measured the data with the sensors, and

**Table 13.** Models Precision, Recall, F1 Scores after LDA Implementation

| LDA Implemented Accuracy | | | | | |
|---|---|---|---|---|---|
| models | label | precision | recall | F1 | accuracy |
| XGBoost | Excellent (3) | 1.00 | 1.00 | 1.00 | 0.75 |
| | Good (2) | 0.50 | 0.43 | 0.46 | |
| | Poor (1) | 0.80 | 0.79 | 0.80 | |
| | Unsuitable (0) | 0.57 | 0.67 | 0.62 | |
| SVM | Excellent (3) | 0.69 | 0.60 | 0.64 | 0.69 |
| | Good (2) | 0.00 | 0.00 | 0.00 | |
| | Poor (1) | 0.99 | 0.69 | 0.81 | |
| | Unsuitable (0) | 0.20 | 1.00 | 0.33 | |
| KNN | Excellent (3) | 0.96 | 1.00 | 0.98 | 0.68 |
| | Good (2) | 0.28 | 0.21 | 0.24 | |
| | Poor (1) | 0.71 | 0.77 | 0.74 | |
| | Unsuitable (0) | 0.55 | 0.49 | 0.52 | |
| Random Forest | Excellent (3) | 1.00 | 0.96 | 0.98 | 0.72 |
| | Good (2) | 0.33 | 0.44 | 0.38 | |
| | Poor (1) | 0.78 | 0.76 | 0.77 | |
| | Unsuitable (0) | 0.60 | 0.58 | 0.59 | |
| Decision Tree | Excellent (3) | 1.00 | 0.87 | 0.93 | 0.72 |
| | Good (2) | 0.42 | 0.45 | 0.43 | |
| | Poor (1) | 0.77 | 0.77 | 0.77 | |
| | Unsuitable (0) | 0.57 | 0.61 | 0.59 | |
| Logistic Regression | Excellent (3) | 0.46 | 0.50 | 0.48 | 0.65 |
| | Good (2) | 0.00 | 0.00 | 0.00 | |
| | Poor (1) | 0.96 | 0.69 | 0.80 | |
| | Unsuitable (0) | 0.23 | 0.47 | 0.31 | |
| Voting Classifier | Excellent (3) | 0.96 | 1.00 | 0.98 | 0.77 |
| | Good (2) | 0.50 | 0.42 | 0.45 | |
| | Poor (1) | 0.78 | 0.85 | 0.81 | |
| | Unsuitable (0) | 0.72 | 0.60 | 0.66 | |

verified the data in the wet lab. The experimental data clearly shows a pH, DO, and COD fluctuation. On the other hand, the temperature level for the sample water is almost the same, but we found the highest temperature to be 28.52°C. Table 16 summarizes the results, and Fig. 4 shows the experimental data analysis using the bar chart representations.

## CONCLUSIONS

All living organisms depend on surface water. The most important resource for our daily necessities is water, which is also necessary for the survival of the majority of living creatures, including humans. To survive and live, pure and sufficient water is important in our daily life. This research focuses on detecting water quality utilizing the water quality index and machine learning algorithms. This research aimed to investigate the machine learning algorithm's performance in these datasets and analyze all algorithm's memory and time complexity. In addition, the hybrid voting method is applied to measure the performance. For this investigation, we used two types of datasets; firstly, we applied SVM, Logistic Regression, XGBoost, Random Forest, Decision Tree, and KNN algorithms without feature reduction. Subsequently, the same machine-learning algorithms were applied after feature reduction. Finally, we have examined different ways for model evaluation, such as statistical evaluation, precision, recall, F1 score, and the AUC-ROC curve. We have found three maximum accuracies for the dataset without feature reduction. The


(a) Voting Classifier (b) Logistic Regression (c) SVM (d) XGBoost (e) Random Forest (f) Decision Tree

**Fig. 1.** Confusion Matrix of XGBoost, SVM, RF, DT, LR, Voting Classifier.

highest accuracy models were SVM at 95%, Logistic Regression at 95%, and XGBoost at 94%. Subsequently, we have used a voting classifier with these three models with a soft voting approach and obtained 96% accuracy.

