Machine Learning-Driven Water Quality Index Prediction: Enhancing Accuracy with Gradient Boosting and Explainable AI for Sustainable Water Monitoring

Md. Jahidul Islam[1], Siraj Us Salekin[2], Asif Anzum[3], Nafis Zaman[1], Abdullah Al Ahad Khan[4], Dilip Sarkar[5], Md. Liton Rabbani[6], Md. Tarek Hossain[6]

Abstract

Background: Water is fundamental to the survival of all life forms, yet access to clean and safe water remains a critical challenge worldwide. Contaminated water is a significant contributor to waterborne diseases, highlighting the need for effective water quality monitoring. The Water Quality Index (WQI) is a standard tool for assessing water quality; however, traditional WQI methods are often constrained by inconsistencies, laboratory inaccuracies, and human error. Methods: This study aimed to overcome these limitations by integrating advanced machine learning (ML) techniques into WQI prediction. Physicochemical parameters, including pH, chloride (Cl^-), sulfate (SO_4^{2-}), sodium (Na^+), potassium (K^+), calcium (Ca^{2+}), magnesium (Mg^{2+}), total hardness, and total dissolved solids, were collected from diverse water sources to form a robust dataset. ML algorithms such as Gradient Boosting, Random Forest, and XGBoost, augmented with explainable AI (XAI), were employed to enhance prediction accuracy. The dataset was split into training (70%), testing (15%), and validation (15%) subsets, and model performance was assessed using RMSE, MSE, MAE, and R^2 metrics. Results: Gradient Boosting outperformed other models, achieving 96% accuracy on the test dataset after, fine-tuning. It demonstrated superior predictive capabilities as evidenced by its performance metrics. These results indicate the potential for ML techniques to address the limitations of traditional WQI methods. Conclusion: This study demonstrates the effectiveness of ML-driven approaches in improving water quality assessments. The integration of Gradient Boosting and explainable AI provides a reliable framework for WQI prediction, enabling better decision-making in environmental health policies and water resource management. This approach offers a pathway to more efficient and accurate water quality monitoring systems.

Keywords: Water Quality Index (WQI), Water Quality Monitoring, Machine Learning Algorithms, Explainable AI (XAI), Predictive Modelling

Significance | This study demonstrates advanced machine learning and Explainable AI techniques for accurate, interpretable Water Quality Index prediction.

*Correspondence. Md. Jahidul Islam, Department of Civil Engineering, Barishal Engineering College, Barishal-8202, Bangladesh.

Editor Md Shamsuddin Sultan Khan And accepted by the Editorial Board October 07, 2024 (received for review August 12, 2024)

Introduction

Most of the freshwater bodies around the world are being contaminated, reducing the suitability of the water. Urbanization has led to an increase in water pollution, posing a severe concern

Author Affiliation.
[1]Department of Civil Engineering, Barishal Engineering College, Barishal-8202, Bangladesh.
[2]Department of Computer Science and Engineering, United International University, Dhaka-1212, Bangladesh.
[3]Institute of Development Studies, University of Sussex, Falmer, Brighton BN1 9RH, UK.
[4]Department of Computer Science and Engineering, United International University, Dhaka-1212, Bangladesh.
[5]Department of Civil Engineering, Dhaka University of Engineering & Technology, Gazipur 1707, Bangladesh.
[6]Assistant Professor, Barishal Engineering College, Barishal-8202, Bangladesh.

Please Cite This:
Md. Jahidul Islam, Siraj Us Salekin, Asif Anzum Nafis Zaman, Abdullah Al Ahad Khan, Dilip Sarkar, Md. Liton Rabbani, Md. Tarek Hossain (2024). "Machine Learning-Driven Water Quality Index Prediction: Enhancing Accuracy with Gradient Boosting and Explainable AI for Sustainable Water Monitoring", Applied Agriculture Sciences, 2(1),1-14,10031

for human life (Agrawal et al., 2021; Chen et al., 2022). Due to the continuous growth in urbanization, only 0.3% of the world's water resources are usable (Kılıç, 2020). Around 785 million people globally do not have access to a safe and reliable water source (Shadabi & Ward, 2022), and around 2.5 billion people do not have proper sanitation. This lack of access to clean and safe water undermines efforts to end extreme poverty and disease in the world's poorest countries (Schweitzer et al., 2020). Water quality assessment plays a crucial role in informing water management decisions. It valuable information about the status of water resources, enabling authorities to identify potential risks and take appropriate measures to protect and improve water quality. Assessment methods, such as the Canadian Council of Ministers of the Environment Water Quality Index (CCME-WQI), the National Sanitation Foundation Water Quality Index (NSF-WQI), the Irrigation Water Quality Index (IWQI) and the Weighted Arithmetic Water Quality Index Method (WAWQI) are commonly used to evaluate water quality (Islam, 2024; Khan et al., 2022). These methods analyse various physicochemical parameters to determine the overall quality of water, including its suitability for drinking, irrigation, and other purposes. By assessing water quality, authorities can identify pollution sources, such as industrial effluents, sewage, and agricultural runoff, and implement measures to mitigate their impact (Lamrini et al., 2022; To, 2020). Water quality assessment also helps in monitoring trends and identifying areas where water quality is deteriorating, allowing for targeted interventions and the development of effective water management strategies (Lee, 2021).

In this article, the authors aim to review existing literature on water quality assessment methods, including traditional indices and the application of machine learning techniques. Additionally, they seek to develop machine learning models for predicting WQI values based on physicochemical parameters, evaluate the performance of these machine learning models against traditional methods in terms of accuracy and efficiency, and investigate the potential of machine learning models to provide dynamic and adapTable solutions for water quality assessment in the face of changing environmental conditions.

In recent years, there has been a growing interest in utilizing machine learning techniques to enhance water quality assessment by predicting WQI values. Machine learning algorithms have shown promise in their ability to analyse vast datasets and generate predictive models (Oreški et al., 2023). These models try to offer a more efficient and accurate way of calculating WQI than traditional methods, which involve manual measurement and analysis of various parameters, by harnessing the power of data-driven modelling (Tabassum et al., 2023). However, multiple studies have revealed that the traditional WQI model produced significantly higher uncertainty in its modelling process (Juwana et al., 2016; Rezaie-Balf et al., 2020; Sutadian et al., 2015; Uddin et al., 2021). As a result, the WQI model needs to reflect accurate water quality status by overestimating and underestimating of WQI values. Some researchers have opted for a non-physical approach to overcome these issues, successfully predicting WQI using artificial intelligence (AI).

However, while the narrative above outlines the current challenges and the potential of machine learning in water quality assessment, it is imperative to ground this discussion in existing scientific literature. Therefore, this study aims to bridge this gap by conducting a comprehensive evaluation of water quality assessment through machine learning, focusing on predicting WQI values.

The application of machine learning techniques in water quality assessment, specifically for the prediction of WQI, is not merely a replacement for deterministic expressions derived from water quality parameters. While it is true that traditional methods offer a straightforward equation for calculating WQI, the complexity and variability inherent in water quality data often necessitate a more nuanced approach. Machine learning models excel in identifying patterns and relationships within large datasets that might not be immediately apparent through deterministic calculations. Furthermore, these models can adapt to changes in water quality parameters over time, offering a dynamic and robust tool for water quality assessment that deterministic methods may not provide. This adaptability is crucial in the face of varying environmental factors and pollution sources, ensuring accurate and timely assessments that support effective water management strategies.

The remainder of this article is as follows. Section 2 explores the related works of the study and a comprehensive review of various machine learning algorithms applications in water quality assessments, including a tabular representation of their recent advancements. Section 3 discussed the step-by-step procedure of the study. Section 4 outlines the result and discussion part. This section also suggested the most reliable model for predictions in our data set analyses. The conclusion part is presented in Section 5.

