Cureus Journal of Computer Science

Open Access Original Article

A Study of Water Quality Monitoring System With Internet of Things and Machine Learning Regression Techniques

Received 03/22/2025
Review began 03/25/2025
Review ended 06/08/2025
Published 06/25/2025

© Copyright 2025
Roy et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI: https://doi.org/10.7759/s44389-025-04028-y

Sujan Chandra Roy [1] , Md Ashraful Islam [2] , Ripa Sarkar [3] , Ratna Rani Sarkar [4] , Ferdaus Anam Jibon [5] , Laila Naznin [2]

1. Department of Computer Science and Communication Engineering, Kishoreganj University, Kishoreganj, BGD 2. Department of Information and Communication Engineering, University of Rajshahi, Rajshahi, BGD 3. Department of Computer Science and Engineering, Uttara University, Dhaka, BGD 4. Department of Computer Science, National University, Gazipur, BGD 5. Department of Computer Science and Engineering, IUBAT—International University of Business Agriculture and Technology, Dhaka, BGD

Corresponding author: Md Ashraful Islam, ras_ice@ru.ac.bd

Abstract

Water is a fundamental resource for sustaining life. However, its quality has been severely compromised due to rapid urbanization and industrialization. Conventional water quality assessment methods rely on costly and time-intensive laboratory analyses, which are inadequate for real-time monitoring. In this paper, we propose a cost-effective, real-time water quality monitoring system integrating Internet of Things (IoT) technology with machine learning (ML) models to address this limitation and enhance predictive accuracy. The system employs four critical water quality parameters, such as temperature, turbidity, pH, and total dissolved solids, that are collected through IoT-enabled sensors, stored on a server, and analyzed using ML techniques. Unlike existing approaches, our model leverages a unique dataset from multiple regions across Bangladesh, ensuring broader applicability. Experimental evaluations demonstrate that the Random Forest Regression model outperforms other ML techniques, achieving high accuracy in alignment with World Health Organization's standards. Furthermore, the system is designed for affordability, utilizing low-cost embedded devices like Raspberry Pi, making it particularly beneficial for resource-constrained and rural communities. The proposed framework presents a scalable and efficient real-time water quality monitoring solution, offering a practical tool for environmental management and public health decision-making.

Categories: IoT Applications, IoT Data Management and Analytics, Machine Learning (ML)
Keywords: machine learning regression, water quality prediction, iot, sensors, raspberry pi 4b

Introduction

Water is a vital resource crucial for sustaining all life forms, but pollution caused by human activities constantly threatens it. Water quality (WQ) has deteriorated alarmingly, making it a widely communicated medium with far-reaching effects due to rapid industrialization. The spread of devastating diseases is a significant consequence of poor WQ, as 80% of diseases are waterborne [1-2]. Waterborne illnesses that occur frequently in Bangladesh include diarrhea, cholera, dysentery, and hepatitis. Among them, diarrhea is the most frequent illness caused by water [3-4]. However, ensuring access to clean water is a fundamental challenge that requires collaboration among governments, organizations, and individuals. Traditional WQ monitoring involves sample collection and lab analysis, which is time-consuming and inefficient, making it less effective in identifying WQ. The severe consequences of water pollution emphasize the urgent need for a quicker, more cost-effective alternative to traditional methods [5-6]. In the past few years, the development of computer technology, such as the internet of things (IoT), machine learning (ML), and deep learning (DL), has had an impact on earlier behaviors. Among them, IoT and ML are emerging technologies that enable researchers from different disciplines to construct systems that benefit human life.

The popularity of IoT technologies has been on the rise worldwide due to the growing number of devices connecting to the Internet. Devices in IoT-based systems typically connect through wireless networks and are expected to reduce operational and logistical costs. On the other hand, ML enables systems to learn from data, improving performance over time and ensuring efficient, human-like decision-making [7-8]. Nowadays, advanced WQ monitoring systems with higher accuracy are being developed by integrating IoT and ML. Therefore, combining IoT and ML will be a powerful solution to address the challenges of WQ monitoring and management challenges, with several advantages over conventional techniques. By taking advantage of these technologies, WQ monitoring systems can be more cost-effective, scalable, and capable of providing real-time insights, ultimately leading to improved decision-making and resource management [9-10].

Therefore, this paper presents a highly efficient real-time system for monitoring WQ that simplifies traditional monitoring methods by integrating IoT and ML approaches. The main goal of this proposal is to identify pure drinking water in real-time, which will provide an efficient and reliable solution to ensure safe water access. The effectiveness of the proposed system is evaluated through a testbed developed using a Raspberry Pi 4B [11-12] and four sensors: total dissolved solids (TDS) [13], turbidity [14], pH [15], and temperature [16] sensors. Data is collected from various locations in Bangladesh and transmitted directly to a server PC for storage. The data is then used to prepare the dataset for predicting WQ. The collected data is subsequently analyzed and processed through the framework's application of ML models. In order to evaluate the data, a number of regression models are used, which include linear regression (LR) [17], XGBoost regression (XGBR) [18], random forest regression (RFR) [19], support vector regression (SVR) [20-22], ridge regression (RR), multilayer perceptron regression (MLPR), AdaBoost regression (ABR), gradient boosting regression (GBR), decision tree regression (DTR), and K-nearest neighbors regression (KNNR) [23-24]. These models ensure that the assessment of WQ is highly accurate and precise. The use of advanced ML techniques in this framework provides actionable insights that allow decision-makers and policymakers to take proactive measures for efficient WQ management. The main contributions of this paper are summarized as follows:

o The proposed system ensures efficient remote monitoring and analysis of WQ without data loss, mitigating health risks for humans and animals.

o Unlike many existing studies that rely on synthetic or publicly available datasets, our system collects real-time, location-specific WQ data from multiple divisions in Bangladesh (Rajshahi, Dhaka, and Chattogram). To the best of our knowledge, this is among the first efforts to compile and utilize such a dataset for ML-based WQ prediction in this region.

o We have developed and tested a fully functional testbed using Raspberry Pi and cost-effective sensors, demonstrating that the proposed model can be feasibly deployed in rural or low-income regions - a vital factor for real-world impact, especially in developing countries.

o The affordability of the system is ensured by utilizing a server-based IoT framework, reducing computational overhead while maintaining high efficiency.

o Finally, the fusion of IoT and ML enables comprehensive WQ assessment and prediction, providing actionable insights for informed decision-making.