The proposed research study demonstrates higher accuracy compared to other research studies. In the future, we aim to expand our work by incorporating more datasets for training models and proper model selection. In addition, we plan to extend our efforts to

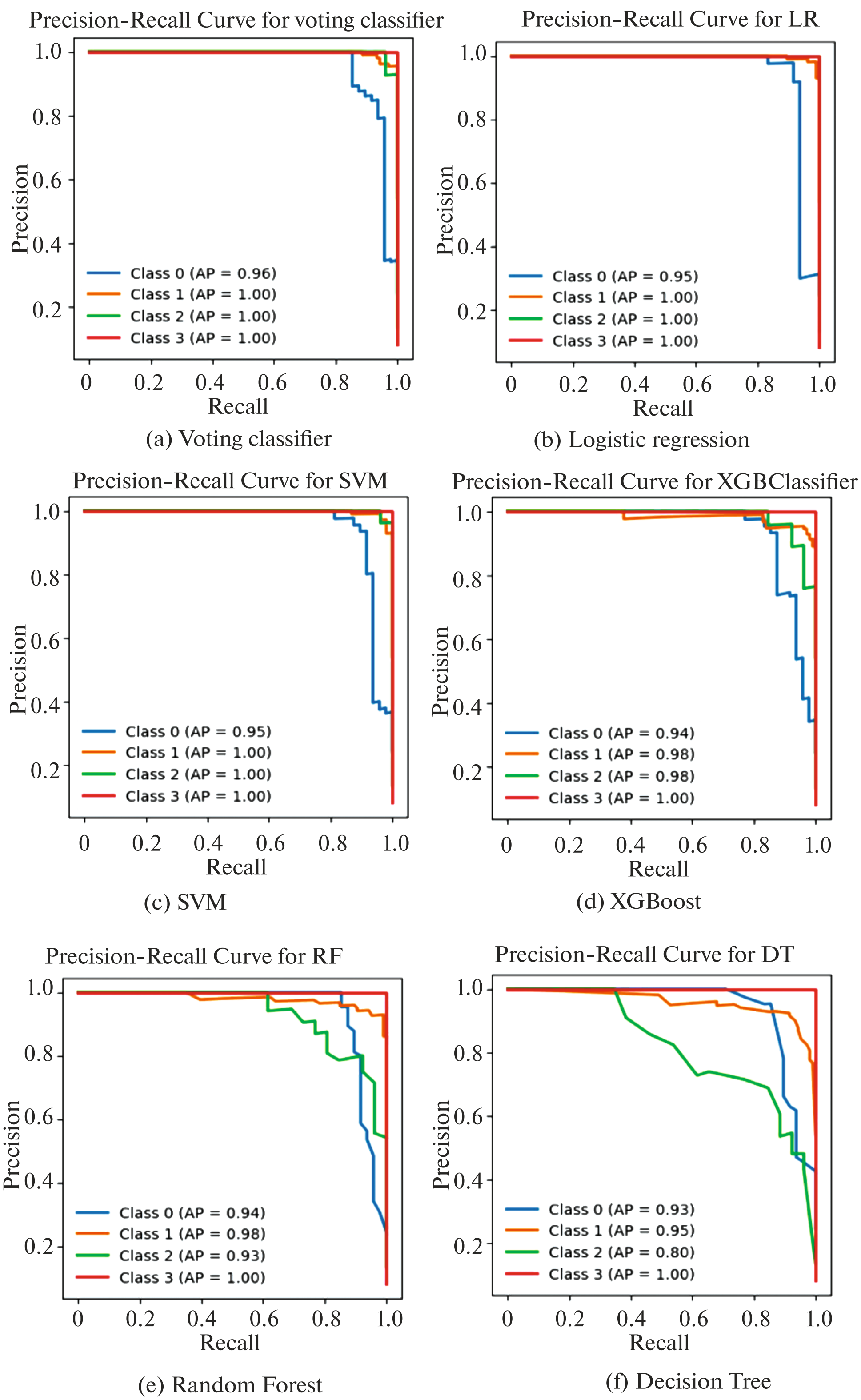


**Fig. 2.** Precision Curve of XGBoost, SVM, RF, DT, LR, Voting Classifier.

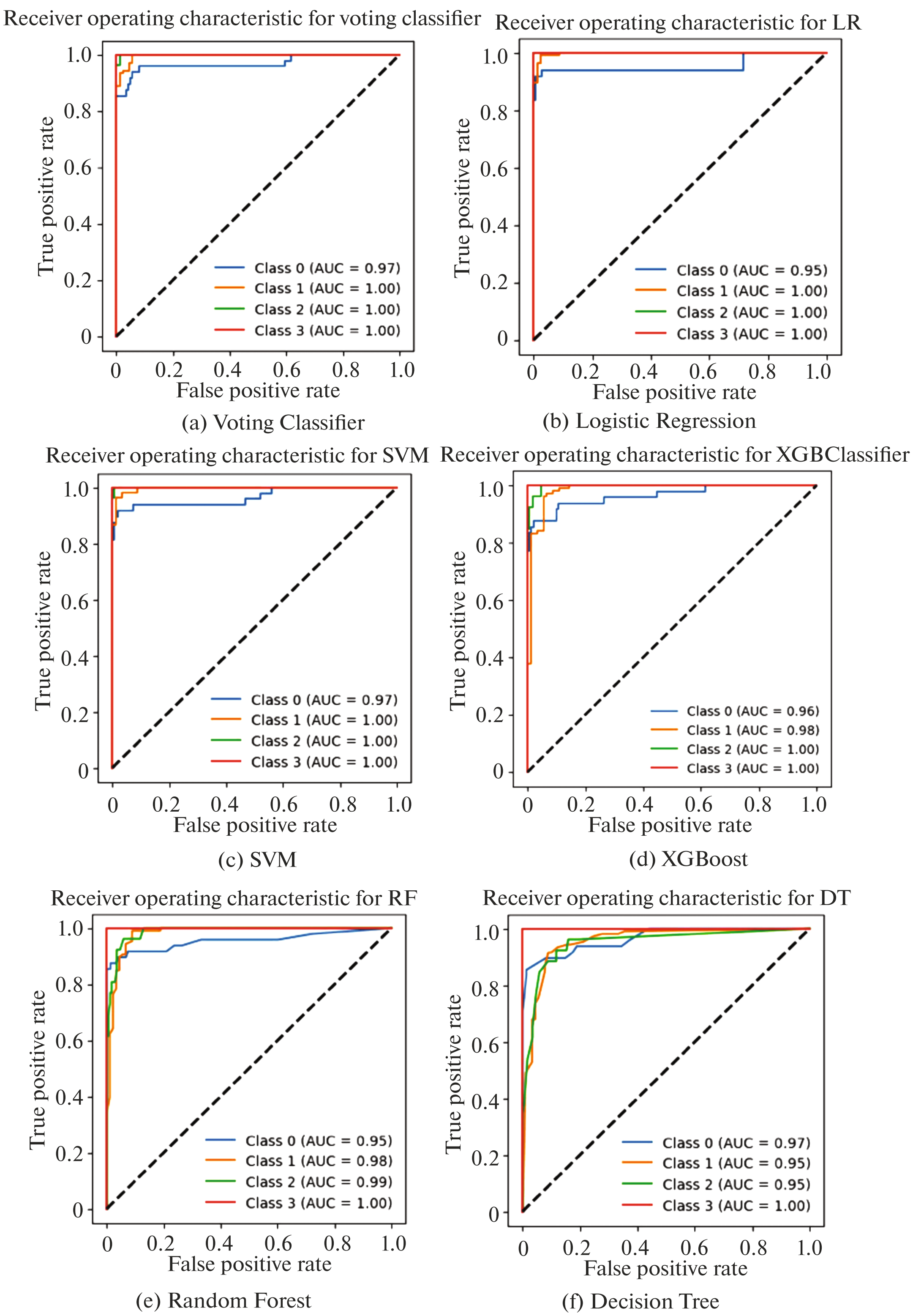


**Fig. 3.** ROC-AUC Curve of XGBoost, SVM, RF, DT, LR, Voting Classifier.

other water locations with different deep-learning models. Moreover, the bayesian optimization can be use at future for hyperparameter tuning in future advancement this study. This method allows for more precise selection of hyperparameters compared to grid or random search. These future works are considered

**Table 14.** Time and Space Analysis

| Models | Without Dimension Reduction | | PCA | | LDA | |
|---|---|---|---|---|---|---|
| | time | memory | time | memory | time | memory |
| XGBoost | 4 μs | 604 KB | 3 μs | 744 KB | 5 μs | 681 KB |
| SVM | 3 μs | 41 KB | 3 μs | 38 KB | 4 μs | 72 KB |
