Applied Environmental Research
- RESEARCH FOR SUSTAINABLE PLANET -

Research Article

The Use of Machine Learning Algorithms for Water Quality Index Prediction in the Sai Gon River, Vietnam

Nguyen Thi Diem Thuy[1,2,*], Nguyen Thi Huynh Mai[1,2], Tran Quang Tra[1,2]

[1] *Faculty of Environment, University of Science, Ho Chi Minh City, Vietnam*
[2] *Vietnam National University, Ho Chi Minh City, Vietnam*

*Corresponding Email: ntdthuy@hcmus.edu.vn

Abstract

Accurate prediction of the water quality index (WQI) lays the groundwork for integrated river basins and sustainable water resource management. Recent and accelerated advances in machine learning have led to various promising applications in water quality assessment. The present study leverages the predictive performance of several ML algorithms, including extreme gradient boosting (XGB), the gradient boosting model (GBM), support vector regression (SVR), and the radial basic function (RBF), to predict the WQI at three monitoring sites on the Sai Gon River from 2015–2019. In comparison, the results indicate that the XGB model outperforms the other models when eight parameters, including DO, BOD_5, COD, N-NH_4^+, P-PO_4^{3-}, pH, temperature, and total coliforms, are input. Specifically, the XGB model exhibited the lowest error rates (RMSE = 1.630 and MAE = 0.782) and highest correlation (R^2 = 0.960 and NSE = 0.953), followed by the GBM, SVR, and RBF models. This study also revealed that model performance decreased substantially when N-NH_4^+ and P-PO_4^{3-} were removed, whereas the exclusion of COD or BOD_5 caused marginal declines in predictive capacity. These findings highlight that parsimonious ML models can minimize the parameters required for WQI prediction but still maintain satisfactory simulations and effectively capture potential relationships between input parameters and derive WQI. Generally, this study provides an analytical framework for simulating WQI based on parsimonious and accurate ML algorithms, which are conducive to water quality assessment and monitoring in developing nations.

ARTICLE HISTORY

Received: 15 Feb. 2025
Accepted: 19 May 2025
Published: 28 May 2025

KEYWORDS

Extreme gradient boosting;
Machine learning algorithms;
Sai Gon River;
Water quality index;
WQI

Introduction

The water quality index (WQI) is a vital tool in water resource management, as it is considered a simple method for assessing and monitoring water quality. Furthermore, the WQI contributes to public health protection by evaluating the suitability of water for use and facilitating the early identification of pollution sources. The WQI was initially proposed by Horton (1965) in the USA. It is calculated on the basis of a combination of physical, chemical, and biological parameters, producing a single value from 0–100. In Horton [1], the procedure for calculating the WQI involves the following key steps: (1) selection of relevant parameters; (2) conversion of raw data into a standardized scale; (3) allocation of weights to the parameters; and (4) aggregation of the subindices to derive the final WQI.

From a global perspective, the WQI has been widely implemented in practice by numerous countries and has been the subject of extensive investigations to refine and adapt various WQI models [2]. For example, in the United States, the National Sanitation Foundation pioneered the development of the NSF–WQI in 1970 [3]. By the mid-1990s, the British Columbia Ministry of Environment, Lands, and Parks introduced a new WQI to evaluate the water quality of diverse water bodies within British Columbia, Canada [4-5]. This model was subsequently refined and adopted in 2001 by the Canadian Council of Ministers of the Environment as the CCME WQI [4,

https://doi.org/10.35762/AER.2025018

6–7]. In Southeast Asia, Malaysia introduced its WQI model in 2007, which incorporates six parameters: DO, BOD, COD, NH_3-N, SS, and pH [8]. In Vietnam, the General Department of Environment – Ministry of Natural Resources and Environment (VEA–MONRE) issued Decision No. 1460/QĐ-TCMT in 2019 [9], providing technical guidelines for calculating and publishing the Vietnam Water Quality Index (VN_WQI), replacing the previous decision made in 2011. According to this decision, the VN_WQI is calculated on the basis of five parameter groups: pH, pesticide residues, heavy metals, organic and nutrient parameters, and microbiological parameters. The conventional approaches to WQI calculation involve deriving a WQI index by aggregating numerous subindices, which requires the measurement of a substantial number of water quality parameters. Moreover, traditional water quality monitoring methods involve the manual collection of water samples followed by laboratory analysis, which is both time-consuming and costly [10]. Consequently, this poses significant challenges to water quality assessment, particularly in developing countries with limited infrastructure and financial resources.

Recently, machine learning (ML)-based techniques have emerged as promising alternatives for WQI calculations in various regions. By leveraging their capacity to process complex, nonlinear relationships within data, ML algorithms can identify underlying mechanisms, enabling accurate predictions of WQI values. Hameed et al. [11] indicated that the RBF model performed more effectively than the Back-Propagation Neural Network (BPNN) model in simulating the WQI over tropical Malaysia and addressed the implications of omitting BOD from WQI predictions because of the high costs associated with analyzing this parameter. Asadollah et al. [12] applied the extra tree regression (ETR) model to predict the monthly WQI in Hong Kong's Lam Tsuen River, which uses 10 water quality parameters and achieves high accuracy (R^2 = 0.98, RMSE = 2.99). However, using all 10 input variables increased monitoring costs. To address this, the study used partial correlation to create different input combinations, identifying a reduced set of parameters (BOD, turbidity, and phosphate concentration) that still performed well. The study also compared output results and quantified input variable uncertainty via the R-factor approach, assessing how the choice of input variables affected prediction uncertainty. Bui et al. [13] applied 16 machine learning algorithms to six years (2012–2018) of monthly data from two stations in the Talar catchment (Iran) and used 10 water quality parameters. Pearson correlation coefficients were used to create 10 different input combinations. The results revealed that the fecal coliform (FC) concentration had the greatest effect on predicting the IRWQI, whereas the total solids (TS) concentration had the smallest effect. The best input combinations varied across algorithms, with variables showing very low correlations generally performing poorly. Kamyab-Talesh et al. [14] employed the support vector machine (SVM) model for WQI prediction and determined the key parameters influencing the WQI via monthly data from December 2007–November 2008 at five stations in the Sefidrud Basin, Iran. Othman et al. [15] employed artificial neural networks (ANNs) to simulate the WQI in the Klang River Basin and further analyzed the sensitivity of the WQI to the parameters needed, identifying DO and pH as the most and least influential inputs, respectively. Another important theme is a comparative analysis of various ML models for WQI prediction. For example, Mohd Zebaral Hoque et al. [16] demonstrated that the linear regression (LR) and ridge models performed most accurately in predicting the WQI among eight regression ML models. Furthermore, Raheja et al. [17] compared the predictive performance of three ML algorithms (i.e., DNN, GBM, and XGB) in analyzing the entropy water quality index (EWQI) and WQI over Haryana State and reported that the deep neural network (DNN) outperformed the other two models. Additionally, electrical conductivity (EC) was demonstrated to be the most determinant parameter, whereas pH appeared to be the least impactful input. Hussein et al. [18] employed the XGB, SVR, and K-nearest neighbor (KNN) models to assess the irrigation water quality index (IWQI) in Naama (southwest Algeria) and reported that each model exhibited unique strengths in WQI prediction for the region. In particular, XGB demonstrated high accuracy, closely aligned reference data, and low variability, whereas the SVR model generated stable and consistent predictions that closely approximated the reference data. The KNN forecasts, on the other hand, aligned well with the reference data, displaying reduced variance and standard deviation. Kamel and Eltarabily [19] applied Bayesian optimization to indicate better performance of XGB in estimating the irrigation water quality index (IWQI) than those generated by random forest (RF) and AdaBoost in El Moghra (Egypt).

In Vietnam, Khoi et al. [20] also demonstrated the superiority of XGB over the other 11 ML models in predicting the WQI in the La Buong River. Nguyen et al. [21] performed a comparative analysis of boosting algorithms (i.e., GBM and XGB) and deep learning models (i.e., RNN and LSTM) and reported the high-performing abilities of boosting models in simulating WQI. This study also employed the Bayesian model averaging (BMA) method to reduce the inputs to three parameters, much fewer than those required by conditional approaches. Recently, Lap et al. [22] integrated filtering methods into RF models to reduce the number of input variables for estimating WQI in the An Kim Hai irrigation system in Vietnam and reported that the optimal models could reduce from ten to four parameters (i.e., coliform, DO, turbidity, and TSS).

Although ML models typically require large datasets to achieve optimal performance, effective strategies

can enable high performance even with limited data. For example, a previous study [14] demonstrated that SVM has been effectively applied to improve accuracy and reduce the risk of overfitting, even with small datasets. In addition, incorporating machine learning models in hybrid models has also shown potential for improving water quality prediction performance, as noted in Bui et al. [13].