The structure of this paper is as follows: The Materials and Methods section examines the system that combines IoT with ML. The results of the real-time experiment obtained during the study are presented in the Results section. The outcomes and their implications are analyzed in the Discussion section, while the Conclusion summarizes the major findings and insights from the research, providing a comprehensive overview of the study.

Materials And Methods

The main aim of this study is to build a system that can monitor drinking WQ consistently in distant locations through the integration of IoT and ML, which is low-power, low-cost, and timesaving. The development of this paper is divided into two major segments:

(i) Implementation of the IoT System: This segment develops and deploys the hardware components necessary for an IoT-based system, which includes sensors, communication modules, and data collection devices.

(ii) Implementation of the ML Models: ML is applied in this segment to analyze collected data and predict the WQ.

In Figure *1*, the flowchart of the proposed WQ prediction system is presented. The details of the IoT framework and machine learning models are explained further in the following subsection.

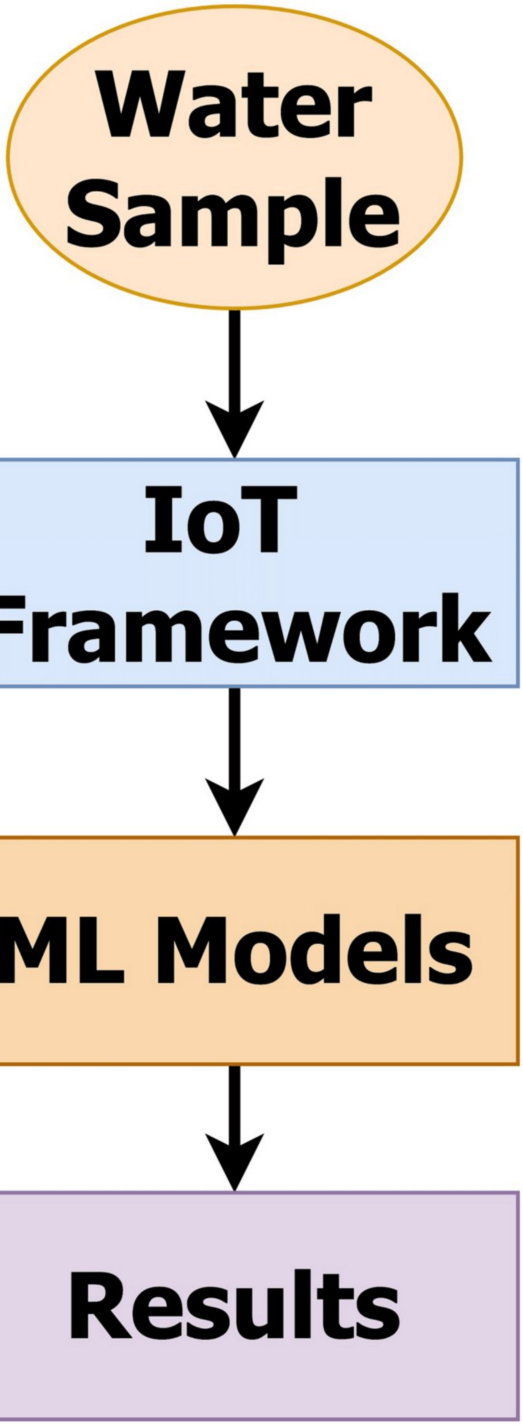

FIGURE 1: Flowchart of the proposed system

IoT, Internet of Things; ML, Machine Learning

Implementation of the IoT system

Figure 2 illustrates the proposed procedure and the visualization of the IoT framework. In the proposed model, the implemented IoT system uses four sensors, including a temperature sensor, pH sensor, turbidity sensor, and TDS sensor, to gather real-time data. Moreover, the IoT system also includes a Raspberry Pi 4B for communication and network establishment purposes, an IoT server, and storage of wirelessly collected data samples. In the following section, the implemented IoT system is described briefly.

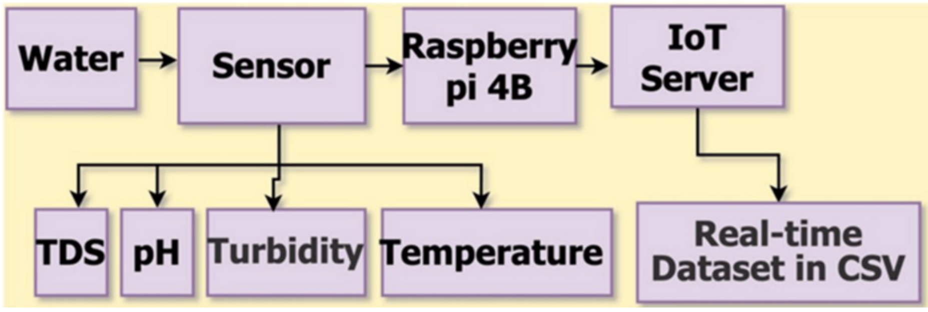

FIGURE 2: IoT framework

IoT, Internet of Things; TDS, Total Dissolved Solids

Sensor for Data Collection

Sensors are capable of detecting changes within their environment by collecting data from it. The initial collection of data from water samples involved measuring four key parameters: TDS, turbidity, temperature, as well as pH.

(i) TDS: The TDS sensor serves as a key indicator of WQ by measuring the concentration of dissolved substances. The standard way to express this measurement is parts per million (ppm) or milligrams per liter (mg/L) [13].

(ii) Turbidity: Turbidity is determined by measuring the number of suspended particles in the water, which can be used to assess its clarity or cloudiness. Turbidity sensors quantify the amount of light scattered by suspended particles, which is then used to determine the haziness of water [14].

(iii) Temperature: A temperature sensor measures water temperature and provides essential data about its quality. It is important to keep the temperature right, as higher temperatures can accelerate bacterial growth in water compared to room temperature conditions. The average human body temperature is around 25-30°C [16].

(iv) pH Sensor: A pH sensor is a device that measures the amount of hydrogen ions in a solution and determines its acidity or alkalinity. The pH scale ranges from 0 to 14, with values lower than 7 indicating acidity, values higher than 7 indicating alkalinity, and a pH of 7 being neutral [15].

Raspberry Pi 4B

The Raspberry Pi 4B is a card-sized, low-power computing device that can perform computation and networking tasks. Its built-in wireless network interface (NIC), which supports IEEE 802.11n and 802.11AC standards, makes it a great choice for IoT applications [11-12].