| K-Nearest Neighbors | 4 μs | 156 KB | 5 μs | 126 KB | 4 μs | 97 KB |
| Random Forest | 4 μs | 2.5 MB | 5 μs | 3.4 MB | 5 μs | 3.4 MB |
| Decision Tree | 4 μs | 9 KB | 4 μs | 11 KB | 5 μs | 11 KB |
| Logistic Regression | 5 μs | 174 KB | 5 μs | 174 KB | 5 μs | 174 KB |
| Voting Classifier | 4 μs | 1.6 MB | 4 μs | 1.7 MB | 5 μs | 8 MB |

**Table 15.** Comparison with existing Models

| Citation | Waterbody | Models | Performance, % |
|---|---|---|---|
| [21] | Mixed | SVM, KNN, NN, and Deep NN | 93 |
| [14] | Mixed | LSTM | 94 |
| [15] | Wastewater | ANN | 79 |
| [2] | Rivers and lakes | ANFIS and FFNN | 96 |
| [22] | River | ANFIS (EC and TDS) | 9 and 92 |
| [7] | Coastal waters | ANN for Chl-a data | 89 |
| [4] | River | ANN | 99 |
| [1] | Lake | MLP | 85 |
| Proposed | River | Voting Classifier with (XGBoost, SVM, LR) | 96 |

to be limitations in this study. However, despite those limitations, this study achieved the objectives and effectively monitored water quality.

