Prediction of Water Quality Index (WQI) Using Machine Learning

Kunyanuth Kularbphettong [1,*], Nareenart Raksuntorn [1], and Chongrag Boonseng [2]

[1] Faculty of Science and Technology, Suan Sunandha Rajabhat University, Bangkok, Thailand
[2] School of Engineering department, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand
Email: kunyanuth.ku@ssru.ac.th (K.K.); nareenart.ra@ssru.ac.th (N.R.); chongrag.bo@kmitl.ac.th (C.B.)
*Corresponding author
Manuscript received August 20, 2024; revised October 8, 2024; accepted October 15, 2024; published January 20, 2025

***Abstract*—The purpose of this project is to assess Water Quality Index (WQI) by using five machine learning techniques including the Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DF), Extreme Gradient Boosting (XGBoost), and Adaptive Boosting (AdaBoost). In this case, we are using Thailand as a base country for assessing water quality of the rivers and canals. The data set was collected from Bangkok Metropolitan Authority of Thailand during the period January 2018 to January 2021. The data set included 43,776 records and each record comprised 12 quantitative measurements related to water quality. Hence, they were used as feature inputs of the assessment model; for instance, pH, DO (Dissolved Oxygen), BOD (Biochemical Oxygen Demand), TP (Total Phosphorus), TCB (Total Coloniform Bacteria), FCB (Fecal Coloniform Bacteria), NO_3-N (Nitrogen-Nitrogen), No_2-N (Nitrogen-Suspended Solid), NH_3-N (Ammonia-Nitrogen), TS (Total Solid), and Total Dissolved Solid (TDS). During the phase of preprocessing K-Nearest Neighbors (KNN) and Random Forest were employed to handle missing data and detecting outliers. KNN imputation was applied to address missing values, while Random Forest was implemented to eliminate outliers, so generating the dataset appropriate for model training. The effectiveness of each machine learning model was assessed employing four principal metrics: accuracy, precision, recall, and F1 score. The findings revealed that all five methodologies excelled in predicting WQI; however, the XGBoost model surpassed the others, attaining the highest values across all metrics, including an accuracy of 91%.**

***Keywords*—water quality index, machine learning, data imputation, SVM, random forest, decision tree, XGBoost**

I. Introduction

Water is one of the most essential necessities for life, and it also plays an important role in human societies; for example, agricultural sectors and all aspects of socio-economic development all need water as the main key for productivity. The quality of water is crucial for the well-being of ecosystems, human health, and various economic activities [1]. The effects of poor water quality can be wide-ranging and impact different aspects of the environment and society. Various factors contribute to the problems of water quality, leading to contamination and degradation of freshwater resources, and water source degradation refers to the decline in the quality and availability of water from natural sources such as rivers, lakes, aquifers, and reservoirs. Various human activities and environmental factors contribute to the degradation of water sources. Water quality is essential for mitigating environmental problems and ensuring the health of ecosystems, human populations, and economies.

Implementing effective water management practices, reducing pollution, and promoting sustainable water use are critical steps in preserving water quality and the overall health of the environment [2]. Therefore, water quality is considered one of the most important environmental problems.

Water Quality Index (WQI) is a numerical expression that represents the overall general quality of water in a particular location. This index is useful for assessing and communicating the quality of water to the public, policy makers, and scientists [3]. In the process of calculating WQI, we have to consider setting parameters in the calculation. These parameters can be varied; however, they often include physical characteristics, chemical characteristics, biological characteristics, and additional parameters. According to the Water Quality Management Bureau, Pollution Control Department [4, 5] of Thailand, there are five parameters: Dissolved Oxygen (DO), Biological Oxygen Demand (BOD), Total Coliform Bacteria (TCB), Fecal Coliform Bacteria (FCB), and Ammonia Nitrogen (NH_3-N) that are used for the assessment of WQI. The WQI illustrates five levels of water quality: very good (91–100), good (71–90), average (61–70), poor (31–60), and very poor (0–30).s

Machine learning is a field of Artificial Intelligence (AI) that focuses on the development of algorithms and models. They enable computers to learn from and make predictions and decisions based on available data. Applying machine learning techniques to predict WQI involves using algorithms to analyze historical water quality data. Not only analyze the data, but also develop the model that can assess the index based on various water quality parameters. Machine learning models can capture complex relationships within the data and provide accurate predictions, especially if trained on high-quality and representative data sets.