IoT Server

Our system utilized a Python script for data collection and storage on the server PC. Its primary purpose is to obtain parameterized data from sensors with regular one-second intervals and generate a CSV file on the server computer for future analysis. Table *1* presents the IoT server PC environment, and Table *2* explains the hardware and software configurations that influence measurements.

Brand	Asus
CPU	Intel Core i7
Memory	8 GB
OS	Ubuntu LTS 14.04

TABLE 1: PC environment

Station	Hardware	Model
Controller	Raspberry Pi	4B
	CPU	Broadcom BCM2711 @1.5 GHz
	RAM	8 GB LPDDR4-3200 SDRAM
	Operating System	Linux Raspbian
	Software	hostapd V 2.10
Sensor	Temperature	DS18B20 Temperature Sensor
	pH	SEN0161 pH sensor
	Turbidity	SEN0189 analog Turbidity sensor
	TDS	KS0429 TDS Meter V1.0

TABLE 2: Hardware and software list

TDS, Total Dissolved Solids

Real-Time Dataset

Real-time data is collected from water samples in different locations of Bangladesh, transmitted wirelessly from the Raspberry Pi 4B to the server PC through the IoT server, and stored in CSV formats on the server PC. The collected data are then utilized for analysis and prediction of WQ.

Implementation of the ML model

The proposed WQ monitoring system utilizes ML models to predict drinking WQ. Figure *3* illustrates the block diagram of the implemented ML model. This figure shows that the collected real-time data are analyzed with different ML models to predict WQ. Figure *4* depicts the systematic framework for predicting drinking WQ. The following section briefly explains the steps.

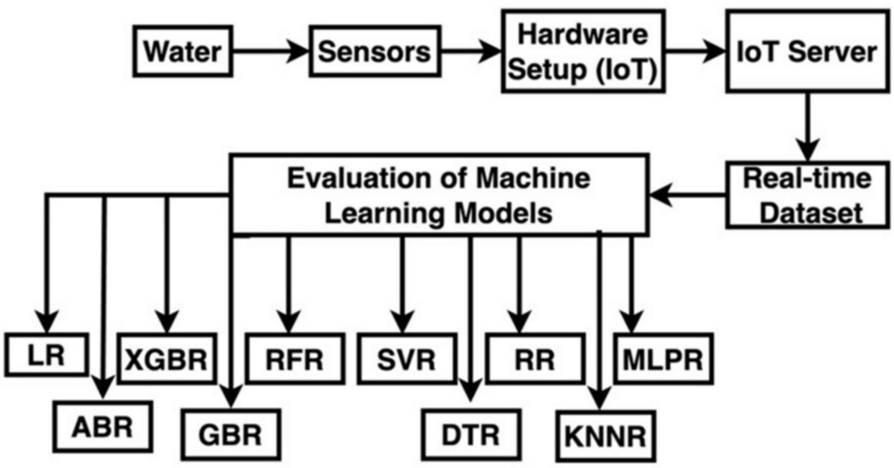

FIGURE 3: A comprehensive diagram of the machine learning models

LR, Linear Regression; XGBR, Extreme Gradient Boosting Regressor; RFR, Random Forest Regressor; SVR, Support Vector Regression; RR, Ridge Regression; MLPR, Multi-layer Perceptron Regressor; ABR, AdaBoost Regressor; GBR, Gradient Boosting Regressor; DTR, Decision Tree Regressor; KNNR, K-Nearest Neighbors Regressor

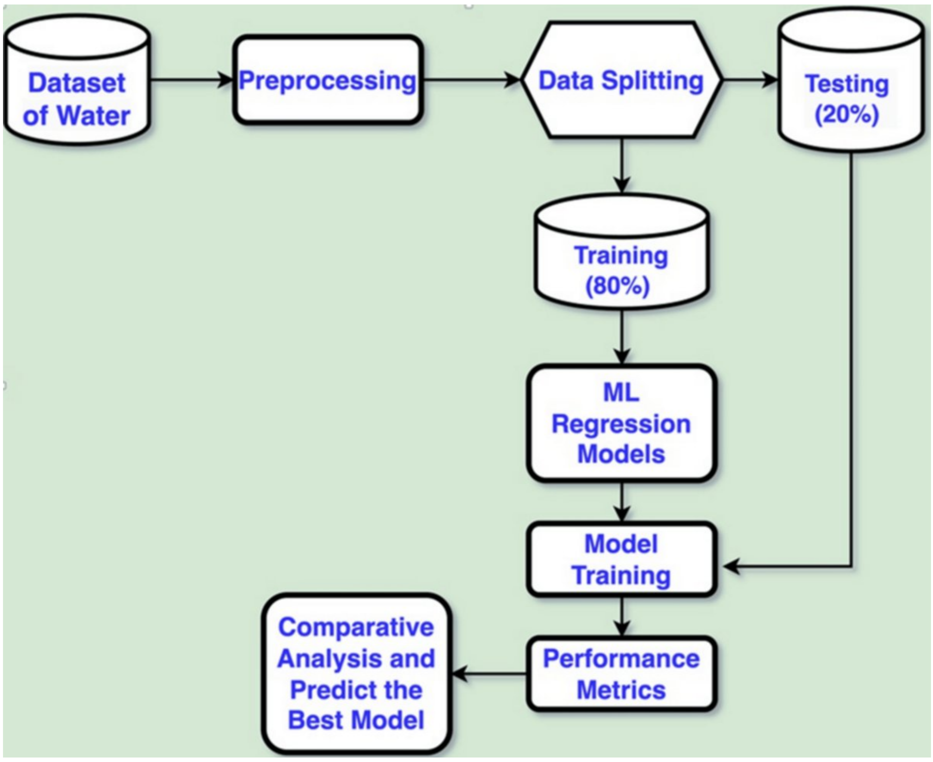

FIGURE 4: ML framework

ML, Machine Learning

Dataset

We label the data obtained from the IoT server in accordance with the standard WQ values for Bangladesh. Table 3 provides a summary of the acceptable boundaries set by the World Health Organization (WHO) for safe drinking water in Bangladesh. For each parameter, 4,762 data points were collected, resulting in a comprehensive dataset for subsequent analysis from Rajshahi, Dhaka, Chittagong Division of Bangladesh [25].