## AUTHOR CONTRIBUTION

The conceptualization, Methodology, implementation, and drafting by all authors. Finally, all authors have reviewed and agreed to the published version of the manuscript.

## FUNDING

This work was supported by ongoing institutional funding. No additional grants to carry out or direct this particular research were obtained.

**Table 16.** Experimental data analysis with IoT-based setup and mean values from the number of samples

| Serial | No. of Sample | pH (1.00–14.00) | DO, mg/L | COD, mg/L | $NO_3$, mg/L | Temperature, °C | $NH_4$, mg/L |
|---|---|---|---|---|---|---|---|
| 1 | 12 | 7.53 | 7.24 | 70.06 | 0.14 | 27.54 | 0.29 |
| 2 | 4 | 8.18 | 7.23 | 27.84 | 0.02 | 28.67 | 0.47 |
| 3 | 5 | 8.24 | 8.63 | 27.82 | 0.14 | 25.88 | 0.51 |
| 4 | 11 | 7.79 | 7.17 | 46.91 | 0.05 | 24.08 | 0.23 |
| 5 | 5 | 8.22 | 8.06 | 50.19 | 0.15 | 25.13 | 0.14 |
| 6 | 4 | 7.85 | 7.04 | 17.07 | 0.37 | 28.18 | 0.27 |
| 7 | 4 | 8.06 | 7.28 | 17.06 | 0.09 | 25.77 | 0.34 |
| 8 | 13 | 7.66 | 7.68 | 70.23 | 0.13 | 25.74 | 0.50 |
| 9 | 5 | 8.10 | 8.48 | 27.52 | 0.03 | 28.52 | 0.27 |
| 10 | 5 | 8.19 | 8.21 | 27.42 | 0.05 | 28.34 | 0.36 |