2 Background Study

2.1 Machine Learning Algorithms

Machine learning algorithms are a set of statistical models and techniques that enable computers to learn and make predictions or decisions without being explicitly programmed. These algorithms learn from data and extract patterns and features to improve their performance. They have become widely used in various applications, such as spam mail classification, image recognition, personalized product recommendations, and natural language processing (Ling, 2023). Machine learning is a subfield of artificial intelligence (AI) and has a subset called deep learning, which uses deep neural networks to learn from complex data. The key characteristics of machine learning algorithms include their ability

to learn from data, make predictions, handle various types of input data, improve through experience and perform
tasks without human intervention (Zhai et al., 2023). Machine learning algorithms work by training data and developing rules based on that learning. These algorithms then evaluate test data to generate results without human intervention. The process involves using various types of input data, such as images, texts, and numbers, to create logical patterns.

Machine learning algorithms can be categorized into several types. Some commonly used machine learning algorithms include support vector machines, logistic regression, decision trees, gradient boosting, random forest and XGBoost (Ren & Du, 2023).

2.2 Water Quality Index (WQI)

A Water Quality Index (WQI) is essential for assessing overall water quality through a single numerical value, reflecting suitability for various uses. The urgency of accurate water assessments is highlighted by the significant marine pollution challenges faced by Bangladesh due to plastic waste (Mim et al., n.d.). WQI is helpful in selecting appropriate treatment techniques and communicating water quality information to the public and decision-makers. Different methods and indices have been developed globally, such as the National Sanitation Foundation's Water Quality Index (NSFWQI), Weighted Arithmetic Water Quality Index (WAWQI), and the British Colombia Water Quality Index (BCWQI) (Mogane et al., 2023). These traditional WQI methods, which rely on grab sampling of physicochemical parameters, can be lengthy and expensive. An enhanced WQI method based on a semi-supervised machine learning technique has been developed to overcome these limitations. This method incorporates a parameter selection step, sub-index calculation, weight assignment, aggregation of sub-indices, and classification (M. Ahmed et al., 2022). By using machine learning algorithms and considering a wide range of parameters, this approach removes uncertainties and improves the accuracy and sensitivity of WQI models (Mueller et al., 2021).

2.3 WQI Detection Using Machine Learning

Machine learning algorithms have been used to predict Water Quality Index (WQI) values in various studies. Bui Quoc Lap et al. (2023) applied feature selection techniques and found that the Random Forest model provided the best accuracy in predicting WQI values from the An Kim Hai system in Vietnam. Md. Galal Uddin et al. (2022) compared eight commonly used algorithms and found that Decision Tree, Extra Tree, Extreme Gradient Boosting, and Random Forest models outperformed others in predicting coastal WQIs in Cork Harbour. Ahmed et al. (2019) investigated the application of machine learning (ML) algorithms for estimating water quality index (WQI) and water quality class (WQC) with a minimal set of input parameters. The proposed method has been developed and tested specifically
for the stream network of the Rawal watershed, and its applicability to other water bodies may need further validation and customization. Wang et al. (2021) proposed a novel approach using a model stacking method to enhance the reliability of beach water quality predictions, which combines the predictions of five base models (MLR, PLS, SPLS, RF, and BN). The study addressed variability challenges across different beaches and consecutive years, achieving robust, cross-validated predictions for beach microbial water quality. Yilma et al. (2018) utilized an artificial neural network (ANN) to predict the water quality index (WQI) of the Little Akaki River, demonstrating promising performance despite generally poor water quality. The ANN model showed promising performance in predicting the CCME-WQI values, achieving an R2 value of 0.93. Sillberg et al. (2021) applied the Attribute-Realization (AR) technique with Support Vector Machine (SVM) for water quality classification, identifying key attributes and showcasing high accuracy. However, the study demonstrated that linear regression outperformed other mathematical functions for classifying river water data, providing a robust foundation for accurate classification. The improvement of Water Quality Index (WQI) prediction accuracy using innovative hybrid machine-learning algorithms was accomplished by Bui et al. (2020). Which emphasizes the superior performance of the BA-RT algorithm. Azad et al. (2018) optimized the Adaptive Neuro-Fuzzy Inference System (ANFIS) with various intelligence algorithms for accurate water quality parameter predictions. Zhang et al. (2019) developed a hybrid statistical model (HANN) combining ANN and genetic algorithm for predicting drinking water treatment plant performance. By utilizing this hybrid model, the researchers sought to provide valuable insights for decision-makers and DWTP managers, allowing them to plan proactively in response to regulatory changes, source water quality fluctuations, and market demand shifts. Khoi et al. (2022) assessed the effectiveness of ML models, highlighting XGBoost's precision in predicting WQI for the La Buong River. Li et al. (2021) developed a robust ML approach using Sentinel-2 MSI data to predict the Water Quality Index (WQI) in the Ebinur Lake Basin, emphasizing spectral modeling indexes' correlation with water quality parameters. The findings are specific to the Ebinur Lake Basin, suggesting potential benefits of higher-resolution data in future studies.

Overall, all the authors discussed and proposed the various uses of machine learning for WQI prediction, and classification. Therefore, an updated machine learning model will be beneficiary in the context of WQI prediction in the current world. Table 1 showcases some of the relevant research on WQI prediction and classification. After going through above-mentioned research, we have found that there is a lack of implementing the Explainable-AI(XAI) model for WQI prediction and classification worldwide. Explainable AI (XAI) is an emerging area of research in the field of Artificial Intelligence

(AI) that focuses on making AI models interpreTable and understandable by humans. XAI techniques aim to provide explanations for how AI systems arrive at their decisions and answers "wh" questions. This is crucial for critical applications, such as prediction, classification, and detection of public health related data, where trust and transparency are essential. So, in this study the author aimed to implement the XAI model into water quality assessment to get several advantages, including real-time monitoring, accuracy in water quality assessing, early detection of water quality deterioration, and informed decision-making for resource management.

3 Methodology

To develop a robust and explainable predictive model, the authors followed a step-by-step process, beginning with the compilation of water quality data from water bodies worldwide, as shown in Figure 1. Subsequently, only the selected parameters were extracted from the datasets to calculate the actual WQI using the WAWQI method to complete the data set. Afterward, the preparation of the data set for analysis is begun. This involves cleaning, which may include handling missing values and outliers, and pre-processing, such as normalization or transformation, to make the data suiTable for modelling.

Prior to modeling, an Exploratory Data Analysis (EDA) was conducted to comprehend the characteristics of the data. This process entailed visualizing distributions, identifying patterns, and exploring potential correlations among variables. Then, the data set is divided into three parts: 70% designated for training purposes, 15% allocated for validation, and the remaining 15% reserved for test sets (Hou et al., 2023). This split is crucial for unbiased model evaluation, allowing the models to be trained and the tuning of hyperparameters without over-fitting. Afterward, different types of regression models on the validation set were assessed to assess the generalization ability of each model without touching the test set. The grid search technique can tune the models' hyperparameters to improve their performance if necessary. The selected models are then trained on the training set, utilizing the entire range of data available for learning. The selected regression models were used, and each model was evaluated based on performance metrics relevant to the project's objectives. Then, the best-performing model has been chosen based on the model's performance on the validation set. Then, the preferred model is finally evaluated on the test set. A rigorous assessment is done on its performance. The authors also generated insights to determine which features most significantly influence the model's predictions. It also helps in understanding and interpreting the model better. The assessment includes a deep dive into the relative importance of each feature. The actionable insights of the model have also been provided. Then, a feature importance plot was generated for individual predictions to understand the model's decision-making process at a granular level.

Finally, a comprehensive assessment of the model was conducted, ensuring that it met the desired criteria and objectives. Throughout this process, authors focus on algorithmic bias and data quality, ensuring the model's predictions are reliable and applicable to real-world scenarios. That is why some outliers, reflected on the test set, are not removed. The outlined workflow facilitates the development of a model that not only performs exceptionally well statistically but also yields actionable insights to empower explainable and interpreTable data-driven decision-making.

