Environmental Systems Research

RESEARCH

Open Access

# Application of artificial intelligence for forecasting surface quality index of irrigation systems in the Red River Delta, Vietnam

Duc Phong Nguyen[1*], Hai Duong Ha[1], Ngoc Thang Trinh[1] and Minh Tu Nguyen[1]

**Abstract**

Water sources for irrigation systems in the Red River Delta are crucial to the socioeconomic growth of the region's communities. Human activities (discharge) have polluted the water source in recent years, and the water source from upstream is limited. Currently, the surface water quality index (WQI), which is calculated from numerous surface water quality parameters (physical, chemical, microbiological, heavy metals, etc.) is frequently used to evaluate the surface water quality of irrigation systems. However, the calculation of the WQI from water quality monitoring parameters remains constrained due to the need for a large number of monitoring parameters and the relative complexity of the calculation. To better serve the assessment of surface water quality in the study area, it is crucial and essential to conduct research to identify an efficient and accurate method of calculating the WQI. This study used machine learning and deep learning algorithms to calculate the WQI with minimal input data (water quality parameters) to reduce the cost of monitoring surface water quality. The study used the Bayes method (BMA) to select important parameters ($BOD_5$, $NH_4^+$, $PO_4^{3-}$, turbidity, TSS, coliform, and DO). The results indicate that the machine learning model is more effective than the deep learning model, with the gradient boosting model having the most accurate prediction results because it has the highest coefficient of determination $R^2$ (0.96). This is a solid scientific basis and an important result for the application of machine learning and deep learning algorithms to calculate WQI for the research area. The study also demonstrated the potential of artificial intelligence algorithms to improve water quality forecasting compared to traditional methods with minimal cost and time.

**Keywords** Machine learning model, Deep learning model, Surface water quality, Red River Delta, Irrigation system

## Introduction

The Red River Delta is the downstream area of the Red River and Thai Binh Rivers in northern Vietnam. The Red River Delta consists of 10 provinces, including 2 cities directly under the central government and 9 provinces with 16 cities under each province. This is the region with the highest population density in Vietnam (1450 people/$km^2$, population is 21,848,913 people).

The area around the Red River Delta is split into three subregions. There are 14 irrigation systems that are different from the areas upstream and in the middle of the Red River Delta. Although the level of water shortage is not severe, tides and saltwater intrusion are factors. There are 2 irrigation systems in the upstream area and 5 irrigation systems in the center of the Red River Delta that are greatly affected by the decline in water sources and are also the areas most affected by socioeconomic development activities, and water pollution is increasing daily. The research results show that the systems in the central delta are more polluted than the upstream and downstream systems. Therefore, the scope of the study was determined to be the irrigation systems representing

*Correspondence:
Duc Phong Nguyen
phongndtv@gmail.com
[1] Institute for Water and Environment (Vietnam Academy for Water Resources), Hanoi, Vietnam

the central part of the Red River Delta, including the Bac Duong and Bac Hung Hai irrigation systems, because these are two typical and serious pollution systems for the study area due to the influence of human activities (discharge) and the impacts of upstream flows (water resources are increasingly limited). Moreover, these are also systems with sufficient data (for a long enough time) to calculate, evaluate, and forecast trends in surface water quality. The geographical location of the study area is shown in Fig. 1.

Water quality in the irrigation systems in the study area is monitored with a frequency of 2–6 times a year, arranged at the time of water supply for the spring crop (from February to April) and the time of irrigation water supply during the crop (from July to September). Therefore, the level of pollution increase, pollution indicators, and causes of pollution were assessed as a basis for proposing solutions to reduce pollution and minimize the harmful effects of water pollution on agricultural production and aquaculture (Chinh 2019).

The general assessment of water pollution in a number of irrigation works that are being watched shows that both the scope and extent of water pollution have grown. Common pollution parameters are DO, $BOD_5$, COD, $NH_4^+$, $NO_2^-$, and coliform. Most of the monitoring points do not meet the standard of water supply for daily life (according to QCVN 08-2015), and approximately 30–50% of monitoring points do not meet irrigation water standards. Companies exploiting irrigation works have to spend a lot of money picking up trash to clear the flow. The water pollution situation in some typical irrigation systems in terms of pollution is as follows:

- The results of water quality monitoring from 2005 to 2018 have assessed water pollution indicators in the Bac Hung Hai irrigation system, including COD, $BOD_5$, $NH_4^+$, $NO_2^-$, $PO_4^{3-}$ and coliform. After more than 10 years, the COD content increased 8.6 times, $NH_4^+$ increased 2.48 times, $PO_4^{3-}$ increased 4.15 times, and coliform increased 91.6 times. The results of water pollution zoning of 83 rivers and canals based on the criteria of the water quality index (WQI), field descriptions of color, smell, and degree of impact on the life of living species in the river and canal show that all rivers have been polluted to different degrees, in which 19/83 rivers and canals are