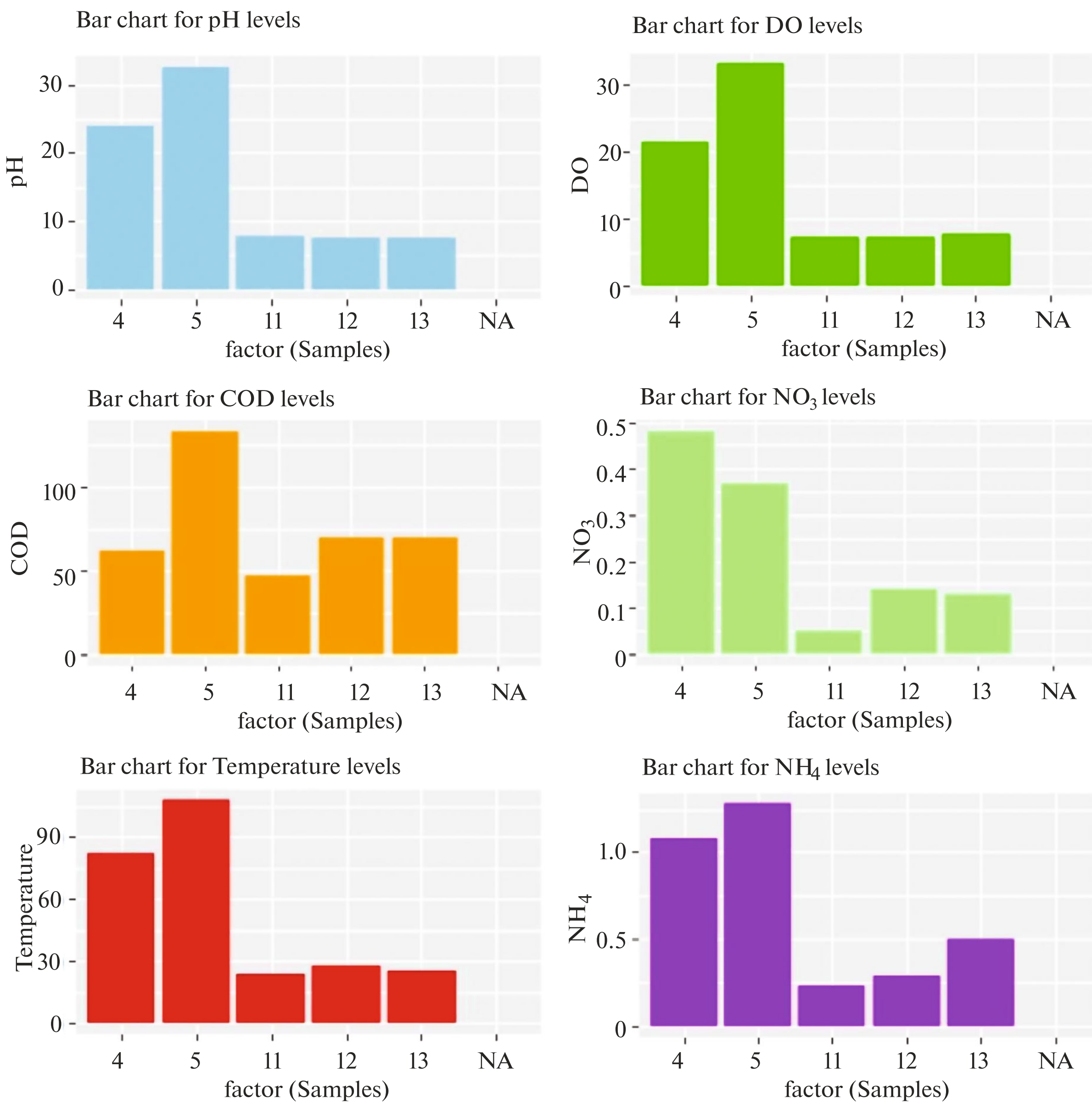


**Fig. 4.** Bar chart of the experimental data analysis with a set of parameters along with the fluctuating values.

## CONFLICT OF INTEREST

The authors of this work declare that they have no conflicts of interest.