3.1 Data obtaining process

Potential of Hydrogen (pH), Total Dissolved Solids (TDS), Chloride (Cl^-), sulphate (SO_4^{2-}), Sodium (Na^+), Potassium (K^+), Calcium (Ca^{+2}), Magnesium (Mg^{+2}), and Total Hardness (T.A) are some of the parameters that were used in this study. Table 2 shows the feature description. A total of 171 monitoring locations' water quality parameters were considered for this study. The selection of water quality parameters for this study was based on the availability of data variables in the water quality assessment research conducted between 2020 and 2022.

3.2 Water Quality Assessment

3.2.1 Water Quality Index Calculation

Water quality indexes are calculated using the Weighted Arithmetic Index Method (Islam, 2024). The steps of this method are given below:

Step: 1 Use this formula to calculate the unit weight (Wn) values for each parameter

$$Wn = \frac{k}{Sn}$$

Where,

$$K = \frac{1}{{}^{1}/_{S1} + {}^{1}/_{S2} + {}^{1}/_{S3} \dots\dots\dots\dots + {}^{1}/_{Sn}} = \frac{1}{\sum {}^{1}/_{Sn}}$$

Sn = Desirable standard value of the n^{th} parameters

Wn = 1(unity) when all selected parameter unit factors are added together

Step: 2 Determine the sub index (Qn) by using the following equation:

$$Qn = \frac{[(V_n - V_0)]}{[(S_n - V_0)]} * 100$$

Where,

V_n = mean concentration of the n^{th} parameters

Sn = Desirable standard value of the n^{th} parameters

V_0 = Real values of parameters in clean water (generally $V_0 = 0$, for most parameters except for pH=7)

$$Q_{pH} = \frac{[(V_{pH} - 7)]}{[(8.5 - 7)]} * 100$$

Step: 3 combining step 1 & step 2, WQI calculate as follows

$$Overall\ WQI = \frac{\sum WnQn}{\sum Wn}$$

3.2.2 Rationale Behind Employing Machine Learning Models

Within the domain of Water Quality Index (WQI) assessment, machine learning presents compelling advantages that complement established WQI equations, even when utilizing laboratory data. The urgency of adapting such advanced methodologies is underscored by the recurring natural disasters in Bangladesh, where floods significantly impact water quality and public health (Islam et al., 2024). Machine learning algorithms with explain-ability possess the remarkable capability to adapt to the inherent complexities of individual water bodies. By leveraging historical data from specific sources like lakes, rivers, or groundwater reserves, these algorithms can incorporate factors such as local pollutants and seasonal variations. This tailored approach can lead to the generation of a more accurate WQI that is highly specific to the water body under study. Furthermore, machine learning offers the potential to identify novel contaminants. By analyzing data for unexpected patterns, these algorithms can signal the presence of emerging contaminants and assess impacts exacerbated by climate change, such as those observed in the coastal regions of Bangladesh, which might not yet be encompassed by the standard WQI equation (Rahman et al., 2024). This ability to provide early warnings of potential water quality issues is a significant strength of machine learning in this context. Additionally, machine learning can optimize the weightings traditionally assigned to various water quality parameters within WQI equations. Through analysis of historical data, machine learning can determine if these weightings require adjustments for specific locations or situations, leading to a more nuanced understanding of water quality. Finally, machine learning models can be integrated with sensor networks that continuously monitor water quality parameters, enabling real-time assessment and facilitating the swifter detection of changes in water quality. It is crucial to recognize that machine learning, in the context of WQI, serves as a refinement tool rather than a complete replacement for existing equations. The WQI equation provides a foundational framework, and machine learning strengthens it by enabling site-specific adaptations and offering the potential for earlier problem identification.

3.3 WQI Classification and Potential Uses

Table 3 shows that the WQI values are divided into five distinct categories, each of which corresponds to a specific water quality status and potential uses. This classification system enables a rapid assessment of water quality, making it a valuable tool for decision-making and management, particularly in situations where quick predictions of water quality are essential for various applications like reducing the health risk (Abdullah et al., 2024).

3.4 Data preparation for modelling

In order to generate the datasets for modelling, the Water Quality Index-WQI of data is considered as dependent variable (Y) whereas other parameters namely Ph, Cl^-, SO_4^{2-}, Na^+, K^+, Ca^{+2}, Mg^{+2}, Total hardness & Total dissolved solids are considered as independent variables X1, X2, X3, X4, X5, X6, X7, X8, and X9, respectively as shown in Table 4. Data presentation is shown in the form of Table 3. Data of the variables were divided into three parts, such as training data set (70%), test datasets (15%) and validating data set (15%). Different dividing strategies of data were carried out to get the best fit for each model. Training data set was then used to learn models whereas validating data set was used to test the models.

3.5 Exploratory data analysis

Exploratory data analysis (EDA) is the process of manipulating data to learn about general patterns and identify specific occurrences that deviate from those patterns (Albert & Rizzo, 2012). EDA consists of statistical models and graphs, incorporating domain knowledge to find information and generate ideas. In this research the EDA report is subdivided into Distribution Analysis, Correlation Analysis, Anomaly Analysis, Pairwise Relationship Analysis.

3.5.1 Distribution Analysis

Distribution analysis involves examining the spatial distribution of data and detecting patterns or clusters within the data. It is a powerful tool for identifying spatial outliers, detecting spatial association, and gaining profound insights into the water quality parameters used in this study. For instance, Figure 2 shows that, the pH distribution appears to closely resemble a normal distribution, albeit with a slight inclination towards higher pH levels. On the other hand, the

Total Dissolved Solids (TDS) distribution is right-skewed, indicating the presence of samples with exceptionally high TDS levels. The distributions of chloride (Cl^-), sulfate (SO_4^{2-}), sodium (Na^+), potassium (K+), calcium (Ca^{+2}), and magnesium (Mg+2) all exhibit a right-skewed pattern, suggesting that most samples cluster within a specific range while outliers with notably high values exist. Total Hardness, interestingly, presents a bimodal distribution, implying the possible existence of two distinct groups within the data, each with different levels of water hardness. Lastly, the Water Quality Index (WQI), our target variable, also follows a right-skewed distribution, providing important insights into its variability within the data set.

Table 1. Research findings based WQI predictive models and its accuracy.

Task	Method	Model	Accuracy	Reference
Prediction	Machine Learning	Random Tree (RT)	94.10%	Bui Quoc Lap Et Al. (2023)
Regression, Classification	Machine Learning	Gradient Boosting (GB)	74.85%	Ahmed Et Al. (2019)
Prediction	Machine Learning	Random Forest (RF)	90.4%	Wang Et Al. (2021)
Prediction	Machine Learning	Artificial Neural Network (ANN)	93.0%	Yilma Et Al. (2018)
Classification	Machine Learning	Support Vector Machine (SVM)	86.0%-95.0%	Sillberg Et Al. (2021)
Prediction	Machine Learning	Adaptive Neuro-Fuzzy Inference System (ANFIS)	95.0%	Azad Et Al. (2018)

Table 2. Attributes name with its description and standard value suggested by authorities

Attributes Name	Description	Standards of Parameters
pH	A measure of the acidity or alkalinity of water.	6.5-8.5[a]
Cl^-	The concentration of chloride ions, which can indicate salinity or contamination.	200-250 mg/L[a]
SO_4^{2-}	The concentration of sulfate ions, which can indicate the presence of sulfates in water.	200-250 mg/L[a]
Na^+	The concentration of sodium ions, which can be relevant for assessing water salinity.	100-200 mg/L[a]
K^+	The concentration of potassium ions, which may be indicative of certain water quality characteristics.	12 mg/L[b]
Ca^{+2}	The concentration of calcium ions, which can influence water hardness and quality.	0-100 mg/L[b]
Mg^{+2}	The concentration of magnesium ions, which can also affect water hardness.	20 mg/L[b]
T.A	The overall capacity of water to neutralize acids, indicating its buffering capacity.	100-200 mg/L[c]
TDS	The total concentration of dissolved inorganic and organic substances in water.	1000 mg/L[c]