The selection of Dhaka, Chattogram, and Rajshahi as study regions was based on their critical importance to Bangladesh's water security challenges. Dhaka, as the densely populated capital with over 21 million residents, represents the urgent need for monitoring urban water systems affected by industrial pollution and high demand. The city's status as the nation's administrative and economic centre makes WQ maintenance essential for both public health and continued development. Chattogram was included due to its dual role as Bangladesh's commercial hub and premier tourist destination. As the country's main seaport handling 90% of international trade, the region faces unique water contamination risks from industrial and shipping activities. Simultaneously, its coastal location and proximity to popular beaches like Patenga require rigorous WQ monitoring to protect both local communities and visitors. Rajshahi provides a distinct perspective as a culturally significant region and popular tourist destination known for its historical landmarks and agricultural production.

WQ Parameters	Bangladesh Standards	WHO guide lines
pH	6.5–8.5	6.5–8.5
Turbidity	10 NTU	5 NTU
TDS	1,000 ppm	1,000 ppm
Temperature	10–22°C	10–22°C

TABLE 3: WQ parameters for drinking water

WHO, World Health Organization; WQ, Water Quality; TDS, Total Dissolved Solids; NTU, Nephelometric Turbidity Unit; PPM, Parts Per Million

Data Pre-Processing

To improve its reliability and quality, it is crucial to pre-process the data before using it for training models. This step is crucial when dealing with raw datasets gathered from primary sources, as they can have values that are inconsistent or irrelevant. The dataset was carefully examined, and it was found that there were extraneous attributes like "-" and "_" . To ensure accuracy and consistency for further analysis, unnecessary values were removed and replaced by either missing or non-numeric values. To solve missing values, which made up less than 2% of the total records, we used forward/backward filling (FFILL/BFILL), a method chosen for its effectiveness in maintaining temporal continuity in hydrological time-series data. Due to the significant short-term autocorrelation observed in our WQ parameters, this approach was particularly appropriate and aligned with the established monitoring protocols of the WHO. A noise detection and removal method was utilized to enhance data quality. Errors or outlier data points were recognized and eliminated using domain-specific thresholds. StandardScaler normalization was utilized to ensure dataset uniformity and prevent scale-related distortion, changing water levels into a 0,1 range.

The exploratory phase of our analysis relies heavily on graphical visualization to reveal insights and patterns within the CSV dataset. Initially, a heatmap is used to present Spearman correlation coefficients between various WQ parameters, indicating the strength and direction of monotonic relationships as shown in Figure 5. In the heatmap, WQ parameters in both rows and columns are represented by their color intensity, which is a reflection of the correlation strength between the intersecting parameters.

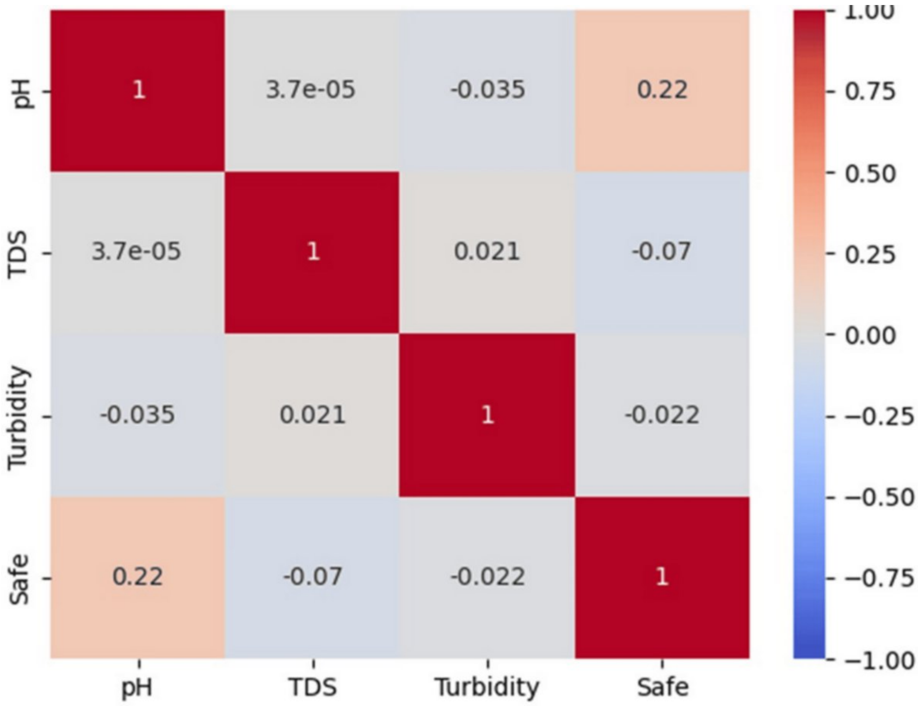

FIGURE 5: Correction matrix of the collected dataset

TDS, Total Dissolved Solids

Data Splitting

To ensure unbiased and effective model evaluation, the dataset was randomly divided into training and testing subsets. Specifically, 80% of the entire dataset was used for training the machine learning models, while the remaining 20% was reserved for testing. The data split was conducted across the complete dataset, without considering temporal factors such as hourly, daily, or weekly intervals. This approach allows the model to learn from a diverse distribution of samples collected in real time from different locations and environmental conditions. Future work may explore time-based data splitting to assess the temporal generalizability of the model.

ML Regression Models (MLR)

In this study, we evaluated 10 MLR models to predict WQ. The models that were evaluated are LR, RFR, XGBR, MLPR, SVR, RR, ABR, GBR, DTR, as well as KNNR.

(i) LR: A fundamental statistical method that models the relationship between a dependent variable and independent variables using a linear function [17].

(ii) XGBR: A high-performance gradient boosting framework that employs tree-based learning with regularization to enhance accuracy [18].

(iii) RRFR: An ensemble learning method that constructs multiple decision trees to improve robustness and reduce overfitting [19].

(iv) SVR: The SVR is a promising method in ML that is used obtained from SVR and used to solve many nonlinear issues such as regression, convex quadratic programming issues. Nevertheless, the training dataset must be manually annotated, and the SVR technique's three variables must be changed using prior information [20-22].

(v) RR: A linear regression variant incorporating an L2 regularization term to address multi-collinearity and prevent overfitting.

(vi) MLPR: A feedforward artificial neural network trained via backpropagation to capture complex nonlinear patterns.

(vii) ABR: A boosting algorithm that iteratively refines weak learners to minimize residual errors and improve predictive performance.