## REFERENCES

1. Ahmed, U., Mumtaz, R., Anwar, H., Shah, A.A., Irfan, R., and García-Nieto, J., Efficient water quality prediction using supervised machine learning, *Water (Switzerland)*, 2019, vol. 11. https://doi.org/10.3390/w11112210

2. Al-Adhaileh, M.H. and Alsaade, F.W., Modelling and prediction of water quality by using artificial intelligence, *Sustainability (Switzerland)*, 2021, vol. 13. https://doi.org/10.3390/su13084259

3. Balachandran, T., Abreu, T., Naloufi, M., Souihi, S., Lucas, F., and Janne, A., IoT and transfer learning based urban river quality prediction, In: 2022 IEEE Global Communications Conference, GLOBECOM 2022—Proc., pp. 257–262. Institute of Electrical and Electronics Engineers Inc, 2022. https://doi.org/10.1109/GLOBECOM48099.2022.10001249

4. Barzegar, R., Adamowski, J., and Moghaddam, A.A., Application of wavelet-artificial intelligence hybrid models for water quality prediction: a case study in Aji-Chay River, Iran, *Stochastic Environ. Res. Risk Assess.*, 2016, vol. 30, pp. 1797–1819. https://doi.org/10.1007/s00477-016-1213-y

5. Cai, H., Zhang, C., Xu, J., Wang, F., Xiao, L., Huang, S., and Zhang, Y., Water quality prediction

based on the KF-LSTM encoder-decoder network: a case study with missing data collection, *Water (Switzerland)*, 2023, vol. 15.
https://doi.org/10.3390/w15142542