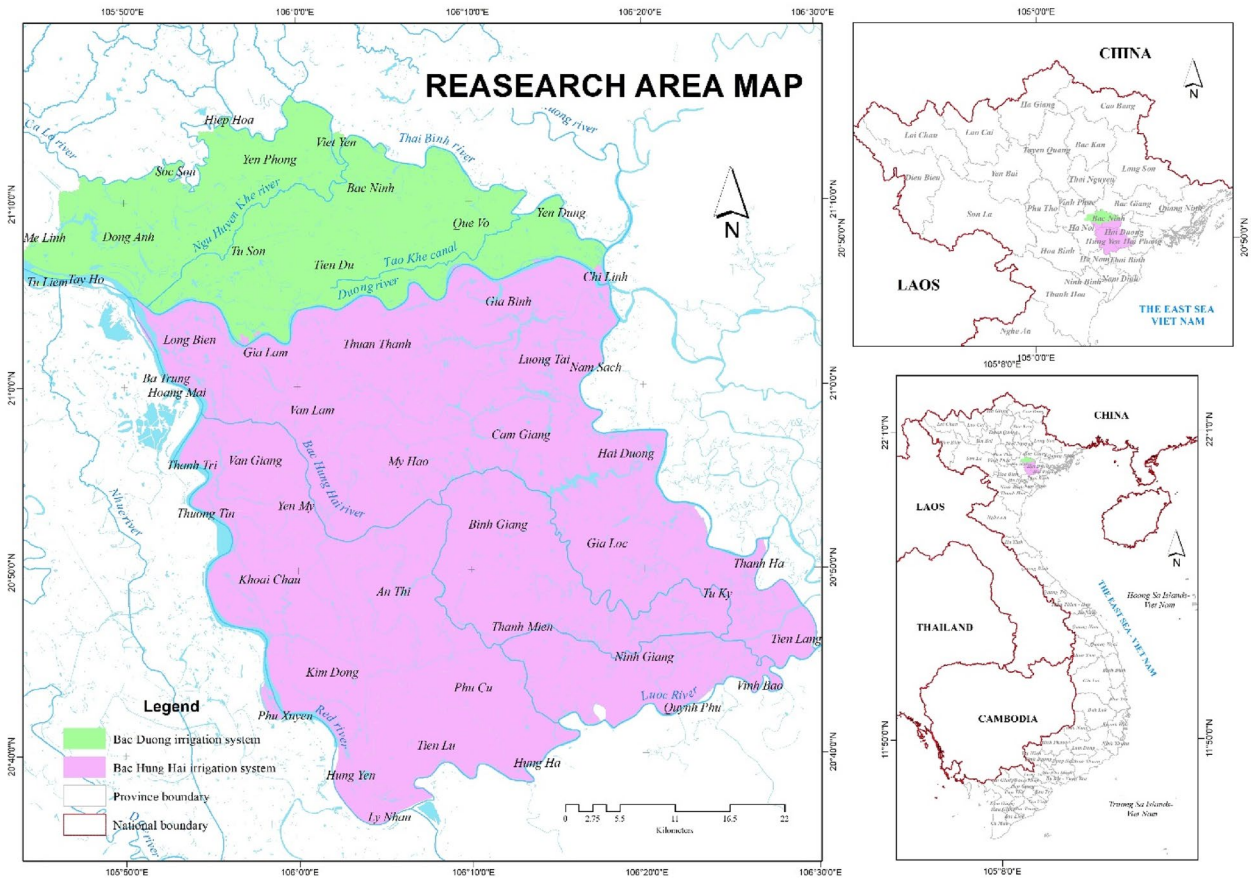


**Fig. 1** Geographical location of the study area

very seriously polluted, 21/83 rivers and canals are severely polluted, 23/83 rivers and canals are moderately polluted, and 20/83 rivers and canals are slightly polluted (Huong 2018).

- The water quality of the Bac Duong irrigation system is in a state of serious deterioration at many locations on the Ngu Huyen Khe River and some locations on the canal system. The results of monitoring the water quality in the Bac Duong irrigation system from 2007 to 2018 show that the water source is polluted mainly by the parameters DO, COD, $BOD_5$, $NH_4^+$, and coliform. The number of points with dissolved oxygen content lower than the allowable standard accounts for 30–100%; the percentage of points with COD content exceeding the standard is from 30 to 100%; the number of points with BOD5 content exceeding the standard through the monitoring sessions ranged from 30 to 90%; and the number of points with ammonium content exceeding the allowable standard is from 36.3 to 100%. Water quality in the dry season months, especially February, March, and April, is heavily polluted at all monitoring points. The results of calculating the water quality index (WQI) between the sampling periods show that at 50–94% of the monitoring points, the water quality is assessed as seriously polluted (Chinh 2019).