Currently, there are machines and techniques that have been developed for evaluating the quality of water. The integration of a Support Vector Machine (SVM) and WQI can be successfully achieved for the assessment of groundwater quality (0.90 accuracy). The integrated approach presented a lower percentage of the area in the excellent class and estimated groundwater quality with high accuracy values [6, 7]. The combination of Machine Learning (ML) with remote sensing data substantially improves the accuracy of global climate model predictions. This method utilizes sophisticated algorithms to enhance data assimilation, cloud modeling, and overall forecast precision. Five machine learning classifier algorithms, including SVM, Naive Bayes (NB), Random Forest (RF), K-Nearest Neighbor (KNN), and Gradient Boosting (XGBoost), were applied to determine the best classifier for assessing water quality classes. XGBoost and KNN were both demonstrated to be better in correcting classified water quality [8]. SVM and XGBoost were applied to assess WQI from seven parameters, and they yield accurate

predictions of water classification. The performance of the XGBoost model was better than that of the SVM. The accuracy obtained from XGBoost was about 94%, while that from SVM was 67%. Hence, XGBoost is a preferred choice for water quality classification [9]. According to [10], ensemble models greatly improve accuracy and outperform simple decision tree solutions, especially when they combine a large number of trees, as in the case of random forests and boosting methods. In fact, many Thai researchers have previously studied methods for predicting the water quality in a range of contexts, such as water bodies, and have identified some of these methods—such as [11–15].

The premise of this work proposes that machine learning models can accurately predict water quality outcomes leveraging the 12 quantitative water quality measurements obtained from five stations along the Chao Phraya River. We hypothesize that models like Random Forest and Support Vector Machines will accurately classify water quality levels, pinpointing important predictors such as BOD, DO, and TDS as the most significant. Furthermore, we anticipate the models to identify temporal variations, namely seasonal fluctuations in pollutant concentrations, with higher values observed during the rainy season as a result of increased runoff. The models' performance will be assessed using measures like accuracy, precision, recall, and F1-score to validate the hypothesis. This research aids decision-makers in comprehending the temporal effects of diverse pollutants, while prediction models may inform policy interventions, prioritize remediation sites, and enhance resource allocation to tackle water quality issues.

II. Data And Methods

A. Data Source

The dataset was gathered from five monitoring stations situated along the Chao Phraya River: Wat Krang, Wat Sopharam, Wat Chang, Wat Choeng Len, and the Royal Irrigation Department station (Fig. 1). These stations denote distinct parts of the river, each displaying diverse degrees of urban, industrial, and agricultural activities that may affect water quality. Consequently, geographic diversity may add biases into the data, thereby influencing generalizations regarding the overall water quality of the river.

The data was gathered from January 2018 to January 2021, encompassing both dry and rainy seasons. The Chao Phraya River has seasonal oscillations, resulting in considerable variations in water quality measures such as turbidity, dissolved oxygen, and nutrient levels across different seasons. There were 43,776 records and 12 quantitative measurements related to water quality included pH, DO, BOD, TP, TCB, FCB, NO_3-N, No_2-N, NH_3-N, TS, and TDS as shown in Table I. The results from this dataset have pragmatic implications for water quality management along the Chao Phraya River. By concentrating on critical metrics such as pH, dissolved oxygen, and nutrient concentrations, authorities can formulate precise monitoring plans, particularly during high-risk intervals like the rainy season, when runoff from urban and agricultural regions typically escalates. These findings may assist in prioritizing river regions for enhanced water quality regulations. Also, the study offers a comprehensive dataset comprising 43,776 records and 12 distinct water quality metrics, reflecting varied environmental circumstances. The incorporation of critical indicators such as pH, dissolved oxygen (DO), biochemical oxygen demand (BOD), and diverse nutrient concentrations renders the system applicable to different field-scale scenarios necessitating comprehensive, multi-variable water quality assessment. This comprehensive dataset facilitates scalability for field investigations requiring continuous, multi-parameter data gathering to accurately assess water systems.