[a]World Health Organization

[b] Ministry of Environment- Al Najaf Environment Directorate

[c] Department of Environment, Bangladesh

Table 3. WQI range, condition, and potential uses of the water sample

WQI	Water quality status (WQS)	potential uses
0–25	Excellent	Human consumption, agriculture and industrial
26–50	Good	Human consumption, agriculture and industrial
51–75	Poor	Agriculture and industrial
76–100	Very Poor	Agriculture
>100	Not suitable for human consumption and fish culture	Needs to be treated properly before use

Source: Brown et al. (1972)

Table 4. Data Presentation

No.	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**	**X9**	**Y**
1	8.44	60	41	67.6	3.36	41	20	180	399	74.22
2	8.29	120	80	97.3	29	42.71	25.65	200	572.67	138.40
3	8.56	110	78	102	17.2	24	33.87	200	556	124.4
4	8.33	110	86	97.8	21	16.07	25	140	487.19	117.11
…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…
168	7.75	175.35	2977.5	122.65	7.25	168.4	43.5	601	1106.5	131.42
169	7.85	239	615.8	183.8	11.25	174.8	48	634	1382.5	130.04
170	7.62	147.774	613.417	105.78	6.488	149.317	44	534.278	984.29	102.52
171	7.6	146.44	321.05	120.15	5.05	112.8	52.68	498	894	99.86

Table 5. Performance Value of All Machine Learning Models

Model	MSE	MAE	RMSE	R^2
Linear Regression	403.25	15.23	20.08	0.58
Random Forest	35.98	4.80	6.00	0.96
KNN	662.72	17.57	25.74	0.31
GB	24.26	3.20	4.93	0.97
XGB	56.32	4.52	24.39	0.94
LGBM	132.48	8.29	11.51	0.86
SVM	206.47	6.93	14.37	0.78
Adaboost	40.98	5.30	6.40	0.96
Catboost	82.10	4.42	9.06	0.91
Ann	623.24	18.41	24.96	0.35

Table 6: Performance Value of Gradient Boosting (Fine Tuned)

Dataset Type	MSE	MAE	RMSE	R^2
Validation Set	14.58	2.64	3.81	0.98
Test Set	162.22	5.75	12.93	0.96