Currently, localities in the study area often use the water quality index (WQI) to assess surface water quality and the usability of water sources for different purposes and must rely on many parameters to calculate the WQI, and the calculation process is relatively complicated. According to Decision No. 1460/QD-TCMT issued by the Vietnam Environment Administration on technical guidance for the calculation and publication of Vietnam's water quality index (VN_WQI), the data to calculate VN_WQI must include at least 3/5 of the parameter groups, of which group IV (organic and nutritional parameters group) is required and there must be at least 3 parameters. In fact, localities often use 3 groups of parameters: Group I (pH); Group IV (DO, $BOD_5$, COD, TOC, N-$NH_4$, N-$NO_3$, N-$NO_2$, P-$PO_4$), and Group V (Coliform).

For the calculation of the WQI, it is necessary to monitor at least 10 of the above parameters. The monitoring of surface water quality in irrigation systems is still performed using the traditional method of collecting water samples, which are then analysed in the laboratory by various chemical and biological tests. These methods are often time consuming and labor intensive and can be expensive, especially when a large number of water samples are collected from different locations. In addition, this method can only provide water quality data at transient points in time, making it difficult to assess changes over time and space.

In recent years, machine learning and deep learning algorithms have been increasingly applied worldwide in calculating and forecasting water quality indices because of their ability to process large amounts of data and make predictions with high precision. Machine learning and deep learning algorithms can handle nonlinear relationships between water quality parameters and handle missing data and multidimensional data efficiently. Additionally, these algorithms can learn from data in real time and continuously improve their predictions as new data become available. This method has been shown to have many outstanding advantages (compared to traditional methods) for modelling complex nonlinear equations.

Forecasting the quality of surface water using machine learning models has been used in many places around the world. A decade-long research review on water quality indices in the field of artificial intelligence was carried out to develop the most feasible or most appropriate models and methods to be applied by researchers. In the future, in the field of water quality (Aminu 2022), the use of AI has increased dramatically in the last decade, yet there is still enough room for researchers to become involved and improve the calculations, projections, etc., of the water quality index. Some case studies, such as the prediction of the irrigation water quality index based on the machine learning and regression model of Mokhtar et al. (2022), have predicted the irrigation water quality index of the Bahr El-Baqr region. Egypt's research results indicate that the best model for prediction is the stepwise regression model, followed by principal component regression (PCR) and partial least squares regression (PLS) (Egypt). The prediction of river water quality index by data mining techniques (k-nearest neighbor, decision tree, naive Bayes, artificial neural network, support vector machine) was developed by Babbar and Babbar (2017) The results show that decision trees and support vector machine classifiers are considered to be the best predictive models. The IoT-based water quality index prediction for farm irrigation by Yadav et al. (2021) used 5 water quality parameters to calculate the irrigation water quality index (IWQI). The correlation analysis method was used to reduce five parameters to three. The results show that the random forest classification model is the best classification model for predicting water quality. Prediction of irrigation water quality indices based on machine learning algorithms in semiarid environments has also been applied; the study used five machine learning models to predict irrigation water quality indicators. which the SVM model is most suitable for all irrigation indicators (Dimple et al. 2022).

Improved water quality index prediction has also been made; a study by Mohd Zebaral Hoque et al. (2022) used eight machine learning regression models based on historical data from rivers in India to predict the water quality index. The results show that the linear and ridge regression models give the best performance. Improved prediction of water quality indices by a new hybrid machine learning algorithm studied by Bui et al. (2020), which used 4 independent machine learning algorithms and 12 hybrid algorithms to predict only water quality indicators. surface water quality in Iran. The results show that the best input matching models and the BA-RT matching algorithm outperform the others. Ibrahim et al. (2023) used integrated water quality indices, machine learning models and GIS approaches to predict groundwater quality for irrigation, and several irrigation water quality indices (IWQIs) and geographic information systems (GIS) were used to assess the groundwater (GW) quality for agricultural land in the El Kharga Oasis, Western Desert of Egypt. Two machine learning (ML) models (i.e., adaptive neuro-fuzzy inference system (ANFIS) and support vector machine (SVM)) were developed for the prediction of eight IWQIs. The performance of the simulation models was evaluated based on several prediction skill criteria, which revealed that the ANFIS model and SVM model were capable of simulating the IWQIs with reasonable accuracy. Abu El-Magd et al. (2023) integrated a machine learning-based model and WQI for groundwater quality assessment using support vector machines (SVMs) integrated with water quality indices (WQI) to assess groundwater quality. The SVM-WQI model shows a low percentage of the area for excellent class compared to the SVM model and WQI. Overall, the integrated ML model and WQI provide an understanding of water quality assessment, which may be helpful in the future development of such areas.