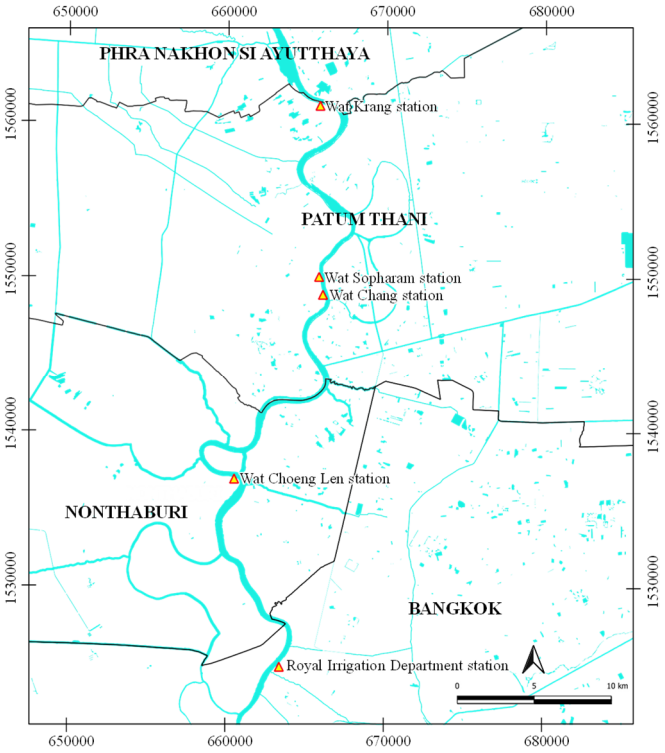

Fig. 1. Location of the study area.

B. Data Pre-Processing

Data pre-processing is an important step for data analysis and machine learning. In general, quality of data affects to algorithms' performance and highly dependent on data pre-processing schemes. In most data pre-processing step, raw data will be cleaned and transformed into a format that can be easily understood and utilized by algorithms. Other tasks include handling missing data, dealing with outliers, data normalization and standardization, and data splitting.

1) KNN and RF Imputations

K-Nearest Neighbors (KNN) and Random Forest (RF) are two notable approaches for imputing missing data. KNN imputation performs by predicting a missing value using the nearest available data values. The algorithm specifically detects a collection of encompassing points that demonstrate the greatest similarity (in terms of distance) to the data point with the missing value, and subsequently estimates the missing value based on the average or weighted average of these neighbors. KNN is more effective when the data reveals local similarity or clustering patterns, rendering it a dependable technique for imputing missing values in environmental datasets, such as water quality data. Fig. 2 illustrates the KNN imputation methodology, wherein each input variable (e.g., pH, TDS, DO) is depicted as a column vector, and missing values are estimated based on the nearest accessible neighbors within the dataset.

KNN Imputation

1. Initialization: Copy the original dataset to avoid altering it directly.

2. Loop over Features: Process each feature (column) separately. For this dataset, features include pH, TDS, DO, BOD, and related features.

3. Identify Missing Values: For each feature, find the indices where the values are missing.

4. Calculate Distances: For each missing value, compute the Euclidean distance to all other data points that have non-missing values. The distance is calculated using all features that are not missing for both the target point and the candidate points..

5. Sort Distances: Identify the k-nearest neighbors based on the smallest distances.

6. Impute Missing Value: Replace the missing value with the mean of the k-nearest neighbors' values for the same feature. For categorical features, use the mode instead of the mean.

7. Return Imputed Data: Return the dataset with all missing values imputed.

Fig. 2. The standard KNN imputation's procedure.

Random Forest Imputation

1. Preprocessing:
- **Data Collection:** Gather dataset containing the variables: include pH, TDS, DO, BOD, and related features.
- **Identify Missing Values**: Mark the missing values in each column.

2. Initial Imputation:
- **Simple Imputation:** Impute missing values with simple statistics (e.g., mean, median).

3. Random Forest Imputation:
- **Iterative Imputation Process:**
  1. **Order the Columns:** Arrange the columns based on the number of missing values in ascending order.
  2. **Impute Each Column:**
     - **Select the Target Column:** Identify the current column with missing values to be imputed.
     - **Select Feature:** Use the remaining columns as features to predict the missing values in the target column.
     - **Train the Random Forest:**
       - Split the dataset into two parts: rows where the target column has values and rows where it does not.
       - Use the rows with known values to train a Random Forest regression model, with the target column as the output and the other columns as input features.
     - **Predict Missing Values:**
       - Use the trained Random Forest model to predict the missing values in the target column.
       - Replace the initial imputed values with the predictions.
  3. **Repeat:** Iterate over all columns with missing values.
  4. **Convergence Check:** Continue iterating through the columns until the imputed values stabilize.

4. Post-Imputation:
- **Validation:** Validate the imputation by checking the coherence and consistency of the imputed values.
- **Analysis**: Perform any additional analysis or use the imputed dataset.

Fig. 3. RF imputation's procedure.

Random Forest imputation (Fig. 3) is an accurate approach for addressing missing data. Random Forest function produce multiple decision trees to estimate missing values, exploiting other obtainable factors in the dataset to predict the missing ones. Random Forest is recognized for its resilience to outliers and its capacity to manage non-linear relationships in the data. Each input variable in the dataset is regarded as a column vector, and missing values are calculated based on the similarity of observations distributed across the trees. Random Forest is especially advantageous for analyzing intricate datasets characterized by a combination of linear and non-linear connections among variables.