ML algorithms can effectively simulate the WQI. Most previous investigations have focused mainly on determining superior models [16–18, 20–22], defining essential and decisive factors [15, 17, 19], and reducing the number of required parameters compared with traditional methods [11, 20–22]. Nevertheless, while the WQI is widely used to assess the pollution levels of water sources through a categorical rating scale, prior studies have focused predominantly on analyzing WQI's quantitative outcomes as specific values and overlooked variations within water quality classes. In Vietnam, most previous studies have reduced the number of input parameters mainly on the basis of the outcomes of ML models (i.e., feature importance) and paid less attention to standards and regulations in the established guidelines for WQI calculation, thereby limiting practical applicability. Additionally, although there is a wide range of ML algorithms for water quality assessment, selecting the most appropriate techniques for specific regions is of concern. This study selected four ML algorithms for WQI prediction on the basis of their representation of different learning paradigms, including a neural network model (RBF), a kernel-based method (SVM), and ensemble boosting techniques (XGB and GBM), and their proven effectiveness in previous studies. Specifically, the RBF [11], SVM [12, 14, 18], XGB [18–20], and GBM [21] methods have demonstrated strong performance in similar tasks.

The present study strove to address these gaps and assess the water quality of the Sai Gon River, a vital river network in the critical economic region of southeast Vietnam, in which the river water quality has highly deteriorated due to industrial activities and massive urbanization. Specifically, by employing (parsimonious) machine learning algorithms, this study aimed to minimize the number of input variables required for calculating the WQI while adhering to the established standards and regulations. Furthermore, this study not only evaluates prediction results on the basis of individual WQI values but also incorporates an analysis within the context of WQI classification ranges. This integrated approach offers a more comprehensive and accurate evaluation of the model's effectiveness in predicting WQI, providing deeper insights into its overall performance.

Materials and data used

The study was conducted following the steps illustrated in Figure 1. Water quality data, including parameters such as BOD_5, DO, COD, temperature, $N\text{-}NH_4^-$, $P\text{-}PO_4^{3-}$, total coliform, and pH, were collected and used to calculate the VN_WQI in accordance with Decision No. 1460/QĐ-TCMT. The dataset was normalized via the StandardScaler method and divided into training and testing subsets. Several machine learning models—the RBF, XGB, GBM, and SVR—were developed and optimized via GridSearchCV. The best-performing model was selected via metrics (e.g., RMSE, MAE, R^2, and NSE) and a Taylor diagram. Scenario analyses were then conducted to predict WQI under reduced input parameters.

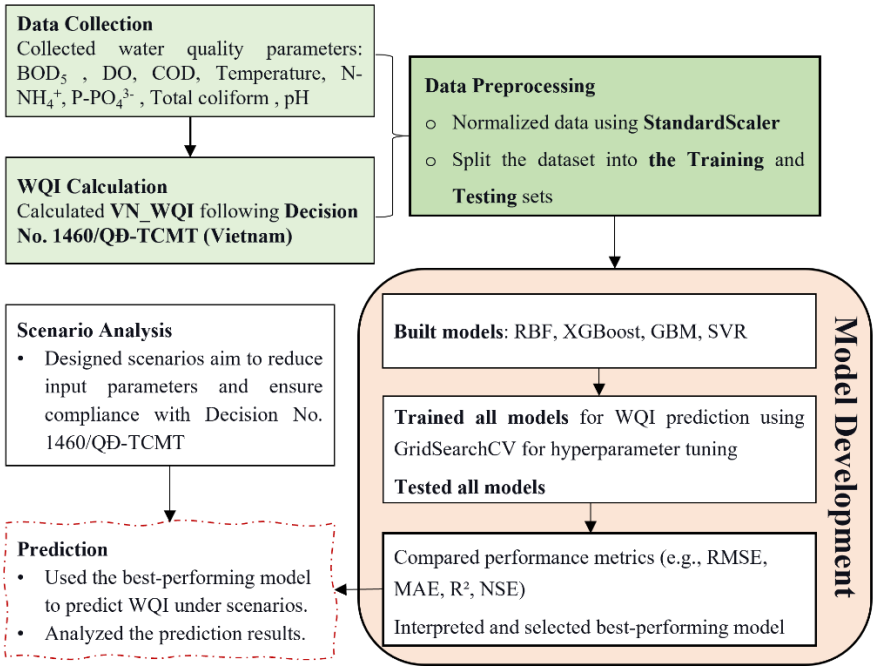


Figure 1 General flow chart of the research.

1) Study area

The study area is located in the lower basin of the Sai Gon River, at a longitude of 10°30'–11°30'N and a latitude of 106°15'–107°15'E (Figure 2). The study area covers approximately 3,200 km² and passes through the provinces of Binh Phuoc, Binh Duong, Tay Ninh, Long An, Dong Nai, and Ho Chi Minh City.

The climate of the study area is tropical monsoon, with a relatively high average annual rainfall of approximately 1,800 mm. There are two distinct seasons: the rainy season (from April to October) and the dry season (from November to March of the following year), with rainfall during the rainy season accounting for approximately 80–85% of the annual total. Additionally, the lower Sai Gon River flows through Ho Chi Minh City, Dong Nai, and Binh Duong Provinces, which are considered economically rich and dynamic areas and are among the leading economic driving forces of Vietnam now and in the coming years [23]. Under the influence of massive urbanization and industrial activities, surface water pollution has become increasingly problematic and continues to be exacerbated in this area. Therefore, it is imperative to conduct a comprehensive analysis of water quality in the Sai Gon River.

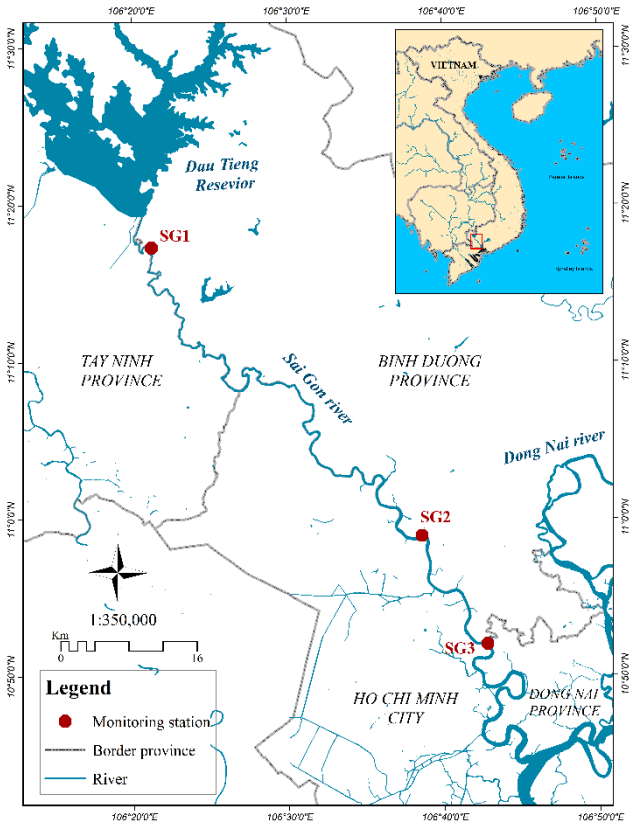


Figure 2 Geographical location of the study area.

2) WQI calculation

In this study, the WQI index was calculated on the basis of Decision No. 1460/QD-TCMT, issued by the Vietnam Environmental Administration – Ministry of Natural Resources and Environment (VEA–MONRE) in 2019, regarding the promulgation of technical guidelines for the calculation and disclosure of the Vietnam Water Quality Index (VN-WQI). As outlined in Decision No. 1460/QD-TCMT, VN_WQI requires data from at least three of the five designated parameter groups, with Group IV (the group containing organic and nutrient parameters) being a mandatory inclusion, comprising a minimum of three parameters. Specifically, Group I includes the pH parameter. Group II encompasses pesticide indicators, such as Aldrin, BHC, Dieldrin, and various DDT compounds (p,p'-DDT, p,p'-DDD, p,p'-DDE), as well as heptachlor and Hepta chlorepoxide. Group III focuses on heavy metal indicators, including arsenic (As), cadmium (Cd), lead (Pb), hexavalent chromium (Cr^{6+}), copper (Cu), zinc (Zn), and mercury (Hg). Group IV includes organic and nutrient indicators, incorporating measures such as dissolved oxygen (DO), 5-day biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), total organic carbon (TOC), and nitrogen compounds ($N–NH_4$, $N–NO_3$, $N-NO_2$) and phosphate ($P–PO_4$). Finally, Group V included microbial indicators such as coliform and E. coli.

According to Decision No. 1460/QD-TCMT, the VN-WQI index is calculated via eight water quality parameters, including pH (Group I), DO, BOD_5, COD, $N-NH_4$, $P-PO_4$ (Group IV), temperature, and total coliforms (Group V). The formula for calculating the VN-WQI index in this study, characterized by organic and nutrient pollution, is expressed as follows (Eq.1):

$$\mathrm{VN\_WQI} = \frac{\mathrm{WQI_I}}{100} \times \left[\left(\frac{1}{k} \sum_{i=1}^{k} \mathrm{WQI_{IV}} \right)^2 \times \frac{1}{l} \sum_{i=1}^{l} \mathrm{WQI_V} \right]^{1/3} \quad \text{(Eq. 1)}$$

where WQI_I represents the WQI value for pH; WQI_{IV} represents the WQI value for organic and nutrient parameters (including DO, BOD_5, COD, $N-NH_4$, and $P-PO_4$); and WQI_V represents the WQI value for microbiological parameters, including coliform.