(viii) GBR: An ensemble technique that builds decision trees sequentially, optimizing predictions by minimizing loss gradients.

(ix) DTR: A non-parametric model that partitions data into hierarchical structures based on feature splits.

(x) KNNR: A distance-based method that estimates target values by averaging the outputs of the k-nearest training samples [23-24].

Model Training

The training of MLR models to predict WQ performance was done as previously described in previous sections. Ten MLR models were trained and evaluated. The real-time validation scores were utilized to monitor the performance of the trained models, guaranteeing their convergence to optimal solutions.

Performance Metrics

The evaluation of MLR techniques for WQ prediction was carried out using different performance metrics, as indicated below:

(i) Mean Absolute Error (MAE): It measures regression accuracy by summing up the absolute values of errors and then dividing the summation by the total number of values. MAE calculating formula is shown in Equation (1), where A_{ac} refers to the actual value, B_{pre} refers to the predicted value, and z refers to the number of considered samples [26].

$$MAE = \frac{\sum(\|A_{ac} - B_{pre}\|)}{z}$$

(ii) Mean Squared Error (MSE): It summarizes error squares and divides them by the total number of predicted values. This attributes more significant weight to larger errors. This is particularly useful in problems with larger errors and larger weights. It is measured by Equation (2), where A_{ac} refers to the actual value, B_{pre} refers to the predicted value, and z refers to the number of considered samples [26].

$$MSE = \frac{\sum(A_{ac} - B_{pre})^2}{z}$$

(iii) Root Mean Squared Error (RMSE): It is just the square root of MSE and is estimated from Equation (3), where A_{ac} refers to the actual value, B_{pre} refers to the predicted value, and z refers to the number of considered samples [26].

$$RMSE = \sqrt{\frac{\sum(A_{ac} - B_{pre})^2}{z}}$$

(iv) Coefficient of Determination (R^2 Score): R-squared error (RSE), also called the coefficient of determination, denoted as R^2, measures the goodness of fit of a model. It quantifies the proportion of variance in the dependent variable that can be attributed to the independent variable(s), as depicted in Equation (4). Higher RSE values signify that the independent variables effectively explain a substantial portion of the variance in the dependent variable [27].

$$R^2 = 1 - \frac{Explained\ variation}{Total\ \ variation}$$

where Explained Variation represents how much of the total variance is accounted for by the model. It is the sum of squared differences between the actual value, A_{ac}, and the predicted value, B_{pre}. Total Variation represents the total variance in the actual data. It is the sum of squared differences between the actual value, A_{ac}, and the mean of actual value $\bar{A_{ac}}$.

(v) Mean Absolute Percentage Error (MAPE): MAPE measures the accuracy of forecasting by taking into account the mean absolute percentage deviation between predicted and observed WQ. According to the theory, models that perform better have lower MAPE values [28]. Equation (5) is used to measure it.

$$MAPE = \frac{100}{} \sum (A_{ac} - B_{pre})/A_{ac}$$

where A_{ac} represents the actual value, B_{pre} represents the predicted value, and z represents the total number of samples considered.

(vi) Explained Variance score (EVS) : The EVS measures the proportion of variance in the observed WQ that the model accounts for, ranging from 0 to 1. A higher EVS value signifies better model performance, which reflects the model's ability to explain data variability regardless of scale [29]. The EVS is calculated using the following Equation (6).

$$EVS = 1 - \frac{var(A_{ac} - B_{pre})}{var(A_{ac})}$$

(vii) Cross-Validation Score: This score uses k-fold cross-validation to estimate model generalization by evaluating performance across multiple training and testing splits [30].

Results

Experimental setup for IoT system and MLR models

Figure *6* presents the real-time testbed system of our proposed solution, which comprises a Linux laptop PC serving as the server. It gathers important data for WQ monitoring and manages the sensors. In the testbed setup, the sensors are wired connected to the Raspberry Pi 4B, and the server PC is also wirelessly connected to the Raspberry Pi [13-16]. This setup makes it possible to collect data from the Raspberry Pi from any location within seconds, and to avoid fluctuations in measurement, the average of the collected data is calculated for each water sample over a 5-minute period, with readings taken every 10-second interval. The experiment was conducted using the Python programming language with Visual Studio (VS.) Code on a conventional laptop equipped with a Windows 11 operating system, an i7 processor, and 8 GB of RAM to train MLR models.

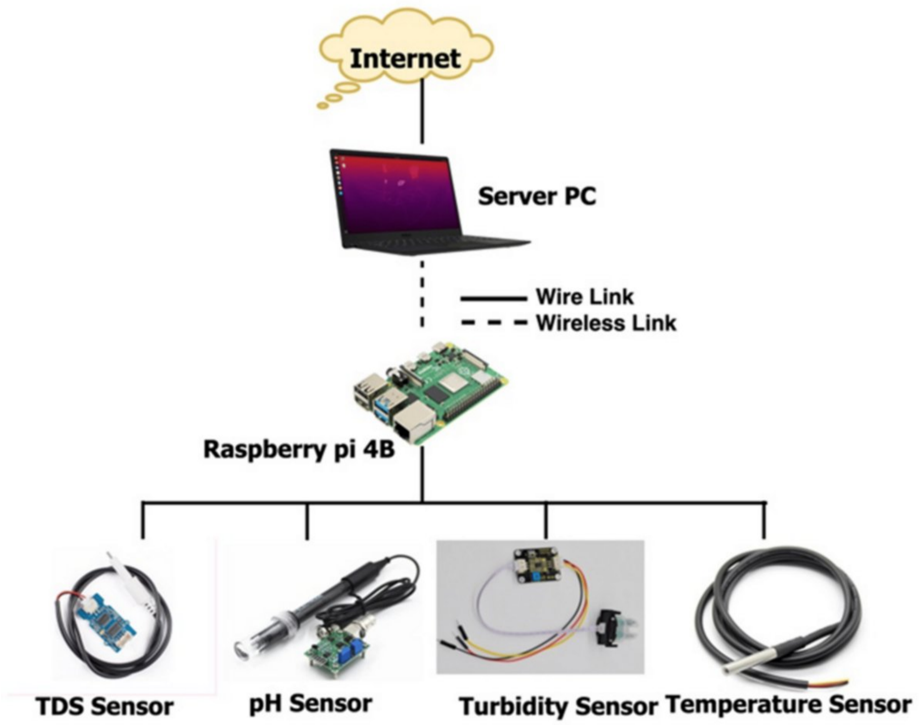

FIGURE 6: Testbed system for IoT framework

Evaluation of different MLR models by utilizing various performance metrics

Ten MLR algorithms were used to analyze real-time values. Table *4* compares performance metrics such as MAE, MSE, RMSE, R^2 score, EVS, MAPE and cross-validation score.