Fig. 4 compares the results of KNN and RF imputation with the original data. The distribution graphs indicate that both imputation approaches yield results that closely resemble the original data. To assess the precision of these imputations, we employed Mean Square Error (MSE), a commonly used performance metric that quantifies the average squared deviation between the actual values and the imputed values. Reduced MSE values signify enhanced imputation accuracy. Our investigation indicated that KNN imputation produced a smaller MSE than RF, illustrating that KNN was more proficient in preserving the structure and accuracy of the water quality data. Besides addressing missing data, Random Forest was employed for outlier detection. Owing to its ensemble characteristics and capacity to manage intricate, non-linear data, RF is particularly adept at detecting outliers within the dataset. Outliers were determined using the residuals (the disparity between predicted and actual values) produced by the Random Forest model. Data points that displayed significantly large residuals, relative to the remainder of the dataset, were identified as probable outliers. The identified outliers were subsequently either eliminated or modified based on their impact on the dataset and the comprehensive study. Consequently, KNN was employed for imputing missing values, while RF was utilized for outlier detection, so assuring that the dataset utilized in this study was both precise and representative, devoid of the impact of missing or erroneous data points.

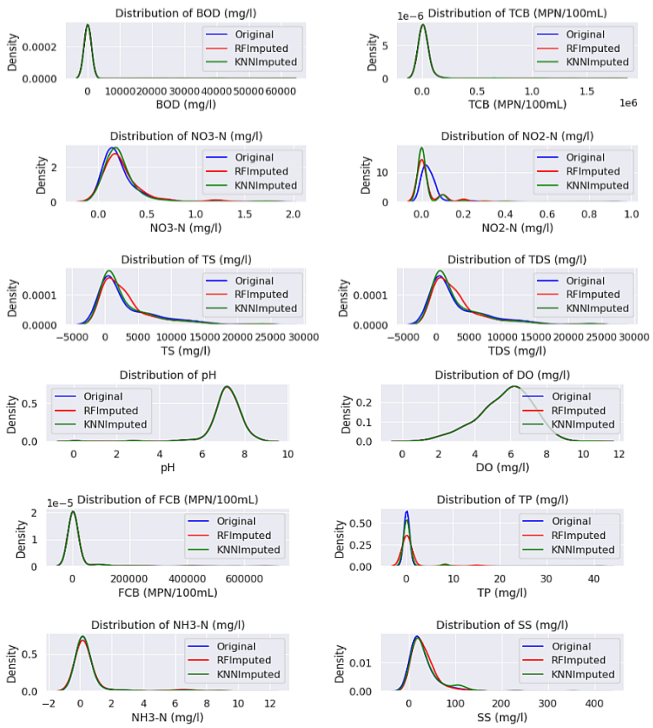

Fig. 4. The Comparison of RF and KNN Imputation.

Table 1. Descriptive statistic of features

Features	Units	Mean	STD
pH	pH unit	7.12226	0.8788
DO	mg/l	5.7283	1.4927
BOD	mg/l	406.5947	4044.2733
TCB	(MPN/100mL)	3.862942e+04	1.526422e+05
FCB	(MPN/100mL)	16252.7445	61145.7514
TP	mg/l	0.4690	2.5958
NO_3-N	mg/l	0.2300	0.2102
NO_2-N	mg/l	0.0389	0.0606
NH_3-N	mg/l	0.6236	1.5700
SS	mg/l	35.9787	41.6093
TS	mg/l	3192.9549	4412.9103
TDS	mg/l	3167.0765	4457.4754

After imputation, the correlation coefficients between the variables in a dataset are displayed graphically in a

correlation heat map as shown in fig 5, which is a depiction of the relationship among features via correlation coefficients. The intensity and direction of the correlations between pairs of variables are shown by these coefficients, which have a range of -1 to 1. Mean and standard deviation (STD) of all the features are shown in Table 1. It is obviously seen that values of features are great different. Hence, we normalized all the features so that no feature was dominated by strong features.

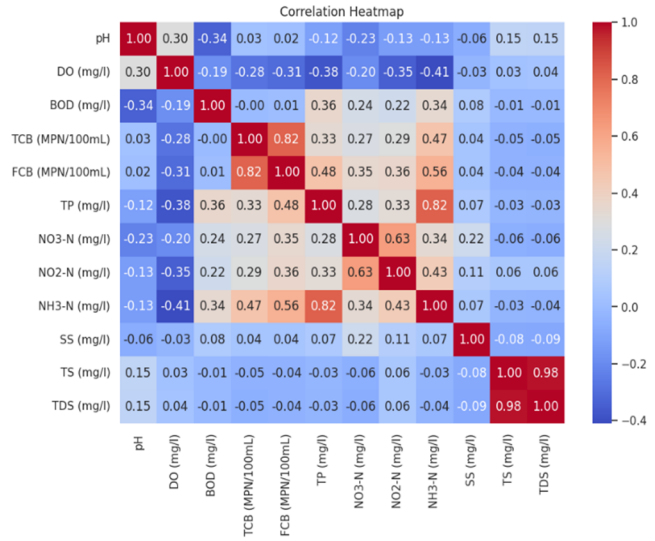

Fig. 5. Correlation Heat map.