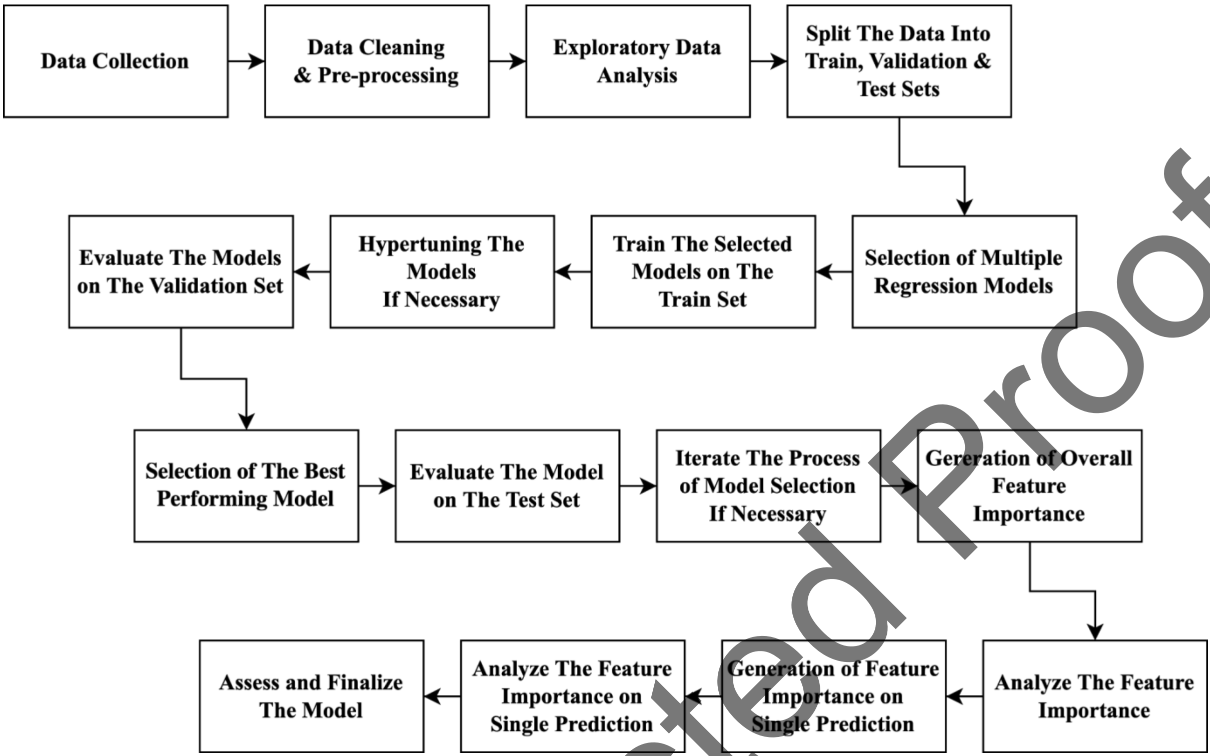

Figure 1. Methodology

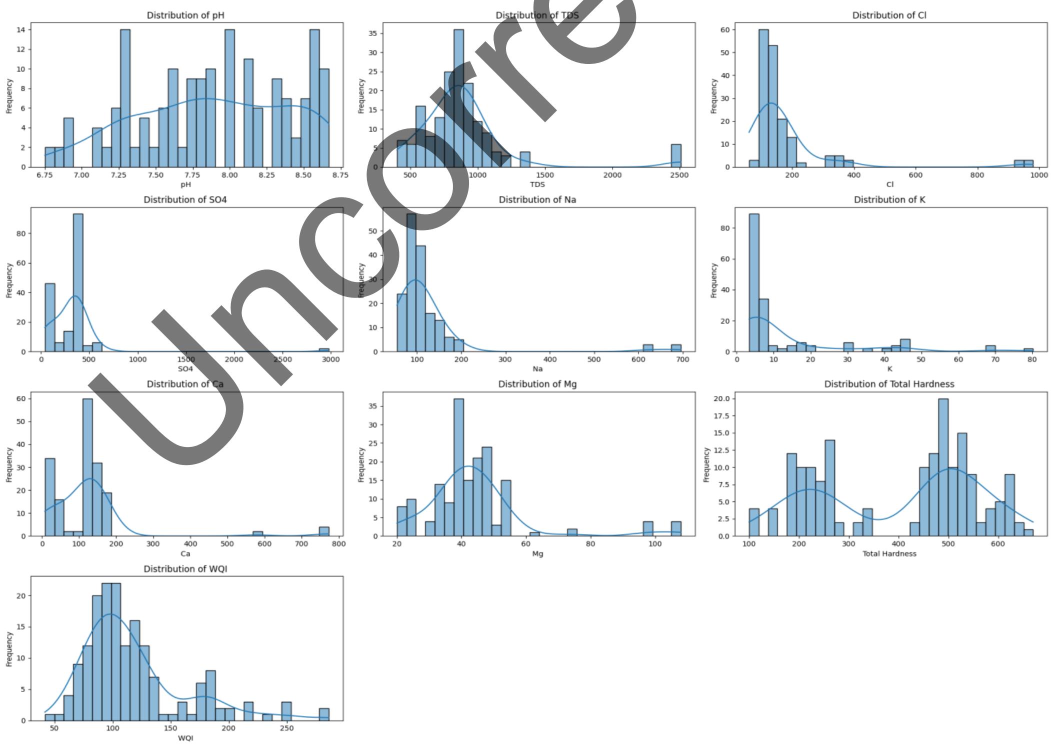

Figure 2. Distribution analysis graphs of the variables

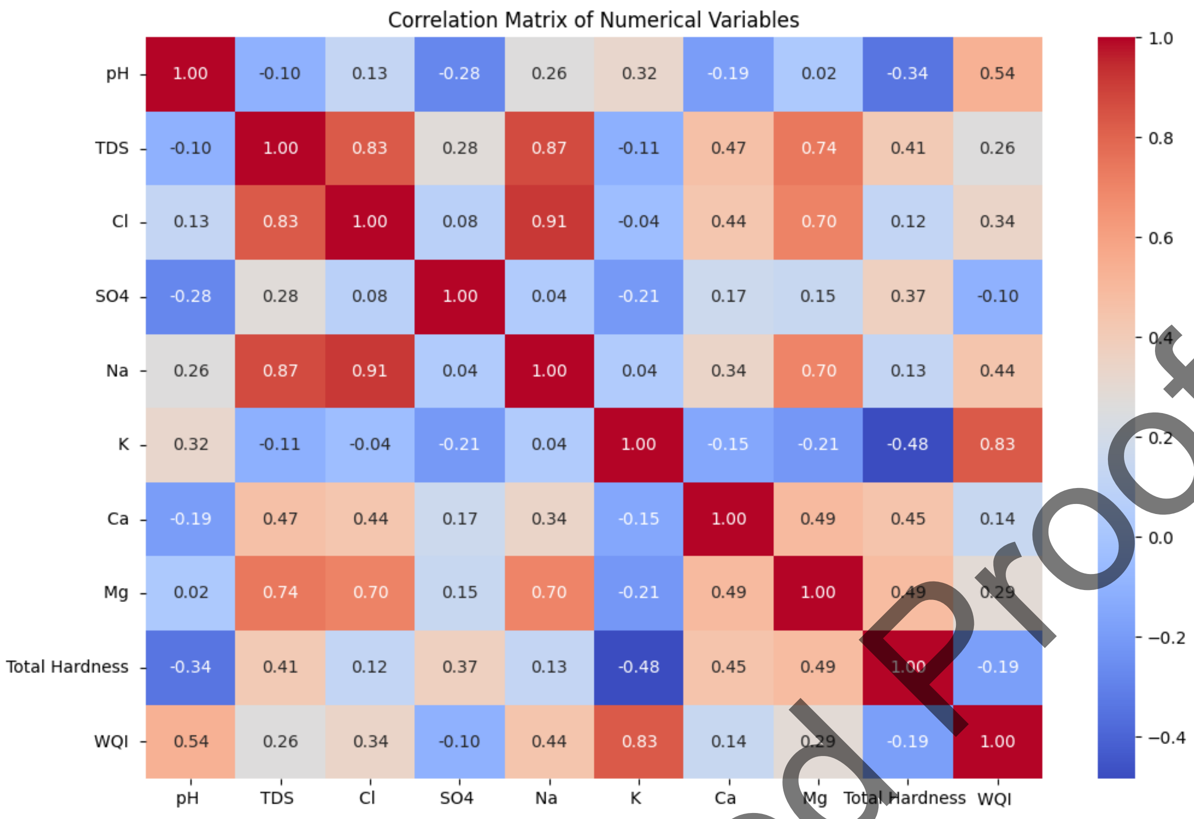

Figure 3. Correlation matrix of variables

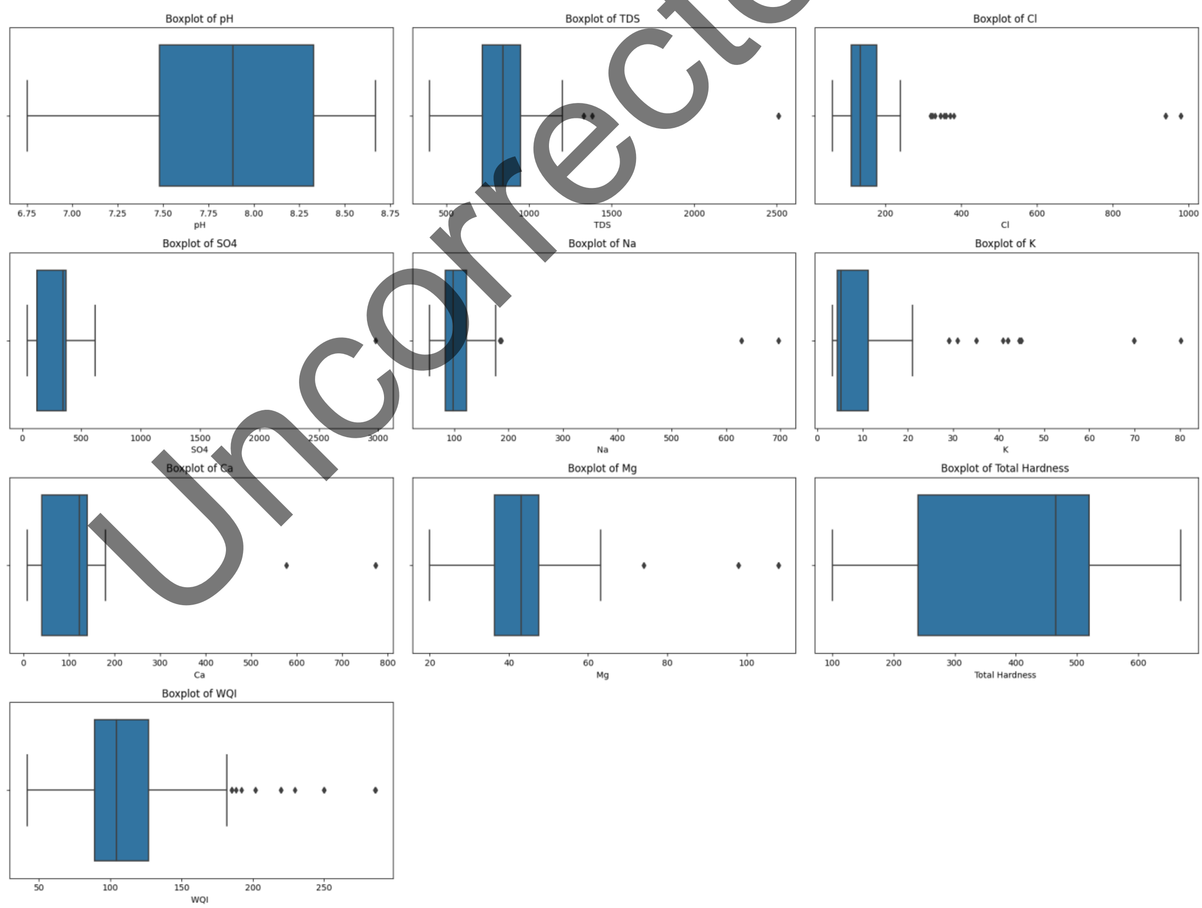

Figure 4. Anomaly Analysis Graphs of variables

Figure 5. Pairwise Relationship Analysis Graphs of variables

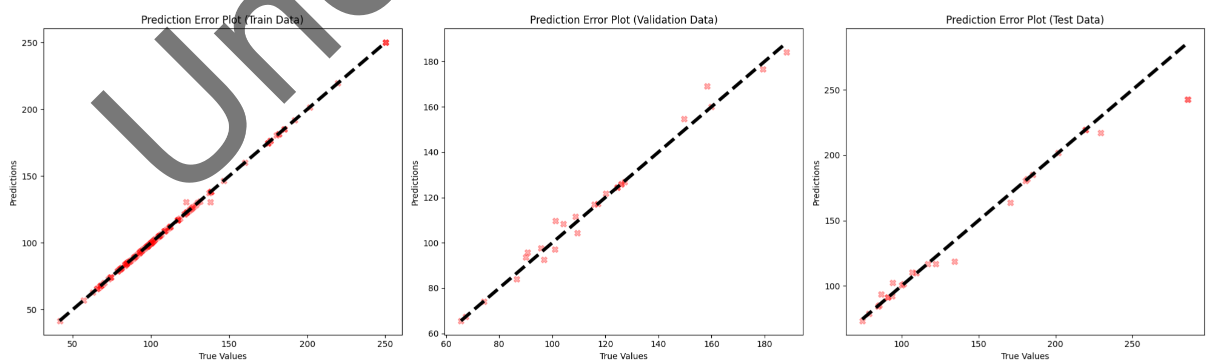

Figure 7. Prediction Error Plot

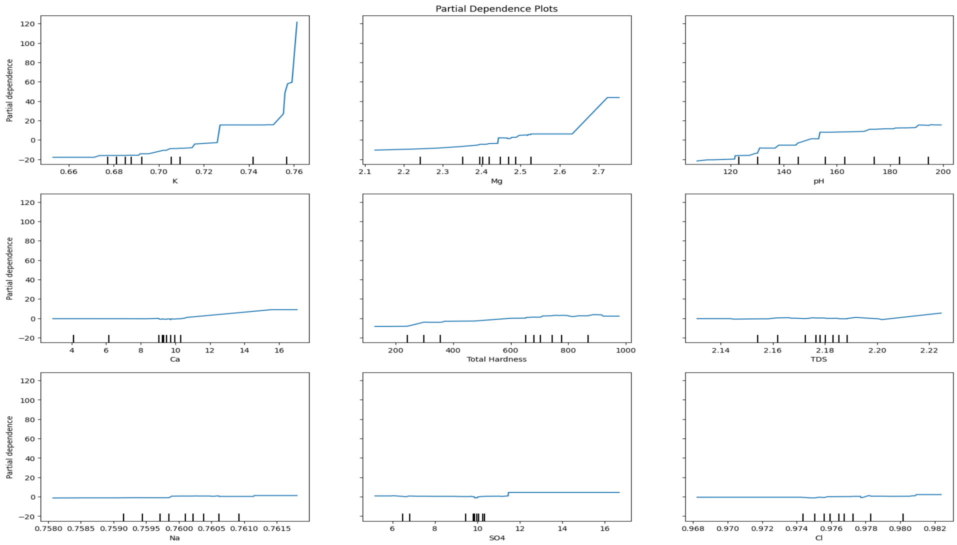

Figure 6. Partial Dependence Plot of GB

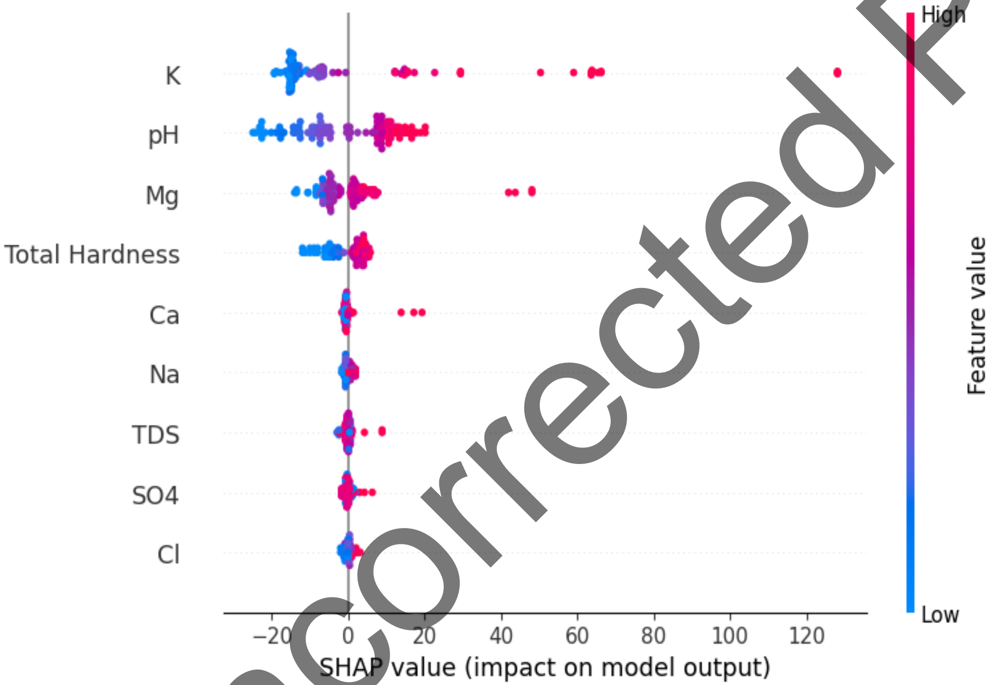

Figure 8. SHAP Summary Plot (Global Importance)

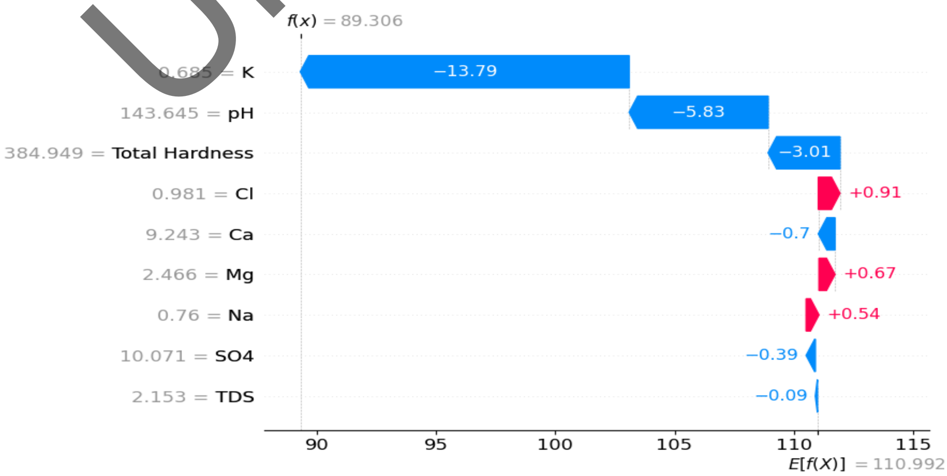

Figure 9. SHAP Summary Plot (Local Importance)

3.5.2 Correlation Analysis

In Figure 3, strong positive correlations were observed between various water quality parameters. TDS and Total Hardness exhibited a very strong positive correlation (0.83), while TDS and CI displayed a strong positive correlation (0.91). Variables such as TDS (0.56) and Total Hardness (0.49) show a moderate positive correlation with the target variable WQI. This suggests that they have important features for predicting WQI.

3.5.3 Anomaly Analysis

In the context of anomaly analysis, the examination of the boxplots in Figure 4 discloses the existence of outliers within the data set. Specifically, TDS, Cl, SO4, Na, K, Ca, and Mg exhibit data points that extend beyond the boxplot whiskers, signifying potential outliers. Total Hardness also displays a few outliers, while the variable WQI indicates the presence of outliers through extreme values in the higher range.

3.5.4 Pairwise Relationship Analysis

The Pairwise Relationship Analysis of the selected parameters in relation to the target variable WQI is shown in Figure 5. It finds that potassium (K) has a somewhat linear relationship with water quality index (WQI), while the relationship between magnesium (Mg) and WQI is less clear. pH levels do not show a strong linear relationship with WQI, but there might be a nonlinear relationship. The relationship between sodium (Na) and WQI doesn't appear to be strong, but there's a noticeable pattern.