A comprehensive description and step-by-step procedure for calculating the WQI are documented in Decision No. 1460/QD-TCMT.

The final VN-WQI values are rounded to the nearest integer on a scale from 0 to 100 and are classified into six distinct categories, each representing a different level of water quality. The water quality is considered excellent, with a WQI between 91 and 100, indicating optimal quality. Values between 76 and 90 fall into the good category, whereas a WQI of 51 to 75 indicates moderate quality. Water quality is categorized as poor when the WQI ranges from 26 to 50 and very poor when it falls between 10 and 25. WQI values below 10 indicate extreme pollution. The classification of water quality levels on the basis of VN_WQI values and their suitability for intended use is presented in Table 1.

Table 1 WQI water quality index rating scale

WQI ranges	Water quality	RGB color code	Suitable usage
91 – 100	Excellent	51;51;255	Suitable for domestic water supply
76 – 90	Good	0;228;0	Suitable for domestic water supply with appropriate treatment
51 – 75	Moderate	255;255;0	Suitable for irrigation and other similar purposes
26 – 50	Poor	255;126;0	Suitable for waterway transport and other similar uses
10 – 25	Very poor	255;0;0	Heavily polluted, requires treatment for future use
< 10	Extreme pollution	126;0;35	Toxic water, requires remediation and treatment

3) Description of ML algorithms

3.1) Radial basic function neuron network

The radial basic function (RBF) was introduced by Lowe and Broomhead [24] and is a type of feedforward neural network. The architecture of an RBF neural network consists of three layers: input, hidden, and output. The input layer receives the raw data (feature vectors) and passes it to the hidden layer. Each neuron in the hidden layer computes the resemblance between the input data and the prototype stored in that neuron. It uses a Gaussian activation function defined by the following formula:

$$\phi(\|x - c_j\|) = e^{-\left(\frac{\|x-c_j\|^2}{2\sigma_j^2}\right)} \quad \text{(Eq. 2)}$$

where *x* is the input vector, c_j represents the center of the Gaussian function, and σ_j specifies the width of the Gaussian function of the j^{th} neuron.

The output is determined via the weighted average approach, expressed via the following formula:

$$y_i = \sum_{j=1}^{n} W_{ij}\, \phi_j(x) \quad \text{(Eq. 3)}$$

where W_{ij} is the i^{th} weight between the hidden and output layers and *n* is the number of neurons in the hidden layer.

The classical RBF process primarily depends on three factors: the prototypes within each RBF neuron and how they are optimally chosen, the beta value, and the weights connecting the hidden layer and the output layer (which influence the final decision) [25].

3.2) Support vector regression

The SVR is a supervised learning algorithm for regression tasks derived from the support vector machine algorithm. Introduced by Vapnik et al. [26], SVR is grounded in statistical learning theory and has demonstrated marked effectiveness [27]. It is valuable for modeling data with nonlinear relationships and complex patterns, making it a robust choice for various predictive modeling challenges.

In the context of an SVR model, the connection between the target variable and the predictive variables (x) is captured through a regression function. According to Smola and Schölkopf [28], the regression function can be expressed as follows (Eq.4):

$$\hat{y} = f(x) = \omega * \varphi(x) + b \quad \text{(Eq. 4)}$$

where φ is a set of functions that replaces complex nonlinear relationships with simpler linear relationships and where ω and b denote the regression function weight and bias, respectively, which are determined by minimizing the difference between the predicted function $f(x)$ and the observed value (y).

Several hyperparameters of the SVR play a critical role in its predictive performance. The regularization parameter (C) regulates the trade-off between minimizing training errors and maintaining a large margin. The epsilon (ϵ) defines the tolerance margin around the predicted hyperplane where errors are not penalized. The kernel function transforms input data into a higher-dimensional space to handle nonlinear relationships. The RBF is the most common kernel function used to transform the input data into a higher-dimensional space [29].

3.3) Gradient boosting model

The gradient boosting model (GBM), originally introduced by Friedman [30], is an advanced ensemble-supervised algorithm. The GBM has become widely used in both regression and classification tasks because of its ability to handle nonlinear data and its high flexibility. The GBM is designed to construct robust predictive models by iteratively combining a series of weak learners, typically decision trees. As each new weak learner is added, it fits a model that enhances the accuracy of predicting the response variable. These additional learners are aligned with the negative gradient of the loss function, ensuring that they effectively reduce error across the entire ensemble. The model predicts values for the structure y = F(x), minimizing the mean squared error as described in Eq. 5 [17]:

$$\hat{y} = F(x) = \frac{1}{n}\sum_i (\hat{y}_i - y_i)^2 \quad \text{(Eq. 5)}$$

where y_i is the actual value; $\hat{y}_i$ is the predicted value; i represents the equities over some test data of size; and n represents the number of samples.

3.4) Extreme gradient boosting

Extreme gradient boosting (XGB), originated by Chen and Guestrin [31], represents an optimized version of gradient-boosted decision trees designed for fast and efficient execution. This algorithm excels at handling sparse data and demonstrates strong performance in both classification and regression tasks owing to its flexibility and versatility.

The XGB model uses an additive training approach, incrementally constructing trees in sequence to minimize the defined loss function. At each iteration t, the prediction is updated by adding the output of the newly trained tree ft (xi):

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad \text{(Eq. 6)}$$

The XGB framework optimizes an objective function Obj(θ), which balances the loss between the predicted and actual values, and a regularization term that penalizes model complexity:

$$Obj(\theta) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega(f_k) \quad \text{(Eq. 7)}$$

where $l\left(y_i, \hat{y}_i^{(t)}\right)$ is the loss function, typically the squared error for regression, and where $\Omega(D_t)$ is the regularization term that controls the complexity of the trees. The regularization term is formulated as follows:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{t} \omega_j^2 \quad \text{(Eq. 8)}$$

where T is the number of leaves in the tree, ω_j represents the weight of the j[th] leaf, and γ and λ are regularization parameters.

The theoretical foundation and detailed calculation procedure of the XGB model are comprehensively described by Chen and Guestrin [31].

According to Hussein et al. [18], the advantages of XGB's functionality can be summarized as follows: XGB is a gradient boosting algorithm that enhances predictive accuracy through iterative additive functions, where each decision tree in the sequence refines the model by correcting the errors of its predecessor. It uses decision trees as base learners trained to minimize a loss function. The model iteratively adds trees, each focusing on the residuals from previous trees to correct errors. Additionally, XGB incorporates regularization (L1 and L2) to prevent overfitting and improve generalizability. It utilizes ensemble learning by combining the predictions of all trees and parallel and distributed computing to handle large datasets efficiently. XGB also provides feature importance scores and supports hyperparameter tuning to optimize model performance.

In this study, four models (i.e., XGB, GBM, SVR, and RBF) were implemented in Python 3.10.9 via the Scikit-learn 1.2.1 and TensorFlow 2.12.0 frameworks and executed on a Windows 11 platform. During model training, a 10-fold cross-validation strategy was applied via GridSearchCV to select optimal hyperparameters and ensure robustness. The hyperparameters for each model, following the training process, are presented in Table 2.

Table 2 Hyperparameter tuning

Model	Hyperparameter	Range	Optimal value
RBF	Node of layer	[5, 7, 10, 12, 15]	10
	learning_rate	[0.001, 0.01, 0.03, 0.05, 0.1]	0.001
	batch_size	[1, 4, 8, 16, 32, 64]	4
	epochs	[50, 100, 200, 300, 500]	500
	Early stopping (patience)	[5, 10, 15, 20, 30]	15
	activation	['relu', 'tanh', 'sigmoid']	relu
	optimizer	['adam', 'sgd', 'rmsprop']	rmsprop
SVR	C	[1, 5, 10, 50]	10
	epsilon	[0.01, 0.03, 0.05, 0.07, 0.1]	0.05
	kernel	['linear', 'rbf', 'sigmoid']	rbf
	gamma	['scale', 'auto']	scale
GBM	learning_rate	[0.01, 0.03, 0.05, 0.1]	0.1
	max_depth	[3, 4, 5, 6, 7, 8]	3
	n_estimators	[100, 200, 300, 400, 500]	200
	subsample	[0.6, 0.8, 0.9, 1.0]	0.8
	max_features	['auto', 'sqrt', 'log2']	sqrt
XGB	learning_rate	[0.01, 0.03, 0.05, 0.1]	0.1
	max_depth	[1, 3, 5, 7, 9, 11]	3
	n_estimators	[100, 300, 500, 700, 900]	300
	reg_lambda	[0.01, 0.05, 0.1, 0.5]	0.1

4) Data collection and processing

The data used as input for the WQI prediction model in the lower Sai Gon River area include eight water quality parameters: DO, BOD_5, COD, N-NH_4, P-PO_4, pH, temperature (T), and total coliforms (coliform). Monthly data spanning five years, from 2015–2019, were collected from the Southern Regional Hydrometeorological Center at three water quality monitoring stations: SG1, SG2, and SG3. These stations were strategically selected to represent upstream, midstream, and downstream zones, ensuring the comprehensive capture of pollution dynamics along the river. The geographical locations of the selected sites are shown in Figure 2.