2) Feature scaling by min-max normalization

Features are scaled to a range between 0 and 1 using min-max normalization.

$$v' = \frac{v - min_A}{max_A - min_A} \quad (1)$$

where A is a feature vector, v' is a normalized value, v is a feature value, min_A and max_A are the minimum and maximum values of the feature vector, respectively.

C. Research Methodology

In this section, we present a method used to calculate WQI and describe a brief summary of related techniques.

1) Water Quality Index (WQI) calculation

In order to calculate the Water Quality Index (WQI), 12 parameters involved water quality, like pH, DO, BOD, TCB, FCB, TP, NO_3-N, NO_2-N, NH_3-N, SS, and TS, were used to conduct as the inputs of the model for estimating WQI. According to Scottish Development Department (SDD), WQI is calculated in 4 states as follows:

1. Normalize Each Parameter: Each water quality criterion is standardized to a similar scale, ensuring that elevated values indicate better water quality. Normalization transforms the raw data to enable comparisons among parameters measured in disparate units on a uniform scale.

2. Assign weights to parameters: Each water quality requirements is given a weight that reflects how important it is to overall water quality. The weights are often established by regulatory standards or expert assessments. Dissolved Oxygen (DO) may carry greater significance than Total Dissolved Solids (TDS), as DO is essential for aquatic organisms. These weights can be obtained from established criteria (e.g., WHO or local environmental standards) or by consultations with environmental specialists.

3. Calculate the sub-index for each parameter: The sub-index of parameter is calculated by the multiplication of its weight and normalized value

4. Sum the sub-indices: The total Water Quality Index (WQI) is calculated by summing the sub-indices for all metrics.

$$WQI = \frac{\sum_{i=1}^{N} q_i \times w_i}{\sum_{i=1}^{N} w_i} \quad (2)$$

where N is the total number of features, q_i is the actual value of the parameter (WHO water quality standards [16]), w_i is the parameter's suggested standard value that is suitable for parameter's value in water.

Table 2. Description of WQI level

Scores	Water Quality Level	Water Quality Index (WQI) Classification
85-100	Excellent	1
70-84	Good	2
50-69	Fair	3
20-49	Bad	4
0-19	Worse	5

The sum of all of the sub-indices produces the overall Water Quality Index (WQI), providing an exclusive numerical value that reflects all aspects of water quality. The Water Quality Index (WQI) is typically categorized into classifications like Excellent, Good, Fair, Poor, and Very Poor to interpret levels of water quality. The water quality classification presented in Table II conforms to current water quality criteria established by relevant environmental authorities. These classifications may differ by area and are regularly revised to incorporate developments in water quality science. This study's categories were modified according to the rules established by the Thailand Pollution Control Department (PCD) and international organizations such as the World Health Organization (WHO).

2) Support Vector Machines (SVM)

SVM is one of effective supervised machine learning approaches for regression and classification. SVM is a well-known technique for classification with highly dimensional environments, especially when the number of samples is less than its dimension. The objective function for class i is written as,

$$\min_{\mathbf{w}_i, b_i, \xi_i} \frac{1}{2}\|\mathbf{w}_i\|^2 + C\sum_{j=1}^{n} \xi_{i,j}, \quad (3)$$

subject to:

$$y_{i,j}(\mathbf{w}_i \cdot \phi(x) + b_i) \geq 1 - \xi_{i,j} \quad (4)$$

$\square_{i,j} \square 0$ for $\square_j$,

where w_i, b_i are the weight vector and bias for the i-th classifier, respectively. $\xi^{i,j}$ are slack variables. $y_{i,j}$ =1 if the j-th sample belongs to class i, otherwise $y_{i,j}$= −1.

Decision Function:

$$\hat{y} = \underset{i}{\text{argmax}}(\mathbf{w}_i \cdot \phi(x) + b_i) \quad (5)$$

3) Decision tree

The objective of WQI classification using DT technique is to establish the decision model resembling a tree based on water quality indicators and to categorize water samples into multiple categories, i.e., excellent, good, fair, poor, and very

poor. In order to produce distinctive subsets with regarding to the target variable, DT divides the data into subsets based on feature values.