4 Result and Discussion

Table 5 represents the Mean Squared Error (MSE), The Mean Absolute Error (MAE), The Root Mean Squared Error (RMSE), R-squared (R2) values of all the machine learning models in the validation datasets. Based on these values it can be said that the Gradient Boosting model consistently excels across all metrics, suggesting it is the most reliable for predictions in our dataset. The Random Forest and XGB models also show strong performance but with some trade-offs in specific metrics. On

the other hand, the Linear Regression and KNN models might need reconsideration or parameter tuning, as their performance metrics indicate a weaker fit and predictive accuracy. After the evaluation of the specific model for these datasets the model is fine tuned for more accurate results.

performance value of the Fine tuned Gradient Boosting model in the validation set and test set is given on Table 6.

4.1 Interpretation of the Model's Performance and Behaviour

4.1.1 Partial Dependence Plot

Partial Dependence Plot (PDP) is a tool that is used in machine learning to illustrate the relationship between selected features and the target variable. It also helps in visualizing and interpreting the marginal effect of a feature on model predictions, revealing patterns, such as linear or non-linear relationships. Figure 6 represents Partial Dependence Plots (PDPs) of Gradient Boosting model. This Figure shows the individual plots that indicates the target variable's(WQI) relationship with the respective features. From those plots, it can easily be understood that K, Mg, and pH have a linear relationship with WQI and the rest of the plots have flat lines indicating a non-linear relationship.

4.1.2 Prediction Error Plot

In regression analysis the Prediction error plots or residual plots compare the actual values with the predicted values which reveals the model's accuracy and the distribution of errors. Figure 7 represents the prediction error plots of Gradient Boosting model which shows that the data points close to the line of perfect prediction and residuals are randomly distributed around the zero line, indicating accurate and unbiased predictions. Additionally, these plots help to identify patterns, suggest model limitations, highlight outliers, and provide a visual tool to evaluate and improve model performance.

4.1.3 Shapley Additive Explanations Summary Plot

Figure 8 & 9 exhibits SHAP (SHapley Additive exPlanations) summary plots for global importance & local importance, respectively. It is an insightful visualization tool in machine learning, offering a comprehensive view of how different features influence a model's predictions across all predictions.

4.1.3.1 SHAP Summary Plot (Global Importance)

In the global importance plot, it ranks features by their importance, depicted at the top of the plot, and uses colour coding to indicate the direction of their impact which shows whether high or low values of a feature increase or decrease the prediction. Additionally, each dot on the plot represents an individual data point, demonstrating the varied effects of feature values across the dataset.

4.1.3.2 SHAP Summary Plot (Local Importance)

It is also known as waterfall plots. In these plots offer an in-depth look at how a model arrives at a specific prediction (f(x)=89.306) for an individual observation. The plot begins with a base value, typically the average prediction (E[f(x)] = 110.992) across the dataset, which serves as a starting point. Each feature's contribution is then represented as individual bars, with their length and direction indicating the magnitude and direction of their impact on the prediction. For example, to predict the value 89.306, the K (Potassium) impacted negatively and the Cl (Clorine) impacted positively. The layout in the Figure allows a clear understanding of both the positive and negative influences of each feature.

5. Conclusion

This research aimed to determine the robust algorithm for predicting the water quality index (WQI) accurately in terms of GNB) were tested and validated for predicting WQI. Predictive models were validated using a number of validators, such as RMSE, MSE, MAE, RMSE and R^2. Explainable AI (XAI) is implemented in model interpretability and understandability. To achieve this goal, ten ML algorithms (Linear Regression, Random Forest, KNN, GB, XGB, LGBM, SVM, AdaBoost, CatBoost, ANN SVM, ExT, LR, and this study for model transparency. The most well-known dataset variables, such as Potential of Hydrogen (pH), Total Dissolved Solids (TDS), Chloride (Cl^-), sulphate (SO_4^{-2}), Sodium (Na^+), Potassium (K^+), Calcium (Ca^{+2}), Magnesium (Mg^{+2}), and Total Hardness (T.A) were obtained in this study. However, it can be concluded that the Gradient Boosting (GB) regression model is effective and robust for predicting the WQI. Additionally, the fine-tuned GB model showcased its ability in the validation set with remarkably low MSE, MAE, and RMSE scores alongside a high R2 value, pointing to its precise predictive capabilities. The findings of this research would also have been much more useful in predicting WQIs at each monitoring site more accurately by the power of Explainable AI. The findings revealed that the Gradient Boosting model performed well in forecasting the water quality index; however, the greatest performance was linked with the higher accuracy (R^2 value). Further studies should be carried out in order to validate the other algorithms in terms of predicting WQIs using temporal variability of data attributes.

Author contributions

All authors made equal contributions to the study design, statistical analysis, and drafting of the manuscript. The corresponding author, along with the co-authors, reviewed and approved the final version of the article prior to submission to this journal.

Acknowledgment

The authors were grateful to their department.

Competing financial interests

The authors have no conflict of interest.

References

Abdullah, M. S., Islam, M. J., Hasan, M. M., Sarkar, D., Rana, M. S., Das, S. S., & Hossian, M. (2024). Impact of waste management on infectious disease control: Evaluating strategies to mitigate dengue transmission and mosquito breeding sites – A systematic review. Journal of Angiotherapy, 8(8), 1–12. https://doi.org/10.25163/angiotherapy.889850

Agrawal, K. K., Panda, C., & Bhuyan, M. K. (2021). Impact of urbanization on water quality. In S. K. Acharya & D. P. Mishra (Eds.), Current advances in mechanical engineering (pp. 665–673). Springer. https://doi.org/10.1007/978-981-33-4795-3_60

Ahmed, M., Mumtaz, R., & Anwar, Z. (2022). An enhanced water quality index for water quality monitoring using remote sensing and machine learning. Applied Sciences, 12(24), Article 24. https://doi.org/10.3390/app122412787

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning. Water, 11(11), 2210. https://doi.org/10.3390/w11112210

Albert, J., & Rizzo, M. (2012). Exploratory data analysis. In J. Albert & M. Rizzo (Eds.), R by example: Concepts to code (pp. 133–151). Springer. https://doi.org/10.1007/978-1-4614-1365-3_5

Azad, A., Karami, H., Farzin, S., Saeedian, A., Kashi, H., & Sayyahi, F. (2018). Prediction of water quality parameters using ANFIS optimized by intelligence algorithms (Case study: Gorganrood River). KSCE Journal of Civil Engineering, 22(7), 2206–2213. https://doi.org/10.1007/s12205-017-1703-6

Brown, R. M., McClelland, N. I., Deininger, R. A., & O'Connor, M. F. (1972). A water quality index—Crashing the psychological barrier. In W. A. Thomas (Ed.), Indicators of environmental quality (pp. 173–182). Springer US. https://doi.org/10.1007/978-1-4684-2856-8_15

Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., & Kazakis, N. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Science of The Total Environment, 721, 137612. https://doi.org/10.1016/j.scitotenv.2020.137612

Chen, S. S., Kimirei, I. A., Yu, C., Shen, Q., & Gao, Q. (2022). Assessment of urban river water pollution with urbanization in East Africa. Environmental Science and Pollution Research, 29(27), 40812–40825. https://doi.org/10.1007/s11356-021-18082-1

Hou, R., Lo, J. Y., Marks, J. R., Hwang, E. S., & Grimm, L. J. (2023). Classification performance bias between training and test sets in a limited mammography dataset (p. 2023.02.15.23285985). medRxiv. https://doi.org/10.1101/2023.02.15.23285985

Islam, M. J. (2024). A study on seasonal variations in water quality parameters of Dhaka rivers. Iranica Journal of Energy and Environment, 15(1), Article 1. https://doi.org/10.5829/IJEE.2024.15.01.09