In addition to the classification of the water quality index based on a machine learning model for the Langat River basin (Shamsuddin et al. 2022), the study evaluates the effectiveness of machine learning models for multiclass classification in water quality assessment and evaluation found that SVM is the best model to predict river water quality. Ecosystem water quality index prediction and water quality classification of a heavily polluted river through supervised machine learning by Fernandez del Castillo et al. (2022) used supervised machine learning models. Monitoring can be used to predict the water quality index (SGR-WQI) for the ecosystem, with the number of water quality parameters reduced from 17 to 12 to expand the water quality monitoring program. Current water volume of the Santiago-Guadalajara River (Mexico).

Deep learning algorithms have also been used to predict and sort water quality indices. The study by Tiyasha et al. (2021) used an artificial intelligence model to predict the river water quality index and showed that the H2O deep learning model was the most accurate (for both large-scale watershed datasets small scale and large scale), followed by a random forest model. Hameed et al. (2016) applied artificial intelligence techniques to predict the water quality index. An ANN can be used to accurately predict the water quality index (WQI). The radial basis functional neural network (RBFNN) model is believed to be the most accurate for predicting WQI in tropical environments (Malaysia). The proposed method provides an efficient alternative to calculating and predicting the WQI, as manual calculation methods are very time-consuming. Aldhyani et al. (2020) developed an artificial intelligence (AI) algorithm to predict the water quality index (WQI) and water quality classification (WQC). The results show that the proposed models can accurately predict the WQI and classify water quality. Artificial neural network models (NARNET and LSTM) and machine learning algorithms (SVM, K-NN, and Naive Bayes) can accurately predict the water quality index (WQI) and the water quality classification (WQC). The NARNET model performed slightly better than the LSTM for predicting WQI values, and the SVM algorithm achieved the highest accuracy (97.01%) for WQC prediction. Ahmed et al. (2019) also used a supervised machine learning algorithm to estimate the water quality index (WQI). The results show that gradient enhancement and polynomial regression are the most efficient algorithms (MAE is 1.9642 and 2.7273, respectively). Multilayer perceptron (MLP) is the most effective for water quality grade classification (WQC). The proposed method achieves reasonable accuracy using the minimum number of parameters, making it suitable for real-time water quality detection systems.

In Vietnam, the use of machine learning models to predict the water quality index has been applied in the La Buong River (Khoi et al. 2022). This study evaluates the effectiveness of 12 machine learning models in predicting the water quality index. The results show that all 12 models have good performance in predicting the WQI, but the XGBoost model has the highest accuracy ($R^2 = 0.989$ and RMSE = 0.107). Than et al. (2016) applied an artificial neural network (ANN) to estimate the water quality index in the Dong Nai River flowing through two provinces, Dong Nai and Binh Duong. The research results have demonstrated that the predicted water quality index (WQI) is very significant and has a high correlation coefficient (R = 0.974 and p = 0.0) compared with the actual value of the WQI. Furthermore, ANN models provide

better predictive values than multivariate regression models.

In summary, previous studies on using deep learning in water quality forecasting have mainly focused on forecasting water quality parameters (physical parameters are the main ones) and calculating the water quality index (WQI). Some studies have also combined deep learning algorithms with real-time monitoring networks and have given very positive results. However, no study has applied the method of selecting important parameters from dozens of water quality parameters (monitoring) as input data to calculate the surface water quality index (WQI) by machine learning and deep learning models. Moreover, the above studies are popular worldwide. In Vietnam, there are very few studies evaluating the potential of machine learning algorithms and deep learning in forecasting the surface water quality index (WQI) based on data input (minimum water quality parameter) to reduce the cost of surface water quality monitoring, which is essential for developing countries.

Therefore, the study and application of machine learning models to predict the surface water quality index in the study area are important and necessary. The study will contribute to providing more scientific, effective, and cost-effective methods of calculating the surface water quality index to suit the actual conditions of localities in the Red River Delta. The objectives of the study are as follows:

- Building a scientific basis for calculating the surface water quality index using artificial intelligence;
- Propose a method to calculate the surface water quality index by machine learning and deep learning methods suitable to the actual conditions of irrigation systems in the Red River Delta.

## Materials and methods

### Implementation method

To achieve the stated objectives, the following research methods were used:

#### *Methods of collecting documents and data*

The data collection for this study will mainly focus on the collection of physical, chemical, and microbiological surface water quality data (temperature, pH, DO, $BOD_5$, COD, N-$NH_4$, N-$NO_3$, N-$NO_2$, P-$PO_4$, and coliform). Particularly for the WQI value at the monitoring sites, it is also collected together with data on water quality parameters in the study area from water quality monitoring reports, and data from previous studies will be collected and analysed to serve the construction machine learning and deep learning models.