Information Gain (IG): IG measures the quality of class, in other word; larger IG means the class is more purity. IG is a difference between the entropy before and after class splitting. Entropy (H) and IG can be computed by Eq. (6) and Eq. (7), respectively.

$$H(D) = \sum_{i=1}^{k} p_i \, log_2 p_i \tag{6}$$

where p_i is the proportion of samples in class i

Information gain IG for a split on attribute A is:

$$IG(D, A) = H(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} H(D_v) \tag{7}$$

where D_v is the subset of D where attribute A has value v.

4) Random forest

RF technique is an ensemble learning method producing a class that is the model of the classes (classification) of the individual trees. To increase generalization and robustness, it integrates the predictions of many base estimators. The important concepts of RF are Ensemble Learning, Bootstrap Aggregation (Bagging) and Feature Randomness.

Bootstrapping: Generate B bootstrap samples D_b from the training set D.

Growing Trees:

- For each bootstrap sample D_b
  - Grow a decision tree T_b by recursively splitting the nodes:
    - At each node, randomly select a subset of m features.
    - Split the node using the feature that handles the best split according to a given criterion.

Prediction:

- For classification, each tree T_b outputs a class prediction $\widehat{y_b}$ for an input x.
- The final prediction $\hat{y}$ is the majority vote:

$$\hat{y} = mode(\{\widehat{y_1}, \widehat{y_2}, \ldots, \widehat{y_B}\}) \tag{8}$$

5) XGBoost (Extreme Gradient Boosting)

In supervised learning circumstance, XGBoost is an effective gradient boosting technique. It integrates the predictions of multiple weak learners, typically decision trees. XGBoost can be used to categorize water quality into various categories in the context of WQI classification.

Objective Function: The regularization term and loss function are the components of the objective function in XGBoost. The regularization term penalizes the model's complexity to prevent overfitting, while the loss function indicates how well the model predicts the target variable. Objective function can be calculated by Eq. (9).

$$Obj(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \tag{9}$$

where L is the loss function, Ω is the regularization term, and f_k are the trees in the model.

Additive Model: At each step t, a new tree f_t is added to minimize the objective function.

$$\hat{y_i}^{(t)} = \hat{y_i}^{(t-1)} + f_t(x_i) \tag{10}$$

Gradient Descent: Gradient descent is used by XGBoost to reduce the loss function. New trees are fitted to the gradients of the loss function with respect to the predictions, which are computed.

6) AdaBoost (Adaptive Boosting)

AdaBoost is an ensemble learning technique that constructs an efficient classifier through the integration of several weak classifiers, which are usually decision trees. The objective is to prioritize the challenging instances in subsequent classifiers and concentrate on the mistakes committed by prior classifiers. AdaBoost can be used to categorize water quality into distinct groups according to a range of metrics in the context of WQI classification system.

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t \, h_t(x)) \tag{11}$$

Compute the weighted error of the weak learner:

$$\epsilon_t = \frac{\sum_{i=1}^{n} w_i^{(t)} I(y_i \neq h_t(x_i))}{\sum_{i=1}^{n} w_i^{(t)}} \tag{12}$$

Compute the weight of the weak learner:

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1-\epsilon_t}{\epsilon_t}\right) \tag{13}$$

Update the weights of the training samples:

$$w_i^{(t+1)} = w_i^{(t)} \exp\left(-\alpha_t y_i h_t(x_i)\right) \tag{14}$$

III. Experimental Results

To provide a comprehensive comparison of the results of WQI predictions using SVM, Random Forest, Decision Tree, XGBoost, and AdaBoost, let's outline a typical analysis and results presentation. Preparing the data, training the model, evaluating the metrics, and discussing the outcomes are all included in this. After running the processes, the results for each model were shown and the metrics for comparison include Accuracy, Precision, Recall, and F1 Score as presented in table III and plotted in Fig. 6.

The seasonal fluctuations in the data, resulting from alterations in the Chao Phraya River's water quality during each season of the year, were accurately represented by these machine learning models, illustrating the relevance of machine learning for intricate, dynamic environmental data.

The Table III demonstrates the performance indicators of five machine learning algorithms—SVM, Random Forest, Decision Tree, XGBoost, and AdaBoost—in predicting the Water Quality Index (WQI). XGBoost achieved the highest accuracy (0.91), precision (0.92), recall (0.90), and F1 score (0.91) among the hypotheses, demonstrating its strength in classifying WQI levels. This exceptional performance is a result of XGBoost's potential to identify intricate non-linear relationships in the data using gradient boosting, which progressively enhances the model by rectifying errors from prior iterations. On the other hand, though the Decision Tree revealed the lowest performance across all criteria (accuracy:

0.81, F1 score: 0.83), its simplicity and interpretability could render it an appropriate tool for preliminary exploratory investigation or for stakeholders necessitating comprehensible models. On the contrary, Random Forest, which integrates numerous decision trees to enhance predictive efficacy, attained a commendable equilibrium between accuracy (0.89) and interpretability.