6. Grbčić, L., Družeta, S., Mauša, G., Lipić, T., Lušić, D.V., Alvir, M., Lučin, I., Sikirica, A., Davidović, D., Travaš, V., Kalafatović, D., Pikelj, K., Fajković, H., Holjević, T., and Kranjčević, L., Coastal water quality prediction based on machine learning with feature interpretation and spatio-temporal analysis, 2021.
https://doi.org/10.1016/j.envsoft.2022.105458
7. Hafeez, S., Wong, M.S., Ho, H.C., Nazeer, M., Nichol, J., Abbas, S., Tang, D., Lee, K.H., and Pun, L., Comparison of machine learning algorithms for retrieval of water quality indicators in case-ii waters: A case study of Hong Kong, *Remote Sens (Basel)*, 2019, vol. 11.
https://doi.org/10.3390/rs11060617
8. Hassan, Md.M., Hassan, Md.M., Akter, L., Rahman, Md.M., Zaman, S., Hasib, K.Md., Jahan, N., Smrity, R.N., Farhana, J., Raihan, M., and Mollick, S., Efficient prediction of Water Quality Index (WQI) using machine learning algorithms, *Human-Centric Intelligent Systems*, 2021, vol. 1, no. 86.
https://doi.org/10.2991/hcis.k.211203.001
9. Hu, Z., Zhang, Y., Zhao, Y., Xie, M., Zhong, J., Tu, Z., and Liu, J., A water quality prediction method based on the deep LSTM network considering correlation in smart mariculture, *Sensors*, 2019, vol. 19.
https://doi.org/10.3390/s19061420
10. Islam Khan, M.S., Islam, N., Uddin, J., Islam, S., and Nasir, M.K., Water quality prediction and classification based on principal component regression and gradient boosting classifier approach, *J. King Saud University—Computer and Information Sci.*, 2022, vol. 34, pp. 4773–4781.
https://doi.org/10.1016/j.jksuci.2021.06.003
11. Kouadri, S., Elbeltagi, A., Islam, A.R.M.T., and Kateb, S., Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast), *Appl. Water Sci.*, 2021, vol. 11.
https://doi.org/10.1007/s13201-021-01528-9
12. Leong, W.C., Bahadori, A., Zhang, J., and Ahmad, Z., Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM), *Int. J. River Basin Manag.*, 2021, vol. 19, pp. 149–156.
https://doi.org/10.1080/15715124.2019.1628030
13. Li, L., Jiang, P., Xu, H., Lin, G., Guo, D., and Wu, H., Water quality prediction based on recurrent neural network and improved evidence theory: a case study of Qiantang River, China, *Environ. Sci. Pollut. Res.*, 2019, vol. 26, pp. 19879–19896.
https://doi.org/10.1007/s11356-019-05116-y
14. Liu, J., Yu, C., Hu, Z., Zhao, Y., Bai, Y., Xie, M., and Luo, J., Accurate prediction scheme of water quality in smart mariculture with Deep Bi-S-SRU learning network, *IEEE Access*, 2020, vol. 8, pp. 24784–24798.
https://doi.org/10.1109/ACCESS.2020.2971253
15. Miklos, A.C., Li, C., Sorrell, C.D., Lyon, L.A., and Pielak, G.J., Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters, *BMC Biophys.*, 2011, vol. 4.
https://doi.org/10.1186/2046-1682-4-13
16. Mohd Zebaral Hoque, J., Nor, N.A., Alelyani, S., Mohana, M., and Hosain, M., Improving water quality index prediction using regression learning models, *Int. J. Environ. Res. Public Health*, 2022, vol. 19.
https://doi.org/10.3390/ijerph192013702
17. Najah Ahmed, A., Binti Othman, F., Abdulmohsin Afan, H., Khaleel Ibrahim, R., Ming Fai, C., Shabbir Hossain, M., Ehteram, M., and Elshafie, A., Machine learning methods for better water quality prediction, *J. Hydrol.*, 2019, vol. 578.
https://doi.org/10.1016/j.jhydrol.2019.124084
18. Rand, J.M., Nanko, M.O., Lykkegaard, M.B., Wain, D., King, W., Bryant, L.D., and Hunter, A., The human factor: Weather bias in manual lake water quality monitoring, *Limnol. Oceanogr. Methods.*, 2022, vol. 20, pp. 288–303.
https://doi.org/10.1002/lom3.10488
19. Rustam, F., Ishaq, A., Kokab, S.T., de la Torre Diez, I., Mazón, J.L.V., Rodríguez, C.L., and Ashraf, I., An artificial neural network model for water quality and water consumption prediction, *Water*, 2022, vol. 14.
https://doi.org/10.3390/w14213359
20. Sakizadeh, M., Artificial intelligence for the prediction of water quality index in groundwater systems, *Model Earth Syst. Environ.,* 2016, vol. 2.
https://doi.org/10.1007/s40808-015-0063-9
21. Shafi, U., Mumtaz, R., Anwar, H., Qamar, A.M., and Khurshid, H., Surface water pollution detection using internet of things, 2018.
https://doi.org/10.1109/HONET.2018.8551341
22. Shah, M.I., Abunama, T., Javed, M.F., Bux, F., Aldrees, A., Rehman Tariq, M.A.U., and Mosavi, A., Modeling surface water quality using the adaptive neuro-fuzzy inference system aided by input optimization, *Sustain.*, 2021, vol. 13.
https://doi.org/10.3390/su13084576
23. Shams, M.Y., Elshewey, A.M., El-kenawy, E.S.M., Ibrahim, A., Talaat, F.M., and Tarek, Z., Water quality prediction using machine learning models based on grid search method, *Multimed Tools Appl.*, 2023.
https://doi.org/10.1007/s11042-023-16737-4
24. Suwadi, N.A., Derbali, M., Sani, N.S., Lam, M.C., Arshad, H., Khan, I., and Kim, K. Il, An optimized approach for predicting water quality features based on machine learning, *Wirel Commun. Mob. Comput*, 2022.
https://doi.org/10.1155/2022/3397972
25. Szomolányi, O. and Clement, A., Use of random forest for assessing the effect of water quality parameters on the biological status of surface waters, *GEM*, 2023, vol. 14.
https://doi.org/10.1007/s13137-023-00229-6
26. Talukdar, S., Shahfahad, Ahmed, S., Naikoo, M.W., Rahman, A., Mallik, S., Ningthoujam, S., Bera, S., and Ramana, G.V., Predicting lake water quality index with sensitivity-uncertainty analysis using deep learning algorithms, *J. Clean Prod.*, 2023, vol. 406.
https://doi.org/10.1016/j.jclepro.2023.136885
27. Tan, G., Yan, J., Gao, C., and Yang, S., Prediction of water quality time series data based on least squares