#### *Data processing methods*

Data preparation and preprocessing were important steps in this study to ensure that the data were appropriate to eliminate any confounding factors or outliers that could affect the accuracy of the models. image. Includes the following steps:

- Data cleaning: collected data will be cleaned to address any missing or inconsistent values. Some commonly used methods for data cleaning include the following:
  - Handling missing values: Addressing missing data by imputing values or making decisions on how to handle the missing entries. This can involve techniques such as the mean imputation, regression imputation, or deletion of incomplete cases.
  - Correcting inconsistencies and outliers: Identifying and resolving inconsistencies, errors, or outliers in the data. This can involve data profiling, statistical methods, or domain-specific knowledge to detect and correct anomalies.
- Normalize data: metrics will be normalized to ensure that all variables (parameters) are on the same scale (dimensionless), which is important for the accuracy of algorithms in machine learning and deep learning. In this study, all data were normalized to fall between 0 and 1 to improve the convergence rate of the model and minimize the influence of the absolute scale. The normalization equation is as follows:

$$X_{norm} = \frac{X_0 - X_{min}}{X_{max} - X_{min}} \quad (1)$$

  where the X norm is the normalized value and $X_0$, $X_{min}$, and $X_{max}$ are the real value, the minimum value, and the maximum value of the same variable, respectively.
- Split data: The data will be divided into training datasets and test datasets. The training dataset is used to train the algorithms, while the test dataset is used to evaluate the accuracy of the prediction results. A commonly used ratio is 80:20, where 80% of the data are allocated for training and 20% for testing. This means that the model is trained on 80% of the data and evaluated on the remaining 20% (Joseph 2022).

#### *Bayes method (BMA)*

The Bayes method (BMA) exploits the Bayes factor (BF) and the index to measure the "compromise" between

the model's complexity and predictability (BIC) and choose the optimal model. This is a new method to overcome the problem of redundancy (the variable has no actual impact) in a multivariable linear regression model (Tuan 2020; Hinne et al. 2020).

Assume that there are m possible models with a parameter vector of $\theta_j$ that can explain γ. Suppose $P_j(\theta_j)$ is the probability of vector $\theta_j$. The probability density of γ can be written as:

$$P_j(y) = \int_{0-j} P_j(y|\theta_j) d\theta_j \tag{2}$$

the posterior probability of $\theta_j$ is:

$$P_j(\theta_j|y) = \frac{P_j(\gamma|\theta_j)P_j(\theta_j)}{P_j(y)} \tag{3}$$

If we have two models $M_1$ and $M_2$ and assume that one of them is true, the posterior probability of $M_1$ is:

$$P_j(M_1|y) = \frac{P(\gamma|M_1)P(M_1)}{P(y|M_1)P(M_1) + P(y|M_2)P(M_2)} \tag{4}$$

In fact, we can also compare the two models $M_1$ and $M_2$ through real evidence:

$$\frac{P(M_1|\gamma)}{P(M_2|\gamma)} = \frac{P(\gamma|M_1)}{P(\gamma|M_2)} \times \frac{P(M_1)}{P(M_2)} \tag{5}$$

This ratio is called the Bayes factor (BF). In the above interpretation, BF gives us information that the data are toward $M_1$ or $M_2$. With the BMA method, each study does not have only one model, but there can be many models that can also explain γ.

According to the water quality monitoring results, there are many water quality parameters, such as physical, chemical, and microbiological parameters (pH, TSS, DO, $BOD_5$, COD, $NH_4^-$, $PO_4^{3-}$, and coliform), that determine pollution, that is, the quality and amount of water (WQI). To determine the characteristic parameters for the machine learning model in the study area, the study used the Bayes method to identify variables (water quality parameters) that have a great influence on the WQI. Statistical analysis results by the Bayes method (BMA) will determine the water quality parameters that have a great influence on the WQI value, thereby determining the main parameters affecting the WQI.

#### *Methods of machine learning and deep learning*

*Machine learning algorithms* Based on the results of the overview study, the study uses machine learning models to calculate (predict) WQI with reinforcement learning algorithms because this is a powerful algorithm with many advantages and gives high computational results. algorithms with high accuracy that are easy to understand and easy to implement (Ahmed et al. 2019; Ni et al. 2020; Osman et al. 2021). Some of the main advantages of this algorithm are as follows:

- Interpretability: Gradient boosting easily interprets its predictions, as it follows a synchronous learning process. We can understand how the model makes a particular prediction by analysing the closure contribution of each weak learner model.
- Low risk of overfitting: less risk of overfitting because the algorithm works by gradually fitting data to each new weak learning model, which reduces the risk of overfitting with noisy data;
- Hidden algorithm selection: learn the most important features by continuously dividing the data based on the feature that provides the best division. This reduces the number of features required to make accurate predictions.