To further illustrate the distribution and variability of the water quality parameters across the monitoring stations, Figure 3 presents a series of boxplots for each parameter at SG1, SG2, and SG3.

These visualizations demonstrate that the parameters measured at the three stations varied significantly, especially for organic and microbiological pollution indicators. Apparent differences in median values for COD, BOD_5, N-NH_4, and coliform were observed between the stations, reflecting an increase in pollution levels from SG1 to SG3. The boxplots also highlight the presence of outliers, especially for COD, BOD_5, N–NH_4, coliform, and P–PO_4. In environmental monitoring, outliers may indicate discharges or seasonal variations in pollutant loads rather than errors, providing critical information for model training and decision-making. Therefore, these outliers are retained in the dataset to preserve representativeness in the study area.

Table 3 presents a statistical description of the water quality data in the study area. The statistical analysis reveals significant discrepancies between the maximum and minimum values of the parameters. For example, the maximum value of total coliform bacteria is 4,450 MPN per 100 mL, which is more than 20 times greater than the minimum value at 220 MPN per 100 mL, whereas the maximum and minimum values of COD are 134 mg L^{-1} and 5 mg L^{-1}, respectively. The variance among the parameters is also considerable, with the average value of total coliform bacteria at approximately 1,809 MPN per 100 mL and a temperature average of approximately 29.4°C, in contrast to the other parameters, which present relatively low average values, such as P-PO_4 (0.05 mg L^{-1}) and DO (3.72 mg L^{-1}). Furthermore, the standard deviation of total coliform bacteria is 854 MPN per 100 mL, indicating substantial data dispersion relative to the mean. In addition, the measurement units across the dataset are inconsistent: the temperature is recorded in °C, the total coliform concentration is in MPN per 100 mL, and the remaining parameters are in mg L^{-1}. These inconsistencies in both the magnitude of the values and the units of measurement highlight the need for standardization.

Standardizing the dataset is imperative prior to model training to ensure a consistent value range and unit system, as well as to improve convergence during training. On the basis of the characteristics of the collected dataset, this study employs the StandardScaler method to standardize the model's input data. This is a simple method that standardizes numerical features by scaling them to have a mean of 0 and a standard deviation of 1. The specific formula is expressed as follows:

$$x_i' = \frac{x_i - \bar{x}}{\sigma} \qquad \text{(Eq. 9)}$$

where x_i and x_i' are the original feature value and the normalized feature value, respectively, and where $\bar{x}$ and σ are the mean and standard deviation values of the feature in the dataset.

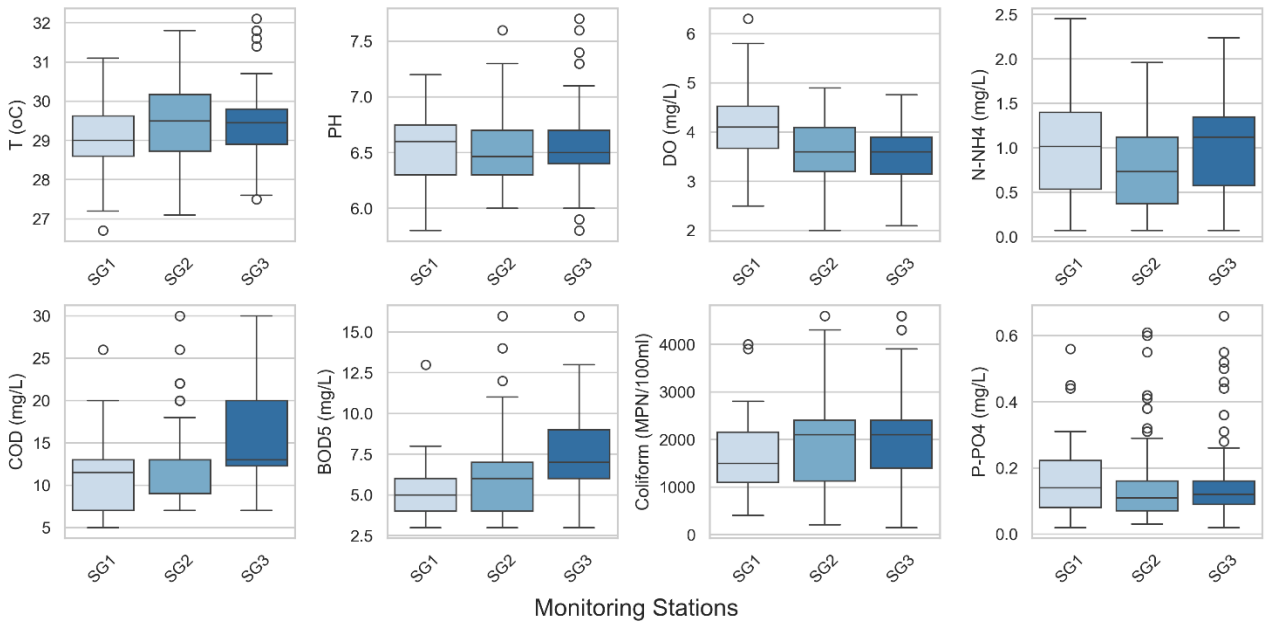


Figure 3 Boxplots of eight water quality parameters at three stations from 2015–2019.

Table 3 Statistical characteristics of the water quality parameters in the study area

Parameters	Unit	Maximum	Minimum	Mean	Standard deviation
T	^{o}C	32.60	27.75	29.40	0.92
pH	-	7.50	6.10	6.57	0.28
DO	mg L^{-1}	5.80	2.40	3.72	0.75
COD	mg L^{-1}	134.00	5.00	14.76	14.95
BOD_5	mg L^{-1}	69.00	3.00	7.18	7.69
$N–NH_4$	mg L^{-1}	10.64	0.07	1.03	1.23
$P–PO_4$	mg L^{-1}	0.19	0.01	0.05	0.04
Coliform	MPN per 100 mL	4,450.00	220.00	1,809.00	854.00

Subsequent to this preprocessing step, the dataset is divided into training and testing sets using an 80:20 split ratio, with a fixed random seed (random state=1) to ensure reproducibility. The data were split without shuffling to preserve the temporal sequence. Specifically, 80% (comprising 202 data points per parameter) are allocated for model training, whereas the remaining 20% (equivalent to 60 data points per parameter from the year 2019) are reserved for model evaluation. This allocation ratio has been widely adopted and has demonstrated substantial effectiveness in several studies employing artificial intelligence models for water quality prediction [11, 14, 16, 21, 32].

5) Model evaluation

Model performance is assessed through both graphical and statistical methods to evaluate the reliability of the predicted results relative to the observed data. In this study, the statistical evaluation metrics include the coefficient of determination (R^2), Nash–Sutcliffe efficiency (NSE) [33], mean absolute error (MAE), and root mean square error (RMSE). These indices collectively provide a comprehensive assessment of model accuracy and predictive reliability. Notably, the NSE measures how closely the plot of observed versus simulated data aligns with the 1:1 line, indicating the model's ability to replicate observed data accurately. The R^2, on the other hand, assesses the strength of the linear relationship between observed and simulated data, providing insight into the degree of association between the two datasets [34]. The RMSE and MAE can be used to determine confidence intervals in model predictions, with the potential to incorporate measurement uncertainty into the analysis [35-36]. The closer the values of R^2 and NSE are to 1, the higher the model performance. Conversely, the closer the values of the MAE and RMSE are to 0, the smaller the model's error is [34]. These statistics are determined via the following formulas:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i^{obs} - Y_i^{sim}\right| \quad \text{(Eq.10)}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{obs} - Y_i^{sim}\right)^2} \quad \text{(Eq.11)}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i-1}^{n}\left(Y_i^{obs} - Y_i^{sim}\right)^2}{\sum_{i-1}^{n}\left(Y_i^{obs} - \bar{Y}^{obs}\right)^2} \quad \text{(Eq.12)}$$

$$R^2 = \left[\frac{\sum_{i=1}^{n}\left(Y^{obs}{}_i - \bar{Y}^{obs}\right)\times\left(Y^{sim}{}_i - \bar{Y}^{sim}\right)}{\sqrt{\sum_{i=1}^{n}\left(Y^{obs}{}_i - \bar{Y}^{obs}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y^{sim}{}_i - \bar{Y}^{sim}\right)^2}}\right]^2 \quad \text{(Eq.13)}$$

where Y_i^{obs} represents the i^{th} observed value; Y_i^{sim} represents the i^{th} simulated value; $\bar{Y}^{sim}$ represents the average simulated value; and $\bar{Y}^{obs}$ represents the average observed value.