From Table 4, we observe that the RFR model is the most high-performing model for performance metrics, achieving the highest R^2 score of 0.9963, MAE of around 0.0011, MSE of around 0.0009, RMSE of 0.0304, and Cross-Validation Score of 0.9992. The GBR model comes in second place, achieving an R^2 score of 0.9959. Furthermore, it has an MAE of 0.0011, MSE of 0.0011, RMSE of 0.0325, and Cross-Validation Score of 0.9990. The DTR model comes in third place, with an R^2 score of 0.9958, MAE of 0.0012, MSE of 0.0010, RMSE of 0.0324, and Cross-Validation Score of 0.9991. Regarding performance metrics, ABR ranks fourth, with an R^2 score of 0.9915, MAE of 0.0021, MSE of 0.0021, RMSE of 0.0458, and Cross-Validation Score of 0.9983. XGBR ranks fifth, with a Cross-Validation Score of 0.9942, MAE of 0.0039, MSE of 0.0027, RMSE of 0.0523, and R^2 score of 0.9890. MLPR ranks sixth, with an MAE of 0.0689, RMSE of 0.1321, MSE of 0.0174, R^2 score of 0.9297, and Cross-Validation Score of 0.0583. KNNR ranks seventh, with an MSE of 0.0274, R^2 score of 0.8898, MAE of 0.0634, Cross-Validation Score of 0.6007, and RMSE of 0.1654. SVR ranks eighth, with an R^2 score of 0.8101, MAE of 0.1643, MSE of 0.0471, RMSE of 0.2171, and Cross-Validation Score of 0.3230. LR ranks ninth, with a Cross-Validation Score of 0.0263, R^2 score of 0.0435, MAE of 0.4738, MSE of 0.2374, and RMSE of 0.4873.

RFR is superior to other models because of its ensemble learning approach, which utilizes multiple decision trees and a voting mechanism to determine the final decision. This democratic algorithm guarantees reliable and precise predictions, making it highly effective for WQ prediction tasks. Figure 7 presents a graphical representation of the experiment based on the data presented in Table 4 by illustrating the comparative performance observed versus predicted values for each model. In this figure, the x-axis indicates the MLR model, and the y-axis indicates the performance evaluation matrices.

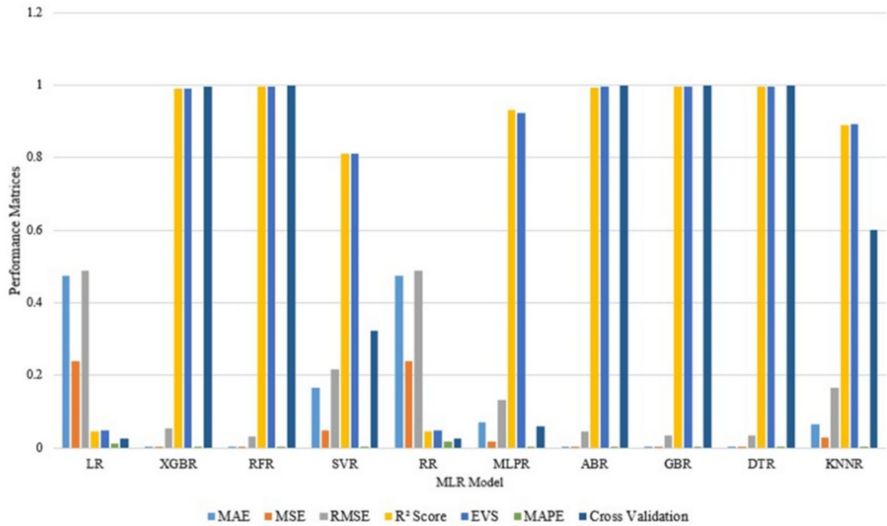

FIGURE 7: Graphical presentation of the experiment

MAE, Mean Absolute Error; MSE, Mean Squared Error; RMSE, Root Mean Squared Error; EVS, Explained Variance Score; MAPE, Mean Absolute Percentage Error; LR, Linear Regression; XGBR, Extreme Gradient Boosting Regressor; RFR, Random Forest Regressor; SVR, Support Vector Regressor; RR, Ridge Regression; MLPR, Multi-Layer Perceptron Regressor; ABR, AdaBoost Regressor; GBR, Gradient Boosting Regressor; DTR, Decision Tree Regressor; KNNR, K-Nearest Neighbors Regressor

Model	MAE	MSE	RMSE	R^2 Score	EVS	MAPE	Cross-Validation
LR	0.4738	0.2374	0.4873	0.0435	0.04865	0.01103	0.0263
XGBR	0.0039	0.0027	0.0523	0.989	0.98902	0.00109	0.9942
RFR	0.0011	0.0009	0.0304	0.9963	0.99593	0.00103	0.9992
SVR	0.1643	0.0471	0.2171	0.8101	0.81123	0.00241	0.323
RR	0.4739	0.2375	0.4873	0.0436	0.04865	0.01602	0.0263
MLPR	0.0689	0.0174	0.1321	0.9297	0.92295	0.00201	0.0583
ABR	0.0021	0.0021	0.0458	0.9915	0.99578	0.00107	0.9983
GBR	0.0011	0.0011	0.0325	0.9959	0.99578	0.00105	0.999
DTR	0.0012	0.001	0.0324	0.9958	0.99578	0.00117	0.9991
KNNR	0.0634	0.0274	0.1654	0.8898	0.8908	0.00213	0.6007

TABLE 4: Performance evaluation using various MLR models

MAE, Mean Absolute Error; MSE, Mean Squared Error; RMSE, Root Mean Squared Error; EVS, Explained Variance Score; MAPE, Mean Absolute Percentage Error; LR, Linear Regression; XGBR, Extreme Gradient Boosting Regressor; RFR, Random Forest Regressor; SVR, Support Vector Regressor; RR, Ridge Regression; MLPR, Multi-Layer Perceptron Regressor; ABR, AdaBoost Regressor; GBR, Gradient Boosting Regressor; DTR, Decision Tree Regressor; KNNR, K-Nearest Neighbors Regressor

Statistical analysis

We used pairwise Wilcoxon signed-rank tests for model comparisons due to their appropriateness for WQ prediction tasks. The nonparametric approach was chosen due to the fact that environmental sensor data frequently shows non-Gaussian distributions and outliers that are contrary to parametric test assumptions. Additionally, the paired design adequately accounts for identical test-set evaluations across models. Furthermore, the method's sensitivity to consistent directional differences is in line with operational needs for reliable contaminant detection. Due to the paired nature of our experimental design, the Wilcoxon test is ideal because we evaluated all ML models on water samples that came from the same location and time point. For multiple comparisons, two-tailed tests were employed [31].