*Gradient boosting algorithms* are a set of superalgorithms that make weaker algorithms better at making predictions by reducing bias and variation in supervised learning problems. The basic principle of the accelerated approach is that it starts by creating a model from the training data and then proceeds to a second model based on the previous model, reducing the bias error incurred when the first model cannot infer relevant patterns from the given data. Every time a new learning algorithm is added, the weight of the data is adjusted again, also known as "reweighting". These models are added sequentially until the training data are reasonably predicted or the maximum number of learners has been added to the ensemble model (Schapire 2003). Full details of these enhancement-based algorithms can be found in Wu et al. (2020) and Bentéjac et al. (2021).

Boosting algorithms combine weak learners, i.e., learners slightly better than random, into a strong learner in an iterative way. Gradient boosting is a boosting-like algorithm for regression. Given a training dataset D = {xi, yi}$^N_1$, the goal of gradient boosting is to find an approximation $\widehat{F}_{(x)}$ of the function F*(x), which maps instances x to their output values y by minimizing the expected value of a given loss function L(y, F(x)). Gradient boosting builds an additive approximation of F*(x) as a weighted sum of functions (6):

$$F_m(x) = F_{m-1}(x) + \rho_m h_m(x) \tag{6}$$

where $\rho_m$ is the weight of the mth function, $h_m(x)$. These functions are the models of the ensemble (e.g., decision trees). The approximation is constructed iteratively.

*Extreme Gradient Boosting (XGBoost)* XGBoost is based on a model that assigns a higher weight to

misclassified data using a gradient boosting method. Boosting algorithm–based regression analysis, wherein each tree is based on a decision tree that is dependent on the previous tree, uses decision partitioning to generate step-by-step functionality. The specified loss function is optimized using the residuals from the previous tree (Shin et al. 2020).

When the first model is generated, the difference between the model predictions and observations is calculated (i.e., residuals or misclassifications). The different tree models can suitably predict the misclassification obtained in the first stage. The residuals remaining after the first two stages are matched to the other trees in the third stage, and the process is repeated several times.

The purpose of the model is simplification through optimizations of the training loss ($l$) and regulations ($\Omega$). $f_k$ is the function of the K-tree. The objective function (J) in round t is given by Eq. (7).

$$J^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \tag{7}$$

In this study, $y_i$ is the observed WQI, and $\widehat{y}_i$ is the obtained final prediction value.

*Deep learning algorithms* To predict water quality parameters for some irrigation systems in the Red River Delta, deep learning algorithms will be chosen based on how well they can find and process complex, nonlinear relationships in the data. Some commonly used algorithms for forecasting are as follows:

*Recurrent neural networks (RNNs)* RNNs are a type of deep learning algorithm that works well with continuous and multivariate data. RNNs are specifically designed to process sequence data, where each input data point depends on the previous data point. RNNs can handle input strings of different lengths. Furthermore, RNNs have the ability to store historical information in a hidden state, allowing them to make decisions based on past inputs. As such, RNNs are designed for sequential data processing and have been shown to perform well for water quality forecasting.

Unlike feed–forward neural networks, RNN delivers information in both directions, and the calculation computed from the initial input is fed back to the network, which is critical in learning the nonlinear relationships between multiple water quality variables. The equation's hidden state, at, is calculated using Eqs. (2–7). In the following equation, $W_1$ is the conventional weight between an input layer and the hidden layer, and $W_2$ is the matrix of recurrent weights between the hidden layer and itself at adjacent time steps. In other words, the RNN can reflect the previous hidden state in the current time process (Shin et al. 2020).

$$a_t = f(W_1 x_t + W_2 a_{t-1}) \tag{8}$$

*Short-long-term memory network (LSTM)* An LSTM network is a type of RNN designed to process time series of data. Input values fed to the LSTM not only pass through several LSTM layers but also propagate over time in an LSTM cell, resulting in a thorough input process in each time step. Overall, LSTM is a powerful tool for sequential data modelling and has several advantages over other RNN architectures in handling long-term dependencies, flexibility, and values. input efficiently, LSTM is also proven to be very effective in real-time water quality forecasting.

LSTM solves the problem using the interactions of three gating units and one memory cell. The input gate controls the degree to which a new value flows into the cell. The memory cell Ct can carry relevant information throughout the processing of the sequence. The memory cell reflects the old state value $C_{t-1}$ by the ratio of the forgotten gate ft and the new state value Cet by the ratio of the input gate. LSTM stores the previous state information in $C_{t-1}$ and uses it to determine the current state $C_t$. Finally, the output gate ot, through which the output is received, serves to adjust the output of the value stored in the memory cell $C_t$. One disadvantage of LSTM, however, is that the model has three gates; therefore, the number of weights and deviation terms required for learning are approximately four times larger. This leads to a long learning time and produces overfitting with less training data (Shin et al. 2020).

Through the above analysis, it is found that recurrent neural networks (RNNs) and LSTM are suitable for this study (Abba et al. 2020; Aldhyani et al. 2020; Ye et al. 2019). Therefore, this study will use this algorithm to build a model to predict the surface water quality index of the irrigation systems in the study area.