6) Scenario

The WQI is traditionally calculated through extensive, time-intensive computational methods and is often prone to occasional errors, particularly during subindex calculations. Additionally, traditional methods for calculating WQI require numerous physical and chemical parameters. To provide an efficient alternative for WQI calculation and prediction, Scenario S8 was developed, incorporating eight water quality parameters, including DO, BOD_5, COD, $N-NH_4$, $P-PO_4$, pH, T, and coliform.

Furthermore, the challenges associated with collecting and analyzing all these parameters, especially those that require precise laboratory conditions, significant time, and high costs, should be recognized. This study investigated the feasibility of reducing the number of input variables while maintaining predictive performance and regulatory compliance. Specifically, a set of reduced-input scenarios (S7-COD, S7-BOD, S7-NH_4, and S7-PO_4) was developed, with each scenario excluding one parameter (COD, BOD_5, $N-NH_4$, or $P-PO_4$, respectively) from the prediction model. These reduced scenarios aim to minimize information loss while still complying with Decision No. 1460/QĐ-TCMT, which requires at least three designated parameter groups (I, IV, and V) for WQI computation, with group IV being compulsory. Furthermore, at least three parameters from group IV must be used in the calculation.

This approach reflects real-world constraints in water quality monitoring and shares a similar purpose of parameter exclusion with previous studies by Hameed et al. [11], Khoi et al. [20], and Nguyen et al. [21]. However, the specific methods for selecting which

parameters to exclude in our study differ, as they are based on practical factors such as laboratory costs, analysis time requirements, and regulatory compliance rather than correlations among parameters. Table 4 presents the specific input parameters used for the scenarios analyzed.

Results

1) Performance evaluation of the ML algorithms

This study evaluated the performance of four machine learning (ML) algorithms, including RBF, SVR, GBM, and XGB, in simulating the WQI using eight input parameters: pH, T, DO, COD, BOD_5, $P\text{-}PO_4$, $N\text{-}NH_4$, and coliform (Scenario S8). To ensure consistent evaluation, a quantitative comparison was conducted Table 5 presents performance metrics, including the MAE, RMSE, NSE, and R^2, which are based on both training and testing results for all the models.

During the training phase, all four models demonstrated excellent simulation performance, characterized by low error rates and high correlation coefficients. In particular, the values of the MAE and RMSE vary from approximately 0.199 to 1.355 and 0.259 to 2.197, respectively. The values of NSE and R^2 exceed 0.9. Additionally, Figure 4 these findings further confirm this excellent performance. The scatter points in the Taylor diagram approximate the reference point very well, indicating similar variations, few errors, and high correlations compared with the observations. In comparison, the GBM model exhibits the best performance, followed by XGB, SVR, and RBF, in that order.

In the testing phase, the RMSE values for XGB, SVR, RBF, and GBM were 1.630, 3.303, 3.926, and 2.656, respectively. The corresponding NSE values were 0.953, 0.807, 0.728, and 0.875, respectively. Among the models, the XGB model showed the best alignment with the actual measured WQI, achieving the lowest error rates (RMSE = 1.630, MAE = 0.782) and the highest correlation (R^2 = 0.960, NSE = 0.953). Its position is closest to the reference point on the Taylor diagram (representing actual measurements) (Figure 4b), underscoring its superior performance. The GBM ranks second, followed by SVR, whereas the RBF model has the lowest simulation performance, as evidenced by its position being farthest from the reference point on the Taylor diagram.

To further evaluate robustness, a stratified analysis by stations and years was also conducted to interpret model performance across spatial and temporal dimensions. As shown in Figure 5, XGB and the GBM consistently outperformed the other models at all three stations, achieving high R^2 values (generally >0.95) and low RMSEs across most years. At SG1, XGB and the GBM exhibited stable performance, whereas the RBF exhibited large fluctuations, particularly in 2016 and 2019. SG2 demonstrated the most consistent results, with XGB and GBM maintaining near-perfect R^2 values and minimal RMSEs throughout the five years. In contrast, SG3 experienced a slight decline in prediction accuracy in 2019 across all the models due to increased data variability and the presence of outliers at SG3. However, XGB maintained higher accuracy than GBM did, with better R^2 values and lower RMSEs. It can be concluded that XGB was the most robust model despite spatial and temporal variability. These results highlight its robustness in modeling complex and variable real-world water quality conditions.

Owing to its superior performance, the XGB model is selected to predict the WQI under various scenarios that are designed to minimize water quality parameters when calculating the WQI within the study area.

Table 4 Scenarios of WQI prediction

No.	Scenario	Input parameters	Excluded parameter
1	S8	pH, T, DO, COD, BOD_5, $P\text{-}PO_4$, $N\text{-}NH_4$, and Coliform	None
2	S7-BOD	pH, T, DO, COD, $P\text{-}PO_4$, $N\text{-}NH_4$, and Coliform	BOD_5
3	S7-COD	pH, T, DO, BOD_5, $P\text{-}PO_4$, $N\text{-}NH_4$, and Coliform	COD
4	S7-PO4	pH, T, DO, COD, BOD_5, $N\text{-}NH_4$, and Coliform	$P\text{-}PO_4$
5	S7-NH4	pH, T, DO, COD, BOD_5, $P\text{-}PO_4$, and Coliform	$N\text{-}NH_4$

Table 5 Efficiency statistics of the ML models in scenario S8

Model	Phases	MAE	RMSE	NSE	R^2
RBF	Training	1.355	2.197	0.917	0.932
	Testing	2.688	3.926	0.728	0.767
SVR	Training	0.435	0.573	0.994	0.995
	Testing	2.030	3.303	0.807	0.830
GBM	Training	0.199	0.259	0.999	0.999
	Testing	1.432	2.656	0.875	0.898
XGB	Training	0.334	0.451	0.997	0.997
	Testing	0.782	1.630	0.953	0.960

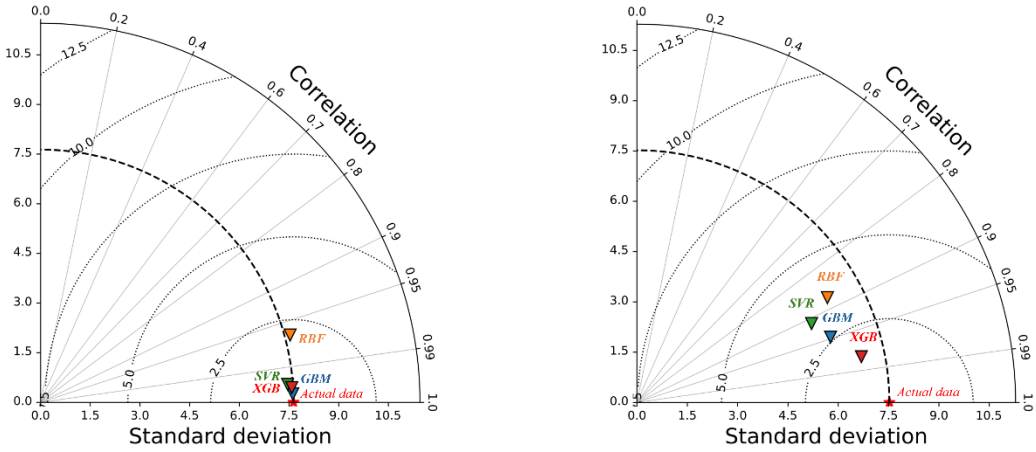


Figure 4 Taylor diagrams comparing simulation results between the ML models in the (a) training and (b) testing phases.

Figure 5 Comparison of model performance by station and year.

2) Performance evaluation of scenarios

Table 6 presents the performance of the XGB model in predicting the WQI of the Sai Gon River under five scenarios designed to test input exclusion, as evaluated by statistical performance measures, including the MAE, RMSE, NSE, and R^2.

The results indicate that in the training phase, Scenarios S8, S7-BOD, and S7-COD yield highly accurate simulations characterized by low errors and high correlations. Specifically, the NSE and R^2 indices exceed 0.990, whereas the MAE and RMSE values remain below 0.5. Conversely, Scenarios S7-PO4 and S7-NH_4 achieve high correlations with NSE and R^2 values above 0.880, although their relative errors are greater than those of the other scenarios, with MAE and RMSE values ranging from 1.803 to 2.644.