Islam, Md. J., Abdullah, M. S., & Alam, M. (2024). Flooding crisis in Bangladesh: Urgent measures required. Biodiversity, 25(2), 95–98. https://doi.org/10.1080/14888386.2024.2330385

Juwana, I., Muttil, N., & Perera, B. J. C. (2016). Uncertainty and sensitivity analysis of West Java Water Sustainability Index – A case study on Citarum catchment in Indonesia. Ecological Indicators, 61, 170–178. https://doi.org/10.1016/j.ecolind.2015.08.034

Khan, I., Zakwan, M., & Mohanty, B. (2022). Water quality assessment for sustainable environmental management. ECS Transactions, 107(1), 10133. https://doi.org/10.1149/10701.10133ecst

Khoi, D. N., Quan, N. T., Linh, D. Q., Nhi, P. T. T., & Thuy, N. T. D. (2022). Using machine learning models for predicting the water quality index in the La Buong River, Vietnam. Water, 14(10), 1552. https://doi.org/10.3390/w14101552

Kılıç, Z. (2020). The importance of water and conscious use of water. International Journal of Hydrology. https://doi.org/10.15406/ijh.2020.04.00250

Lamrini, M., Quevy, Q. A., Yassin Chkouri, M., & Touhafi, A. (2022). Data integrity analysis of water quality sensors and water quality assessment. IECON 2022 – 48th Annual Conference of the IEEE Industrial Electronics Society, 1–6. https://doi.org/10.1109/IECON49645.2022.9968643

Lap, B. Q., Phan, T.-T.-H., Nguyen, H. D., Quang, L. X., Hang, P. T., Phi, N. Q., Hoang, V. T., Linh, P. G., & Hang, B. T. T. (2023). Predicting water quality index (WQI) by feature selection and machine learning: A case study of An Kim Hai irrigation system. Ecological Informatics, 74, 101991. https://doi.org/10.1016/j.ecoinf.2023.101991

Lee, S. (2021). Water quality management. In S. Lee (Ed.), China's water resources management: A long march to sustainability (pp. 191–228). Springer International Publishing. https://doi.org/10.1007/978-3-030-78779-0_6

Li, X., Ding, J., & Ilyas, N. (2021). Machine learning method for quick identification of water quality index (WQI) based on Sentinel-2 MSI data: Ebinur Lake case study. Water Supply, 21(3), 1291–1312. https://doi.org/10.2166/ws.2020.381

Ling, Q. (2023). Machine learning algorithms review. Applied and Computational Engineering, ACE, 4, 91–98. https://doi.org/10.54254/2755-2721/4/20230355

Mim, F. I., Islam, Md. J., & Abdullah, M. S. (n.d.). Plastic tsunami: Bangladesh's maritime ecosystem under siege. Environmental Forensics, 0(0), 1–3. https://doi.org/10.1080/15275922.2024.2330026

Mogane, L. K., Masebe, T., Msagati, T. A. M., & Ncube, E. (2023). A comprehensive review of water quality indices for lotic and lentic ecosystems. Environmental Monitoring and Assessment, 195(8), 926. https://doi.org/10.1007/s10661-023-11512-2

Mueller, J., Varadharajan, C., Wu, Y., & Siirila-Woodburn, E. (2021). Machine learning to enable efficient uncertainty quantification, data assimilation, and informed data acquisition (AI4ESP1097). Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States). https://doi.org/10.2172/1769743

Oreški, D., Pihir, I., & Višnjiú, D. (2023). Comparative analysis of machine learning algorithms on data sets of different characteristics for digital transformation. 2023 46th MIPRO ICT and Electronics Convention (MIPRO), 1428–1433. https://doi.org/10.23919/MIPRO57284.2023.10159910

Rahman, H., Easha, A. A., Fatema, N., Islam, Md. J., & Alam, M. (2024). Climate change adaptation strategy of the coastal indigenous community of Bangladesh. Advances in Civil Engineering, 2024(1), 5395870. https://doi.org/10.1155/2024/5395870

Ren, Z., & Du, C. (2023). A review of machine learning state-of-charge and state-of-health estimation algorithms for lithium-ion batteries. Energy Reports, 9, 2993–3021. https://doi.org/10.1016/j.egyr.2023.01.108

Rezaie-Balf, M., Attar, N. F., Mohammadzadeh, A., Murti, M. A., Ahmed, A. N., Fai, C. M., Nabipour, N., Alaghmand, S., & El-Shafie, A. (2020). Physicochemical parameters data assimilation for efficient improvement of water quality index prediction: Comparative assessment of a noise suppression hybridization approach. Journal of Cleaner Production, 271, 122576. https://doi.org/10.1016/j.jclepro.2020.122576

Schweitzer, R. W., Harvey, B., & Burt, M. (2020). Using innovative smart water management technologies to monitor water provision to refugees. Water International, 45(6), 651–659. https://doi.org/10.1080/02508060.2020.1786309

Shadabi, L., & Ward, F. A. (2022). Predictors of access to safe drinking water: Policy implications. Water Policy, 24(6), 1034–1060. https://doi.org/10.2166/wp.2022.037

Sillberg, C., Kullavanijaya, P., & Chavalparit, O. (2021). Water quality classification by integration of attribute-realization and support vector machine for the Chao Phraya River. Journal of Ecological Engineering, 22(9), 70–86. https://doi.org/10.12911/22998993/141364

Sutadian, A. D., Muttil, N., Yilmaz, A. G., & Perera, B. J. C. (2015). Development of river water quality indices—A review. Environmental Monitoring and Assessment, 188(1), 58. https://doi.org/10.1007/s10661-015-5050-0

Tabassum, S., Kotnala, C. B., Masih, R. K., Shuaib, M., Alam, S., & Alar, T. M. (2023). Performance analysis of machine learning techniques for predicting water quality index using physiochemical parameters. 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), 372–377. https://doi.org/10.1109/ICSCSS57650.2023.10169408

To, T. C. (2020). Water quality assessment of Saigon River for public water supply based on water quality index. Vietnam Journal of Science and Technology, 58(5A), 85. https://doi.org/10.15625/2525-2518/58/5A/15203

Uddin, M. G., Nash, S., Mahammad Diganta, M. T., Rahman, A., & Olbert, A. I. (2022). Robust machine learning algorithms for predicting coastal water quality index. Journal of Environmental Management, 321, 115923. https://doi.org/10.1016/j.jenvman.2022.115923

Uddin, Md. G., Nash, S., & Olbert, A. I. (2021). A review of water quality index models and their use for assessing surface water quality. Ecological Indicators, 122, 107218. https://doi.org/10.1016/j.ecolind.2020.107218

Wang, L., Zhu, Z., Sassoubre, L., Yu, G., Liao, C., Hu, Q., & Wang, Y. (2021). Improving the robustness of beach water quality modeling using an ensemble machine learning approach. Science of The Total Environment, 765, 142760. https://doi.org/10.1016/j.scitotenv.2020.142760

Yilma, M., Kiflie, Z., Windsperger, A., & Gessese, N. (2018). Application of artificial neural network in water quality index prediction: A case study in Little Akaki River, Addis Ababa, Ethiopia. Modeling Earth Systems and Environment, 4(1), 175–187. https://doi.org/10.1007/s40808-018-0437-x

Zhai, C., Sui, Y., & Wu, W. (2023). Machine learning-assisted correlations of heat/mass transfer and pressure drop of microchannel membrane-based desorber/absorber for compact absorption cycles. International Journal of Heat and Mass Transfer, 214, 124431. https://doi.org/10.1016/j.ijheatmasstransfer.2023.124431

Zhang, Y., Gao, X., Smith, K., Inial, G., Liu, S., Conil, L. B., & Pan, B. (2019). Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network. Water Research, 164, 114888. https://doi.org/10.1016/j.watres.2019.114888