The Wilcoxon signed-rank tests in Table *5* demonstrate that RFR has a significantly better performance ($p < 0.001$) than GBR, XGBR, SVR, MLP, and KNN models. Although RFR performs better than ABR ($p = 0.565$) and DTR ($p = 0.525$), it suggests that these ensemble methods have a similar predictive capability for WQ assessment. The performance of all advanced models is considerably better than that of traditional linear approaches ($p < 2.60 \times 10^{-157}$), with GBR having a particular advantage over DTR (p = 2.08 × 10-161). This validates RFR's ability to be the most robust model when compared to other models.

Comparison	W Statistic	p-value	Significant (α = 0.05)
RFR vs GBR	2854.0	9.30×10^{-158}	True
RFR vs XGBR	951.0	2.78×10^{-156}	True
RFR vs SVR	950.0	3.05×10^{-156}	True
RFR vs MLPR	952.0	3.07×10^{-156}	True
RFR vs KNNR	179.0	1.24×10^{-30}	True
RFR vs ABR	3.0	0.565	False
RFR vs DTR	3.0	0.525	False
GBR vs DTR	260.0	2.08×10^{-161}	True
ABR vs GBR	952.5	1.96×10^{-160}	True
LR/RR vs RFR	167.0	2.60×10^{-157}	True

TABLE 5: Statistical significance of model performance comparisons (Wilcoxon signed-rank tests)

LR, Linear Regression; XGBR, Extreme Gradient Boosting Regressor; RFR, Random Forest Regressor; SVR, Support Vector Regressor; RR, Ridge Regression; MLPR, Multi-Layer Perceptron Regressor; ABR, AdaBoost Regressor; GBR, Gradient Boosting Regressor; DTR, Decision Tree Regressor; KNNR, K-Nearest Neighbors Regressor

Discussion

The analysis of 10 MLR models reveals significant variations in predictive performance across different algorithms. The results indicate that the RFR outperforms all other models, achieving the highest R^2 and cross-validation scores with the lowest error metrics (MAE, MSE, RMSE). This superior performance can be attributed to RFR's ensemble learning approach, which aggregates multiple decision trees to enhance prediction accuracy and reduce overfitting. The weakest models in the study are the LR and RR, both showing poor predictive accuracy with R^2 scores close to zero and notably high error values. The low cross-validation scores for these models further confirm their inefficiency in capturing complex relationships within the dataset. Overall, the study highlights the importance of model selection in predictive analytics. The outstanding performance of RFR underscores the effectiveness of ensemble learning techniques in handling real-time data, making it a promising approach for WQ prediction and similar applications. Future research may further explore hyperparameter tuning and hybrid modelling techniques to enhance predictive performance.

The integration of IoT technology significantly enhances cost-effectiveness through its hardware architecture and data acquisition capabilities. Our system employs modular, low-cost sensors, representing a substantial cost reduction compared to traditional industrial-grade monitoring system. The implementation of edge computing on Raspberry Pi devices further reduces operational expenses by minimizing cloud processing needs and bandwidth requirements. The wireless connection between Raspberry Pi devices and the central server provides dual benefits:

i. Eliminating expensive wired infrastructure installations

ii. Enabling real-time remote monitoring from any location through secure web interfaces

ML contributes substantially to ongoing cost optimization and system performance. Our ML implementation leverages an extensive dataset of over 4,000 WQ measurements where all data is collected in real-time, enabling accurate predictions with minimal computational overhead.

This synergistic combination of cost-conscious IoT design and optimized ML processing creates a sustainable framework for WQ monitoring that balances accuracy with affordability. The wireless connectivity and cloud integration provide additional value by delivering accessible, real-time WQ data to any internet-connected device.

Conclusions

This paper presents an IoT-based system for monitoring WQ that can collect and analyze data in real-time, resulting in significant complexity reduction compared to conventional systems. Four sensors were part of the IoT framework: TDS, turbidity, pH, and temperature sensors. The proposed system was tested and evaluated for effectiveness through testbed experiments in Rajshahi, Dhaka, and Chattogram divisions of Bangladesh, where a Raspberry Pi 4B is the central controller. By using this central controller, water supply monitoring teams can easily obtain real-time information, making it a viable solution to ensure the availability of potable water. By incorporating ML algorithms with the collected sensor data, we implemented 10 MLR algorithms. The results confirmed that the RFR model, with an R^2 score of 0.9963, MAE of approximately 0.0011, MSE of approximately 0.0009, RMSE of 0.0304, and a cross-validation score of 0.9992, was the most successful among them. In future, we will strategically incorporate three additional key regions across Bangladesh such as Rangpur Division, Sylhet Division, and Cox's Bazar - selected through a rigorous multi-criteria analysis and creating a hybrid ML model capable of providing more accurate WQ measurements. Moreover, the system will incorporate advanced sensors and compare their performance to other prediction models to strengthen the overall monitoring framework. A comprehensive hyperparameter optimization using GridSearchCV with 10-fold cross-validation is also planned to systematically refine model performance.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Md Ashraful Islam, Sujan Chandra Roy, Ripa Sarkar, Ratna Rani Sarkar, Ferdaus Anam Jibon

Acquisition, analysis, or interpretation of data: Md Ashraful Islam, Sujan Chandra Roy, Ripa Sarkar, Ratna Rani Sarkar, Ferdaus Anam Jibon, Laila Naznin

Drafting of the manuscript: Md Ashraful Islam, Sujan Chandra Roy, Ripa Sarkar, Ratna Rani Sarkar, Ferdaus Anam Jibon, Laila Naznin