In the testing phase, Scenario S8, which integrates all the parameters, achieves the highest predictive accuracy, as evidenced by the lowest MAE (0.782) and RMSE (1.630) values. The strong alignment between the simulated and observed WQI values, demonstrated by high NSE (0.953) and R^2 (0.960) values, underscores the effectiveness of including all eight parameters for reliable prediction. Scenarios S7-COD and S7-BOD also perform satisfactorily, with moderate error levels (MAEs of 1.075 and 1.052 and RMSEs of 1.890 and 2.003, respectively) and high correlations with observed values (NSE and R^2 values above 0.92). Scenarios S7-COD and S7-BOD also perform satisfactorily, with moderate error levels (MAEs of 1.075 and 1.052 and RMSEs of 1.890 and 2.003, respectively) and high correlations with observed values (NSE and R^2 values above 0.92). In comparison, Scenario S7-PO_4, which excludes the P-PO_4 parameter, results in increased error (MAE of 2.676 and RMSE of 3.138) and a slightly reduced correlation (NSE of 0.826 and R^2 of 0.850), indicating a decline in predictive accuracy. Moreover, Scenario S7-NH_4, which omits N-NH_4, produces the highest errors (MAE of 3.718 and RMSE of 4.661) and the weakest correlation (NSE of 0.616 and R^2 of 0.635), reflecting a substantial decrease in predictive performance.

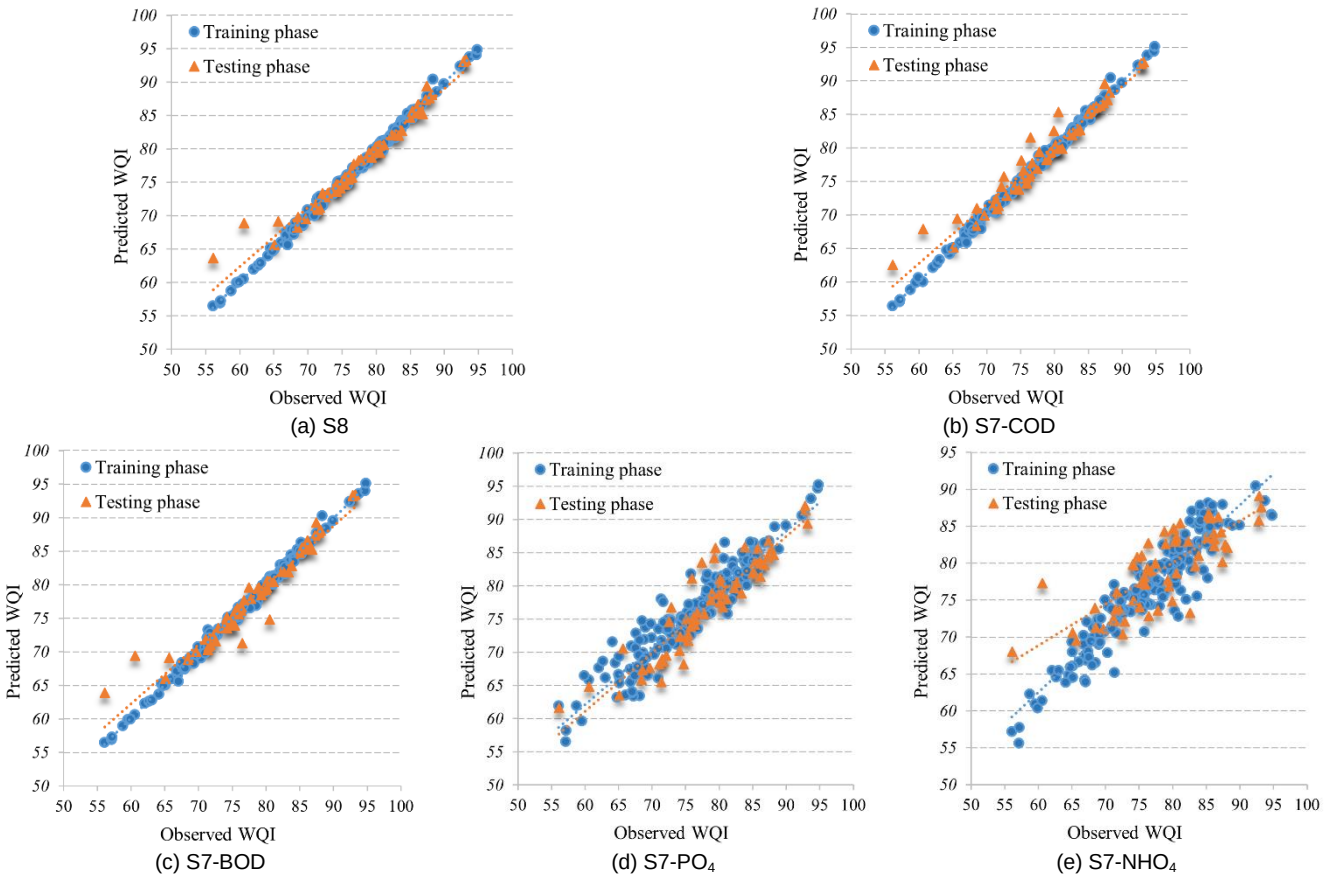
Figure 6 further depicts scatter plots between the observed and simulated WQI values during the training and testing phases across the five scenarios. These plots illustrate the alignment between the observed and simulated WQI values generated by the XGB model. The observed and simulated WQI values during the training phase (blue circles) clearly exhibit less dispersion than those in the testing phase (orange triangles). The correlation between the observed and simulated WQI values is notably lower in scenarios S7-NH_4 and S7-PO_4, whereas scenarios S8, S7-BOD, and S7-COD demonstrate greater correlations.

To compare the simulation performance of the XGB model across five scenarios, this study presents the outcomes for each scenario during the training period (2015–2018) and the testing period (2019) via Taylor diagrams (Figure 7).

An examination of the scatter positions on the diagrams indicates that the scenario incorporating all eight parameters (Scenario S8) achieves the highest correlation with the observed data, along with the lowest error and similar variations to the observations, thus demonstrating the most accurate simulation performance. Scenarios that exclude individual parameters, specifically COD and BOD_5 (scenarios S7-COD and S7-BOD), also perform robustly, maintaining a high correlation with the observed WQI values. Moreover, the scenario excluding P-PO_4 (scenarios S7-PO_4) results in low errors but high biases and reduced correlations. In particular, Scenario S7-NH_4 yields the least effective WQI simulation among all five scenarios across both the training and testing phases, as evidenced by its position being farthest from the reference points.

Table 6 Efficiency statistics of the XGB model under the 5 scenarios of input variable combinations

Phases	Scenarios	S8	S7-BOD	S7-COD	S7-PO_4	S7-NH_4
Training	MAE	0.334	0.372	0.375	1.803	2.030
	RMSE	0.451	0.507	0.494	2.411	2.644
	NSE	0.997	0.996	0.996	0.900	0.880
	R^2	0.997	0.996	0.996	0.901	0.881
Testing	MAE	0.782	1.052	1.075	2.676	3.718
	RMSE	1.630	2.003	1.890	3.138	4.661
	NSE	0.953	0.929	0.937	0.826	0.616
	R^2	0.960	0.932	0.948	0.850	0.635



(a) S8 (b) S7-COD

(c) S7-BOD (d) S7-PO_4 (e) S7-NHO_4

Figure 6 Plots of the predicted and actual WQI values across the five scenarios.

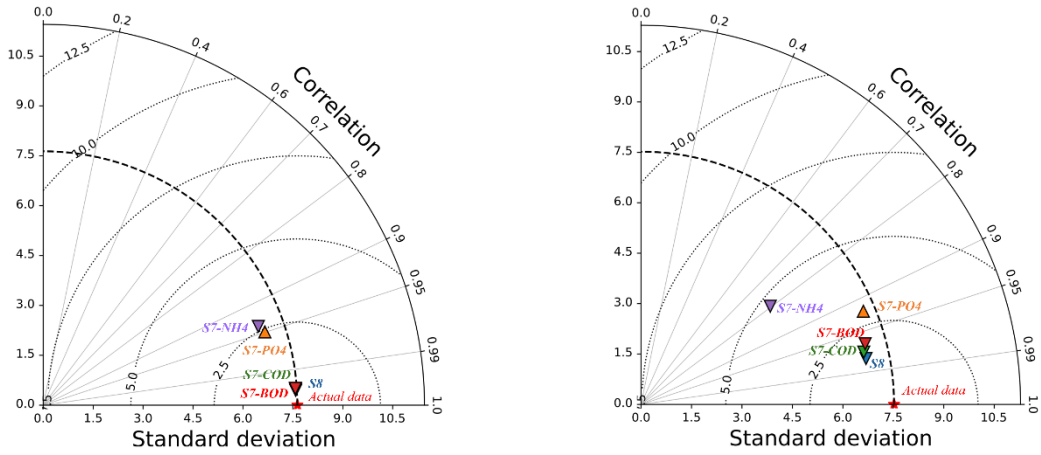


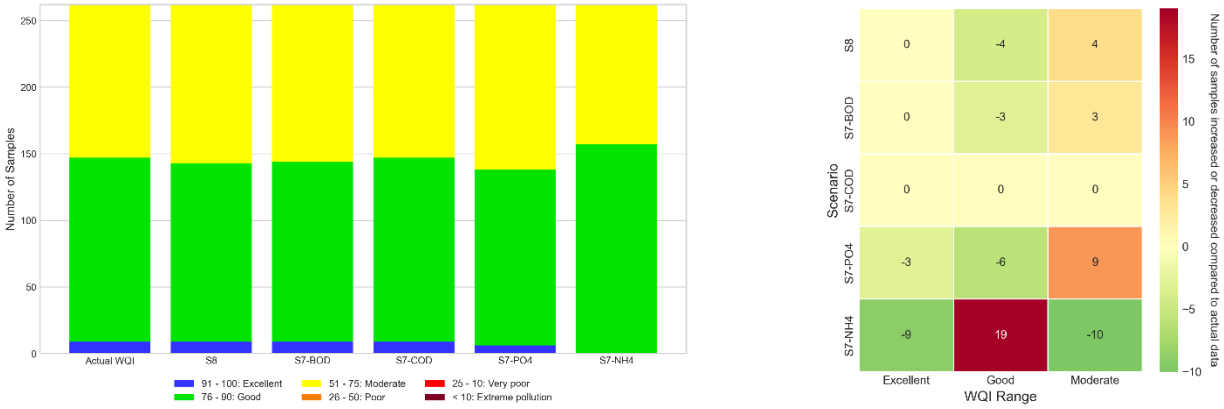
Figure 7 Taylor diagrams comparing simulation results between scenarios in the (a) training and (b) testing phases.