support vector machine, *Procedia Engineering*, 2012, pp. 1194–1199.
https://doi.org/10.1016/j.proeng.2012.01.1162

28. The Water in You, Water and the Human Body, https://www.usgs.gov/special-topics/water-science-school/science/water-you-water-and-human-body (Accessed October 26, 2023).
29. Uddin, M.G., Nash, S., Rahman, A., and Olbert, A.I., Performance analysis of the water quality index model for predicting water state using machine learning techniques, *Process Safery Environ. Protect.*, 2023, vol. 169, pp. 808–828.
https://doi.org/10.1016/j.psep.2022.11.073
30. Wang, K., Band, S.S., Ameri, R., Biyari, M., Hai, T., Hsu, C.C., Hadjouni, M., Elmannai, H., Chau, K.W., and Mosavi, A., Performance improvement of machine learning models via wavelet theory in estimating monthly river streamflow, *Eng. Appl. Comput. Fluid Mech.*, 2020, vol. 16, pp. 1833–1848.
https://doi.org/10.1080/19942060.2022.2119281
31. Wang, X., Li, Y., Qiao, Q., Tavares, A., and Liang, Y., Water quality prediction based on machine learning and comprehensive weighting methods, *Entropy*, 2023, vol. 25.
https://doi.org/10.3390/e25081186
32. Wang, X., Li, Y., Qiao, Q., Tavares, A., and Liang, Y., Water quality prediction based on machine learning and comprehensive weighting methods, *Entropy*, 2023, vol. 25.
https://doi.org/10.3390/e25081186
33. Wong, W.Y., Al-Ani, A.K.I., Hasikin, K., Khairuddin, A.S.M., Razak, S.A., Hizaddin, H.F., Mokhtar, M.I., and Azizan, M.M., Water quality index using modified random forest technique: assessing novel input features, *CMES*, 2022, vol. 132, pp. 1011–1038.
https://doi.org/10.32604/cmes.2022.019244
34. Wu, X., Zhang, Q., Wen, F., and Qi, Y., A water quality prediction model based on multi-task deep learning: a case study of the Yellow River, China, *Water*, 2022, vol. 14.
https://doi.org/10.3390/w14213408
35. Xu, J., Xu, Z., Kuang, J., Lin, C., Xiao, L., Huang, X., and Zhang, Y., An alternative to laboratory testing: Random forest-based water quality prediction framework for inland and nearshore water bodies, *Water*, 2021, vol. 13.
https://doi.org/10.3390/w13223262
36. Yahya, A.S.A., Ahmed, A.N., Othman, F.B., Ibrahim, R.K., Afan, H.A., El-Shafie, A., Fai, C.M., Hossain, M.S., Ehteram, M., and Elshafie, A., Water quality prediction model based support vector machine model for ungauged river catchment under dual scenarios, *Water*, vol. 11.
https://doi.org/10.3390/w11061231
37. Zhou, J., Wang, J., Chen, Y., Li, X., and Xie, Y., Water quality prediction method based on multi-source transfer learning for water environmental IoT system, *Sensors*, 2021, vol. 21.
https://doi.org/10.3390/s21217271

**Publisher's Note.** Pleiades Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
AI tools may have been used in the translation or editing of this article.