Critical review of the manuscript for important intellectual content: Md Ashraful Islam, Sujan Chandra Roy, Ripa Sarkar, Ratna Rani Sarkar, Ferdaus Anam Jibon, Laila Naznin

Supervision: Md Ashraful Islam, Sujan Chandra Roy, Ferdaus Anam Jibon, Laila Naznin

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Vallino E, Ridolfi L, Laio F: Measuring economic water scarcity in agriculture: a cross-country empirical investigation. Environmental Science & Policy. 2020, 114:73-85. 10.1016/j.envsci.2020.07.017
2. Singh M, Ahmed S: IoT based smart water management systems: A systematic review. Materials Today Proceedings. 2021, 46:5211-5218. 10.1016/j.matpr.2020.08.588
3. Water Quality Thematic Report 2021. (2021). Accessed: February 2025: https://psb.gov.bd/policies/wqtr2019.pdf.
4. Bangladesh Major infectious diseases—Demographics. (2023). Accessed: February 2025: https://www.indexmundi.com/bangladesh/major_infectious_diseases.html.
5. Lemessa F, Simane B, Seyoum A, Gebresenbet G: Assessment of the impact of industrial wastewater on the water quality of rivers around the Bole Lemi Industrial Park (BLIP), Ethiopia. Sustainability. 2023, 15:4290. 10.3390/su15054290
6. Beshir A, Reddythota D, Alemayehu E: Evaluation of drinking water quality and associated health risks in Adama City, Ethiopia. Heliyon. 2024, 10:e36363. 10.1016/j.heliyon.2024.e36363
7. Khan MSI, Islam N, Uddin J, Islam S, Nasir MK: Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. Journal of King Saud University - Computer and Information Sciences. 2022, 34:4773-4781. 10.1016/j.jksuci.2021.06.003
8. Wiryasaputra R, Huang C-Y, Lin Y-J, Yang C-T: An IoT real-time potable water quality monitoring and prediction model based on cloud computing architecture. Sensors. 2024, 24:1180. 10.3390/s24041180
9. Chinnappan CV, John William AD, Nidamanuri SKC, et al.: IoT-enabled chlorine level assessment and prediction in water monitoring system using machine learning. Electronics. 2023, 12:1458.

10.3390/electronics12061458
10. Rahu MA, Chandio AF, Aurangzeb K, Karim S, Alhussein M, Anwar MS: Toward design of internet of things and machine learning-enabled frameworks for analysis and prediction of water quality. IEEE Access. 2023, 11:101055-101086. 10.1109/access.2023.3315649
11. Roy SC, Funabiki N, Rahman MM, Wu B, Kuribayashi M, Kao W-C: A study of the active access-point configuration algorithm under channel bonding to dual IEEE 802.11n and 11ac interfaces in an elastic WLAN system for IoT applications. Signals. 2023, 4:274-296. 10.3390/signals4020015
12. Raspberry Pi. (2023). Accessed: January 2025: http://raspberrypi.org.
13. TDS Sensor in Raspberry Pi. (2025). Accessed: June 2025: https://github.com/vezorgoat/-Raspberry-Pi-TDS-Sensor.
14. How To Use a Turbidity Sensor. (2025). Accessed: June 2025: https://www.hackster.io/Mellow_labs/how-to-use-a-turbidity-sensor-e9b3e7.
15. Raspberry Pi With pH Sensor Using Atlas Scientific pH Sensor Kit. (2025). Accessed: June 2025: https://myhydropi.com/connecting-a-ph-sensor-to-a-raspberry-pi/.
16. Raspberry Pi Temperature Sensor Using the DS18B20. (2025). Accessed: June 2025: https://pimylifeup.com/raspberry-pi-temperature-sensor/.
17. Pal M, Bharati P: Introduction to correlation and linear regression analysis. Applications of Regression Techniques. Springer, Singapore; 2019. 1-18. 10.1007/978-981-13-9314-3_1
18. Chen T, Guestrin C: XGBoost: A scalable tree boosting system. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016, 785-794. 10.1145/2939672.2939785
19. Breiman L: Random forests. Machine Learning. 2001, 45:5-32. 10.1023/a:1010933404324
20. Halim Z, Rehan M: On identification of driving-induced stress using electroencephalogram signals: A framework based on wearable safety-critical scheme and machine learning. Information Fusion. 2020, 53:66-79. 10.1016/j.inffus.2019.06.006
21. Chen H, Huang JJ, McBean E: Partitioning of daily evapotranspiration using a modified Shuttleworth-Wallace model, random forest and support vector regression, for a cabbage farmland. Agricultural Water Management. 2020, 228:105923. 10.1016/j.agwat.2019.105923
22. Cheng Y, Peng J, Gu X, et al.: An intelligent supplier evaluation model based on data-driven support vector regression in global supply chain. Computers & Industrial Engineering. 2020, 139:105834. 10.1016/j.cie.2019.04.047
23. Rong S, Zhang B-W: The research of regression model in machine learning field. MATEC Web of Conferences. 2018, 176:01033. 10.1051/matecconf/201817601033
24. Kumar S, Bhatnagar V: A review of regression models in machine learning. Journal of Intelligent Systems and Computing. 2022, 3:40-47. 10.51682/jiscom.v3i1.30
25. Water_Dataset. (2025). Accessed: March 2025: https://github.com/Sujan-Roy/Water_Dataset.git.
26. Ahmed U, Mumtaz R, Anwar H, Shah AA, Irfan R, García-Nieto J: Efficient water quality prediction using supervised machine learning. Water. 2019, 11:2210. 10.3390/w11112210
27. Menard S: Coefficients of determination for multiple logistic regression analysis. The American Statistician. 2000, 54:17-24. 10.1080/00031305.2000.10474502
28. Aldrees A, Khan MA, Tariq MAUR, Mustafa Mohamed A, Ng AWM, Bakheit Taha AT: Multi-expression programming (MEP): water quality assessment using water quality indices. Water. 2022, 14:947. 10.3390/w14060947
29. Gupta HV, Kling H, Yilmaz KK, Martinez GF: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology. 2009, 377:80-91. 10.1016/j.jhydrol.2009.08.003
30. Kohavi R: A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conferences on Artificial Intelligence. 1995, 14:1137-1143.
31. Rainio O, Jarmo T, Riku K: Author Correction: Evaluation metrics and statistical tests for machine learning. Scientific Reports. 2024, 14:15724. 10.1038/s41598-024-66611-y