3) Comparison with actual WQI ranges

This study further analyzed sample distributions across six water quality classification thresholds, as specified in Decision No. 1460/QD-TCMT, dated November 12, 2019. This analysis compares the observed WQI values with the simulated WQI values across various scenarios via the XGB model to assess whether the model predictions significantly diverge from the actual measurements. The statistical findings are presented in Figure 8. Figure 8a compares the WQI ranges across different scenarios, with each range color-coded according to the guidelines outlined in Decision No. 1460/QD-TCMT. Figure 8b illustrates the changes in the number of samples in each WQI classification, indicating whether the number increased (positive value) or decreased (negative value) compared with the actual data, as displayed in the heatmap.

A statistical analysis of WQI values derived from observed data between 2015 and 2019 reveals the following distribution: 3.4% (9 samples) fall within the "Excellent" category (WQI range: 91–100), 52.7% (138 samples) are classified as "Good" (WQI range: 76–90), and 43.9% (115 samples) are classified as "Moderate" (WQI range: 51–75), with no samples below a WQI of 50. These findings suggest that the surface water quality of the Sai Gon River primarily aligns with the "good" and "moderate" classification levels.

Figure 8 also indicates that the distribution of samples across the simulated water quality ranges in scenario S7-COD closely aligns with that of the observed WQI values. This alignment indicates that the COD parameter could be considered nonessential for WQI prediction within the XGB model in this study area. In Scenario S8, which incorporates all eight input parameters, although this scenario has the highest correlation among the five scenarios, there are differences in the distributions of samples across the "good" and "moderate" water quality levels in the simulated results. Specifically, four samples initially classified as "good" shifted to the "moderate" category, with this discrepancy occurring during the testing phase in 2019. Similar differences arise in Scenario S7-BOD, where the BOD_5 parameter is excluded: three samples initially classified as "good" shift to "moderate" in the simulated results. The simulated sample distribution in this scenario aligns with the observed data for the remaining water quality levels.



(a) Comparison of WQI classification across scenarios (b) Change in sample number relative to actual data

Figure 8 Characteristics of observed and simulated WQI values across different scenarios.

For the scenario in which the P-PO_4 parameter is excluded, the model's simulated results diverge significantly from the observed WQI values. Specifically, the number of samples in the "Excellent" and "Good" quality categories decreases to 3 and 6 samples, respectively, whereas the number of "Moderate" quality samples increases by 9. Excluding the N-NH_4 parameter results in the most significant discrepancy between the simulated and observed WQI values: no samples reach the "Excellent" category, the "Good" category has 19 additional samples compared with the observed data, and the "Moderate" category has ten fewer samples. These findings suggest that excluding P-PO_4 and N-NH_4 is inappropriate for WQI prediction, as it significantly affects water quality assessment outcomes in this study area.

Discussion

Research findings demonstrate that boosting algorithms, particularly XGB, perform effectively in predicting the WQI, which is consistent with previous studies [18-20], which also identify XGB as more effective than other ML models.

However, in other cases, Raheja et al. [17] reported that the DNN model outperforms both XGB and the GBM. Moreover, Nguyen et al. [21] reported that the GBM is more effective than XGB and deep learning algorithms. These results suggest that different models excel with particular datasets or areas, as their effectiveness depends on factors such as data distribution, regional traits, and how well the model adapts to these variations. For example, in southern Vietnam, including the Sai Gon River (in this study) and the La Buong River [20], XGB has been identified as the most effective predictive model. Conversely, in northern Vietnam, such as the Red River Basin [21] and the An Kim Hai system [22], the GBM and RF have demonstrated superior predictive capabilities. These findings highlight the necessity of model selection for the specific local context and data characteristics, as no model is universally optimal.

The study also considers the effects of reducing the input water quality parameters when calculating the WQI. The results reveal that the parameters P-PO_4 and N-NH_4 substantially impact WQI outcomes. Conversely, omitting the parameters BOD_5 or COD did not significantly affect the WQI results of the XGB model. Notably, when the COD parameter is excluded, the water quality classification based on the WQI is closely aligned with the measured data. This finding is consistent with those reported by Hameed et al. [11], which also indicates that excluding the BOD_5 parameter had minimal impact on WQI predictions. Furthermore, Kamyab-Talesh et al. [14] identified nitrate and phosphate as the most critical factors influencing the WQI. On the other hand, Mohd Zebaral Hoque et al. [16], Khoi et al. [20], and Lap et al. [22] demonstrated that omitting these parameters still resulted in accurate WQI predictions. The influence of input water quality parameters on WQI calculations varies depending on the model and the study's specific context. As such, it is necessary to quantify each parameter's contribution to WQI prediction through approaches such as sensitivity analysis in future research. This would allow for the exclusion of less influential parameters while maintaining model accuracy.

In addition, despite the limitations of sample size, several factors support their suitability for ML modeling and contribute to the reliability of the results. First, the three monitoring stations represented hydrological zones, upstream, midstream, and downstream, allowing a comprehensive understanding of the pollution dynamics along the river. Second, the monthly data spans multiple dry and rainy seasons across five years, providing sufficient temporal variability for modeling water quality trends. Additionally, 10-fold cross-validation was applied to enhance model generalizability and reduce potential bias due to the limited dataset size. The XGB model's performance indicated reliability despite the modest dataset size, further supporting its effectiveness in handling limited data, as demonstrated in previous studies [18–20].

Similar studies in the literature have effectively utilized ML models with datasets of comparable or even smaller scales, achieving high predictive accuracy. Table 7 summarizes the selected studies for comparison. These findings suggest that, when carefully designed, ML models can deliver reliable results even with limited datasets, particularly when high-quality input features and appropriate validation strategies are used. This is meaningful in water quality management in developing countries, where water quality monitoring faces significant challenges due to the lack of monitoring sensors and high costs, resulting in sparse and infrequent datasets. Nevertheless, increasing the spatial coverage and extending the time series would undoubtedly further improve the generalizability and robustness of the models. Future studies should explore the integration of additional stations and longer data records to enhance predictive capabilities and support broader application, as well as integrate site-specific contextual information at monitoring stations, such as hydraulic conditions or point-source discharge locations, to enhance interpretability.

Conclusions

An informed decision on water quality management requires tailored tools for accurately predicting the water quality index. State-of-the-art machine learning algorithms have offered various efficient toolkits for assessing water quality. In an attempt to simulate the WQI in the Sai Gon River from 2015–2019, the present study applied four ML algorithms (i.e., XGB, GBM, RBF, and SVR) to monthly data of eight water quality parameters, including DO, BOD_5, COD, N-NH_4, P-PO_4, pH, temperature, and total coliforms. Moreover, five scenarios with varying input

parameters were designed to identify the best-performing model and further explore the possibility of minimizing the number of input parameters for WQI calculation in this study area.

On the basis of the same scenario with all eight input parameters (i.e., Scenario S8), this study first identified the optimal model by comparing the simulation performance of the four models in predicting the WQI during the training and testing phases. The optimal model was subsequently adopted to simulate the remaining scenarios to evaluate the possibility of reducing the number of input parameters for WQI calculation. The results demonstrate that the XGB model effectively simulates the WQI for this study, achieving the lowest error rates and the highest correlation, followed by GBM, SVR, and RBF. Moreover, this study also indicates that N-NH_4 and P-PO_4 are essential factors in WQI prediction in the study area, whereas omitting BOD_5 results in only minor decreases in model performance. Therefore, COD or BOD5 may be excluded when the WQI is calculated via the XGB model in cases where these data are not available or where it is desirable to reduce the cost and time of analyzing these parameters in the laboratory.