*Training and testing the model* The training process uses the training dataset that will be used to train the algorithms to recognize the parameters and their relationships in the dataset. The validation process involves using a test dataset to evaluate the accuracy of the algorithm. The following steps will be taken to train and validate the model:

- Data: surface water quality monitoring data from 2018 to 2022 (72 data points) including 8 parameters: $BOD_5$, $NH_4^+$, $PO_4^{3-}$, turbidity, TSS, Coliform, DO and WQI. The training set included 7 parameters (independent variables): $BOD_5$, $NH_4^+$, $PO_4^{3-}$, turbidity, TSS, coliform, and DO; the test set was WQI (dependent variable).

- Model training: machine learning and deep learning algorithms (RNN) will be trained using the training dataset to minimize the prediction error between actual water quality parameters and their forecast.
- Hyperparameter tuning: the parameters of machine learning and the deep learning algorithm will be adjusted to further improve accuracy.
    - Gradient boosting contains five tuning parameters that focus on the following: the distribution parameter specifies the distributional assumption for the response variable, which in this case is Gaussian or normally distributed; cv.folds indicates the number of cross-validation folds to conduct during model fitting; cv.folds indicates the number of cross-validation folds to perform during model fitting; cv.folds indicates the number of cross-validation folds. The shrinkage parameter governs the learning rate or step size at each boosting iteration. It calculates each tree's contribution to the overall model. A lower value, such as 0.01, usually results in higher model performance but may necessitate more iterations; n.minobsinnode, tt provides the minimal number of observations required in each terminal node of the boosted trees. Nodes with fewer than this number of observations will not be split further. The number of boosting iterations or trees to grow is given by n.trees.
    - eXtreme Gradient Boosting (XGBoost) contains three tuning parameters that focus on the following: the number of trees (nround); the shrinkage parameter (eta in the params), a small positive value; and the shrinkage parameter (eta in the params). This determines how quickly boosting learns. Typical values are 0.01 or 0.001, with the correct decision depending on the problem. To obtain good performance, a very small value of B may be needed. The number of splits in each tree that determines the complexity of the boosted ensemble (determined by max.depth).
    - RNNs and long short-term memory (LSTM) have four tuning parameters to focus on: epochs: the number of epochs during which the model should be trained; batch_size: the training batch size; validation_split: the proportion of training data utilized for validation. In addition, verbose: this controls whether or not progress updates are printed throughout training.
- Model testing: validated by the test dataset.
- Model selection: The best-performing algorithm (according to the calibrated parameters) will be selected based on the calibration results. The language used to code is R with Keras packages, which is a high-level neural network API running on top of TensorFlow. It was developed with a focus on enabling easy and fast design of complex deep learning models, as well as making them easier to train. Keras provides convenient methods for loading and preparing data, as well as tools to visualize and interpret training results.

*Evaluation of model accuracy* The following model accuracy indicators will be used to evaluate the model's accuracy in predicting the water quality index in the study area. The criteria for evaluating (calibrating) the models are presented in formulas (9)–(12):

- Mean absolute error (MAE) is the mean difference between the true value and the predicted value. MAE is a popular index to calculate error to evaluate (test) the model for continuous variables, determined by formula (9). where Pi is the predicted value and Mi is the actual measured value. The lower the MAE value is, the more accurate the calculations.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Pi - Mi|. \tag{9}$$

- The mean square error (MSE) of an estimator is the average of the squares of the errors, i.e., the difference between the predicted values and the actual measured values, and is calculated according to formula (10). The lower the MSE value is, the more accurate the calculations.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{10}$$

- RMSE is the square root of the mean of squared errors. The RMSE is a measure of how spread these residuals are; in other words, it tells you how concentrated the data are around the best-fit line. RMSE is the standard deviation of the residuals (prediction error) and is calculated according to formula (11). The lower the RMSE value is, the more accurate the calculation results.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(Q_A^i - Q_P^i)^2} \tag{11}$$

- The coefficient of determination ($R^2$) reflects the percentage of variance of y that can be explained by the model determined by formula (12). where ESS is the

sum of the squared deviations of the residuals and TSS is the sum of squared deviations. The $R^2$ value ranges from 0 to 1, and the closer the $R^2$ value is to 1, the more accurate the calculation results.

$$R^2 = 1 - (ESS/TSS) \quad (12)$$

The flowchart of the study structure is shown in Fig. 2.

## Results and discussion

### Results of collecting surface water quality monitoring data

The results of collecting surface water quality monitoring data from 2018 to 2022 at sampling locations in the study area are summarized in Table 1, and the evolution of some key water quality parameters is shown in Figs. 3, 4, 5, 6, 7 and 8. The surface water in the study area is mainly contaminated with organic matter, nutrients, and microorganisms. The parameters that exceed the allowable standards many times are DO, $BOD_5$, COD, $NH_4^+$ and total coliforms (these are also typical pollution parameters of the study area).