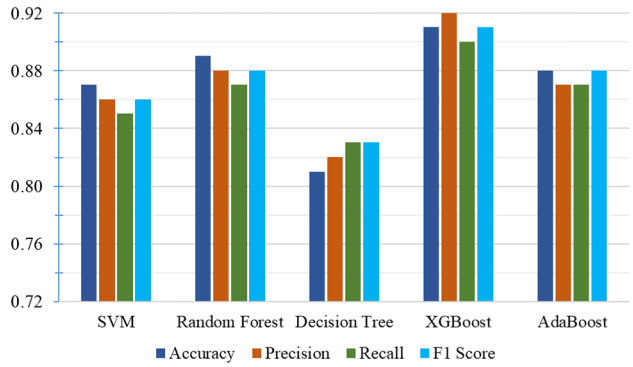

Fig. 6. Visualization of results.

Table 3. The experimantal results

Model	Accuracy	Precision	Recall	F1 Score
SVM	0.87	0.86	0.85	0.86
Random Forest	0.89	0.88	0.87	0.88
Decision Tree	0.81	0.82	0.83	0.83
XGBoost	0.91	0.92	0.90	0.91
AdaBoost	0.88	0.87	0.87	0.88

SVM and AdaBoost displayed comparable performance, obtaining F1 values of 0.86 and 0.88, respectively. The performance of SVM indicates its ability for handling high-dimensional spaces while recognizing complex data patterns, whereas AdaBoost's ability to concentrate on misclassified cases enhances its predictive accuracy in subsequent rounds. Both approaches are productive in contexts where accuracy is essential, such as detecting water sources necessitating treatment. The final choice of an approach for WQI prediction is dependent on the specific requirements of the work. To optimize prediction performance and identify intricate patterns in the data, XGBoost may be the optimal choice. Nevertheless, in real-time monitoring systems where computing resources and model interpretability are crucial Random Forest or AdaBoost may provide more pragmatic solutions.

The decision of a machine learning model in water quality management may lead to substantial practical consequences. High-performing algorithms such as XGBoost can deliver accurate predictions that assist authorities in promptly identifying regions suffering water quality degradation, facilitating corrective measures. Conversely, models like as Random Forest and AdaBoost, however somewhat fewer precise, may provide enhanced computational efficiency and increased transparency, which are crucial for regular tracking and providing information to stakeholders who are not technical. Moreover, the recall of the models is particularly vital in detecting substandard water quality, as it signifies the capacity to accurately classify cases where the Water Quality Index suggests potential health hazards. XGBoost's elevated recall (0.90) indicates it's successful in detecting instances necessitating prompt intervention to mitigate water contamination.

The XGBoost model is the most effective model in this investigation, owing to its capacity to manage the dataset's complexity and nonlinearity, effectively capturing both seasonal and regional fluctuations in water quality. Its strong performance indicators (accuracy, precision, recall, and F1 score) demonstrate its efficacy in minimizing errors, rendering it a dependable option for forecasting the Water Quality Index and facilitating environmental monitoring goals. The model's benefits, including its management of absent data, scalability, and ensemble learning methodology, render it suitable for use in dynamic and varied hydrological contexts such as the Chao Phraya River.

IV. Conclusion

WQI is a typical indicator to evaluate the quality of water in rivers and canals. WQI has a great impact on assessing public health and also environmental management. Several computation techniques have been investigated for WQI estimation. Machine learning can be beneficial in enhancing WQI prediction; also, it helps interpret the variables influencing quality of water. A high level of WQI has significant effects on many scales such as economy, society, environment, and many more. For example, when clean water is provided for a community, there will be improvement on hygiene and health, food production, education, and most importantly; there is less damage to the environment. On the other hand, a lower level of WQI can cause many hazards like health issues. Thus, there are tons of health issues that will occur if the water is contaminated; for instance, infections, chronic illnesses, gastrointestinal disorder, etc. are some of health issues that come with low level of WQI. There are various machine learning techniques that can improve WQI estimation, and offer deeper insight of water quality management. In this work, we used SVM, RF, DT, XGBoost, and AdaBoost techniques to predict WQI. These techniques can be applied to predict modeling, anomaly detection, and parameter optimization. Each model has its strengths and weaknesses, and the preferred method will be decided base on the particular need of the application and the accuracy and efficiency of the outcomes. In this work, XGBoost provided the best accurate results of WQI. Therefore, it is important to understand the reason why XGBoost method is the best method to use in assessing WQI, so that it can be used effectively and efficiently.