The results of this study provide solid evidence supporting the application of ML algorithms in calculating WQI. This study demonstrates that ML models can reduce the number of parameters needed for WQI estimation while ensuring accuracy, providing a practical solution for water quality monitoring in developing countries such as Vietnam. By adopting these techniques, developing nations can enhance their assessment and management of surface water resources more effectively and sustainably.

Table 7 Comparison of this research with previous studies using small datasets

Location	Time (monthly)	No. of stations/samples	Best model	R^2	Ref.
Saigon River, Vietnam	2015–2019	03/262	XGB	0.96	This study
Lam Tsuen River, Hong Kong	1998–2017	01/240	ETR	0.98	[12]
			SVR	0.96	
Northern Iran	2012–2018	02/144	BA-RF	0.94	[13]
Sefidrud Basin, Iran	12/2007–11/2008	05 stations	SVR	0.87	[14]
Naama, Algeria	-	166 samples	XGB (10)	0.96	[18]
			SVR (4)	0.96	
			KNN (5)	0.94	
El Moghra, Egypt	11/2018–12/2019	46/46	XGB	0.872	[19]
La Buong River, Vietnam	2010–2018	02/220	XGB	0.989	[20]

Acknowledgements

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number C2023-18-25.

We would like to express our deepest gratitude to Assoc. Prof. Dao Nguyen Khoi for his pivotal contributions to this research and his unwavering support throughout our academic journey.

References

[1] Horton, R.K. An index number system for rating water quality. Journal of the Water Pollution Control Federation, 1965, 37(3), 300–306.

[2] Chidiac, S., El Najjar, P., Ouaini, N., El Rayess, Y., El Azzi, D. A comprehensive review of water quality indices (WQIs): History, models, attempts and perspectives. Reviews in Environmental Science and Bio/Technology, 2023, *22*(2), 349–395.

[3] Brown, R.M., McClelland, N.I., Deininger, R.A., Tozer, R. G. A water quality index-do we dare. Water and Sewage Works, 1970, 117(10), 339–343.

[4] Lumb, A., Sharma, T.C., Bibeault, J.F. A review of genesis and evolution of WQI and some future directions. Water Quality, Exposure and Health, 2011, 3, 11–24.

[5] Shah, K.A., Joshi, G.S. Evaluation of water quality index for River Sabarmati, Gujarat, India. Applied Water Science, 2017, 7, 1349–1358.

[6] Cash, K., Wright, R. Canadian water quality guidelines for the protection of aquatic life. CCME: Ottawa, ON, Canada, 2001.

[7] Bharti, N., Katyal, D. Water quality indices used for surface water vulnerability assessment. International Journal of Environmental Sciences, 2011, 2(1), 154–173.

[8] Shuhaimi-Othman, M., Lim, E.C., Mushrifah, I. Water quality changes in Chini lake, Pahang, west Malaysia. Environmental Monitoring and Assessment, 2007, 131, 279–292.

[9] MONRE. Decision No. 1460/QD-TCMT on the promulgation of technical guidelines for calculation and disclosure of Vietnam Water Quality Index (VN-WQI). Vietnam Environmental Administration – Ministry of Natural Resources and Environment (VEA–MONRE): Hanoi, Vietnam, 2019.

[10] Wu, Y., Liu, S. Modeling of land use and reservoir effects on nonpoint source pollution in a highly

agricultural basin. Journal of Environmental Monitoring, 2012, 14(9), 2350–2361.

[11] Hameed, M., Sharqi, S.S., Yaseen, Z. M., Afan, H.A., Hussain, A., Elshafie, A. Application of artificial intelligence (AI) techniques in water quality index prediction: A case study in tropical region, Malaysia. Neural Computing and Applications, 2017, *28*, 893–905.

[12] Asadollah, S.B.H.S., Sharafati, A., Motta, D., Yaseen, Z.M. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. Journal of Environmental Chemical Engineering, 2021, 9(1), 104599.

[13] Bui, D.T., Khosravi, K., Tiefenbacher, J., Nguyen, H., Kazakis, N. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Science of the Total Environment, 2020, 721, 137612.

[14] Kamyab-Talesh, F., Mousavi, S.F., Khaledian, M., Yousefi-Falakdehi, O., Norouzi-Masir, M. Prediction of water quality index by support vector machine: A case study in the Sefidrud Basin, Northern Iran. Water Resources, 2019, 46, 112–116.

[15] Othman, F., Alaaeldin, M.E., Seyam, M., Ahmed, A.N., Teo, F.Y., Ming Fai, C., ... & El-Shafie, A. Efficient river water quality index prediction considering minimal number of inputs variables. Engineering Applications of Computational Fluid Mechanics, 2020, 14(1), 751–763.

[16] Mohd Zebaral Hoque, J., Ab. Aziz, N.A., Alelyani, S., Mohana, M., Hosain, M. Improving water quality index prediction using regression learning models. International Journal of Environmental Research and Public Health, 2022, 19(20), 13702.

[17] Raheja, H., Goel, A., Pal, M. Prediction of groundwater quality indices using machine learning algorithms. Water Practice and Technology, 2022, 17(1), 336–351.

[18] Hussein, E.E., Derdour, A., Zerouali, B., Almaliki, A., Wong, Y.J., Ballesta-de los Santos, M., ... Elbeltagi, A. Groundwater quality assessment and irrigation water quality index prediction using machine learning algorithms. Water, 2024, 16(2), 264.

[19] Kamel Elshaarawy, M., Eltarabily, M.G. Machine learning models for predicting water quality index: Optimization and performance analysis for El Moghra, Egypt. Water Supply, 2024, 24(9), 3269–3294.

[20] Khoi, D.N., Quan, N.T., Linh, D.Q., Nhi, P.T.T., Thuy, N.T.D. Using machine learning models for predicting the water quality index in the La Buong River, Vietnam. Water, 2022, 14(10), 1552.

[21] Nguyen, D.P., Ha, H.D., Trinh, N.T., Nguyen, M. T. Application of artificial intelligence for forecasting surface quality index of irrigation systems in the Red River Delta, Vietnam. Environmental Systems Research, 2023, 12(1), 24.

[22] Lap, B.Q., Hong, P.T.T., Du Nguyen, H., Quang, L.X., Hang, P.T., Phi, N.Q., … Hang, B.T.T. Predicting WQI by feature selection and machine learning: A case study of An Kim Hai irrigation system. Ecological Informatics, 2023, 74, 101991.

[23] Hang, H.T.M., Hung, N.T., Van Dung, N. The management of pollution sources in Dong Nai river system basin. Science and Technology Development Journal, 2006, 9 (SI: TN&MT), 5–17.

[24] Lowe, D., Broomhead, D. Multivariable functional interpolation and adaptive networks. Complex Systems, 1988, 2(3), 321–355.

[25] Aljarah, I., Faris, H., Mirjalili, S., Al-Madi, N. Training radial basis function networks using biogeography-based optimizer. Neural Computing and Applications, 2018, 29, 529–553.

[26] Vapnik, V., Golowich, S., Smola, A. Support vector method for function approximation, regression estimation, and signal processing. In: Mozer M.C., Jordan M.I., and Petsche T. (Eds.) Advances in Neural Information Processing Systems 9, MA, MIT Press, Cambridge, 1997, 281–287.

[27] Vapnik, V. Three remarks on the support vector method of function estimation. In: Sch¨olkopf B., Burges C.J.C., and Smola A.J. (Eds.), Advances in Kernel Methods—Support Vector Learning, MIT Press, Cambridge, MA, 1999, 25–42.

[28] Smola, A.J., Schölkopf, B. A tutorial on support vector regression. Statistics and Computing, 2004, 14, 199–222.

[29] Li, J., Abdulmohsin, H.A., Hasan, S.S., Kaiming, L., Al-Khateeb, B., Ghareb, M.I., Mohammed, M.N. Hybrid soft computing approach for determining water quality indicator: Euphrates River. Neural Computing and Applications, 2019, 31, 827–837.

[30] Friedman, J.H. Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 2001, 29(5), 1189-1232.

[31] Chen, T., Guestrin, C. Xgboost. A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2016, 785–794.

[32] Bedi, S., Samal, A., Ray, C., Snow, D. Comparative evaluation of machine learning models for groundwater quality assessment. Environmental Monitoring and Assessment, 2020, 192, 1–23.

[33] Nash, J.E., Sutcliffe, J.V. River flow forecasting through conceptual models Part I—A discussion of principles. Journal of hydrology, 1970, 10(3), 282–290.

[34] Moriasi, D.N., Gitau, M.W., Pai, N., Daggupati, P. Hydrologic and water quality models: Performance

measures and evaluation criteria. Transactions of the ASABE, 2015, 58(6), 1763–1785.

[35] Harmel, D.R., Smith, P.K. Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modeling. Journal of Hydrology, 2007, 337(3), 326–336.

[36] Harmel, R.D., Smith, P.K., Migliaccio, K.W. Modifying goodness-of-fit indicators to incorporate both measurement and model uncertainty in model calibration and validation. Transactions of the ASABE, 2010, 53(1), 55-63.