To sum up, WQI will keep growing as long as data collection and machine learning algorithms continue to progress. Water quality management may undergo a revolution due to merging of AI and WQI, which will provide more sophisticated capabilities and wider applications. We could potentially expect better monitoring system, predictive capacity, and decision making process as a result of improving AI, IoT, and data analysis. This will then ultimately result in better reservoirs, public health, and environmental sustainability.

Conflict of Interest

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

KL conducted the research; NR analyzed the data; CB guided in the whole research; and all authors had approved the final.

FUNDING

This research was funded by Suan Sunandhe Rajabhat University.

ACKNOWLEDGMENT

Authors are grateful to the enormous guidance and support received from Suan Sunandha Rajabhat University.

REFERENCES

[1] D. Grey and C. Sadoff, "Sink or swim? Water security for growth and development," *Water Policy,* vol. 9, no. 6, pp. 545–571, 2007.

[2] World Health Organization. (2017, Feb. 14). Water quality and health -Review of turbidity: Information for regulators and water suppliers. [Online]. Available: https://www.who.int/publications/i/item/WHO-FWC-WSH-17.01

[3] W. Simachaya. (2020, Aug. 22). *Water Quality Management in Thailand.* [Online]. Available: https://www.pcd.go.th/wp-content/uploads/2020/04/pcdnew-2020-04-22_08-59-15_424617.pdf

[4] Pollution Control Department, Report on the operation of the Water Quality Management Division, Ministry of Natural Resources and Environment (only Thai), Bangkok, Thailand, Nov. 2020.

[5] Pollution Control Department. (2022, Aug. 23). Water quality index: WOI. [Online]. Available: https://www.pcd.go.th/wp-content/uploads/2022/08/pcdnew-2022-08-23_03-47-16_304672.pdf

[6] S. A. Abu El-Magd, I. S. Ismael, M. A. Sh El-Sabri, M. S. Abdo, and H. I. Farhat, "Integrated machine learning–based model and WQI for groundwater quality assessment: ML, geospatial, and hydro-index approaches," *Environmental Science and Pollution Research*, vol. 30, no. 18, pp. 53862–53875, 2023, doi: 10.1007/s11356-023-25938-1.

[7] S. Koley, "Augmenting efficacy of global climate model forecasts: machine learning appraisal of remote sensing data," *International Journal of Engineering Trends and Technology*, vol. 72, no. 6, pp. 442–502, 2024, doi: https://doi.org/10.14445/22315381/IJETT-V72I6P139.

[8] M. G. Uddin, S. Nash, A. Rahman and A. I. Olbert, "Performance analysis of the water quality index model for predicting water state using machine learning techniques," *Process Safety and Environmental Protection*, vol. 169, p. 828, 2023, doi: https://doi.org/10.1016/j.psep.2022.11.073.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735. PMID: 9377276.

[10] H. I. H. Yusri, A. A. A. Rahim, S. Hassan, I. Halim and N. E. Abdullah, "Water quality classification using SVM And XGBoost method," in *Proc. 13th IEEE Control and System Graduate Research Colloquium (ICSGRC)*, Shah Alam, Malaysia, Jul. 2022, pp. 231–236, doi: 10.1109/ICSGRC55096.2022.9845143.

[11] C. V. Sillberg, P. Kullavanijaya and O. Chavalparit, "Water quality classification by integration of attribute-realization and support vector machine for the Chao Phraya River," *Journal of Ecological Engineering*, vol. 22, no. 9, 2021.

[12] M. Wongaree, "Water quality assessment by using of water quality index for Mak Khaeng Canal, Udon Thani Province, Thailand," *Environmental Asia*, vol. 12, no. 2, pp. 96–104, 2019.

[13] W. Khanitchaidecha, "Spatial and seasonal variation in surface water quality of Nan River, Thailand," *Naresuan University Engineering Journal*, vol. 14, no. 1, Article 1, 2019.

[14] R. Sukthanapirat, S. Suttibak and P. Jaikaew, "Comparison of water quality in community and rural areas in Mekong River, Thailand by using water quality index," *Thai Environmental Engineering Journal*, vol. 31, no. 3, pp. 1–11, 2017.

[15] S. Nuanmeesri, L. Poomhiran, P. Kadmateekarun and S. Chopvitayakun, "Improving the water quality classification model for various farms using features based on artificial neural network," *TEM Journal*, vol. 12, no. 4, pp. 2144–2156, 2023, doi: 10.18421/TEM124-25.

[16] WHO (World Health Organization), *Water Quality for Drinking: WHO Guidelines*, Springer: Berlin/Heidelberg, Germany, 2011.