### Feature selection for machine learning and deep learning models

Feature selection is the method of reducing the input variable to the model by using only relevant data and eliminating noise in the data. According to the results of the correlation analysis between surface water quality parameters, COD and $BOD_5$ have the highest correlation (0.99); the correlation between DO and WQI is 0.48; and the correlation between $NH_4^+$ and $PO_4^{3-}$ is 0.47. In particular, the correlation between the WQI and parameters is not high, ranging from 0.05 to 0.48 (Fig. 9). Therefore, choosing (optimal) parameters to calculate the WQI by machine learning and deep learning models will be difficult.

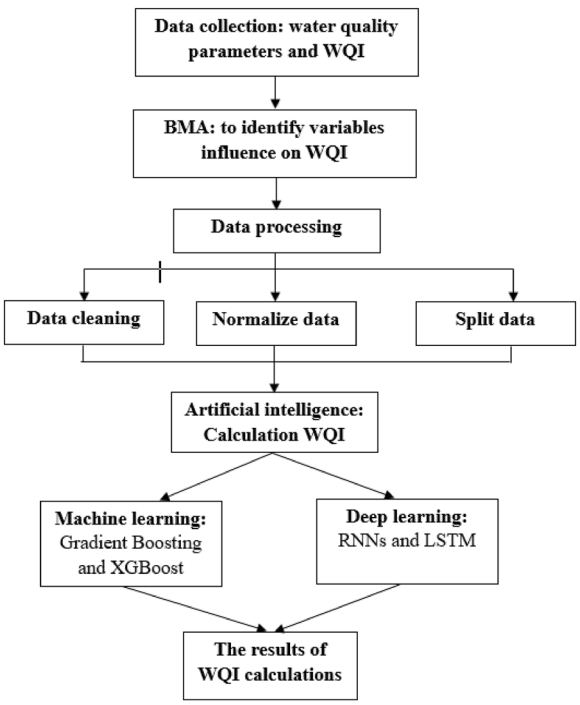


**Fig. 2** Flowchart of study structure

It can be seen that the correlation between the WQI and parameters is nonlinear; there are many water quality parameters, such as physical, chemical, and microbiological, that determine pollution, that is, water quality (here is the WQI value). To select the optimal parameters, the study applied the Bayes method (BMA). The results of the statistical analysis by BMA are shown in Table 2 and Fig. 10.

The results of selecting important water quality parameters by the BMA method are as follows:

- The probability of occurrence (according to the selected model) of each parameter affecting the WQI is as follows: $NH_4^+$, DO (100%), coliform (96.6%), $PO_4^{3-}$ (92.5%), $BOD_5$ (92.1%), turbidity (71.9%), and TSS (62.8%);
- There are 5 optimal models selected as follows:
  - Model 1: Seven parameters were selected as $BOD_5$, $NH_4^+$, $PO_4^{3-}$, turbidity, TSS, coliform and DO (posterior probability was 20.9%);
  - Model 2: 6 parameters were selected as $BOD_5$, $NH_4^+$, $PO_4^{3-}$, turbidity, coliform and DO (posterior probability was 19.3%);
  - Model 3: 6 parameters were selected as $BOD_5$, $NH_4^+$, $PO_4^{3-}$, TSS, Coliform and DO (posterior probability was 19%);
  - Model 4: 7 selected parameters are $BOD_5$, $NH_4^+$, $PO_4^{3-}$, turbidity, pH, coliform and DO (posterior probability is 5.4%);
  - Model 5: Seven parameters were selected as $BOD_5$, $NH_4^+$, $PO_4^{3-}$, TSS, pH, coliform and DO (posterior probability was 4.7%).

Based on the above analysis, model 1 is the best because it has the highest posterior probability (20.9%) and has found water quality parameters that have a large effect on the WQI value are $BOD_5$, $NH_4^+$, $PO_4^{3-}$, turbidity, TSS, coliform and DO. Therefore, Model 1 is chosen to calculate (predict) the WQI by machine learning and deep learning algorithms (which will be done in "Research on calculating the surface water quantity index by machine learning and deep learning methods" section).

**Table 1** Summary of typical values of surface water quality in the study area

| Parameters | Bac Duong (N = 384) | Bac Hung Hai (N = 395) | Overall (N = 779) |
|---|---|---|---|
| $BOD_5$ | | | |
| Mean (SD) | 28.8 (31.1) | 25.5 (19.1) | 27.1 (25.8) |
| Median [Min, Max] | 19.8 [4.90, 294] | 21.5 [0, 183] | 20.4 [0, 294] |
| *COD* | | | |
| Mean (SD) | 75.8 (85.6) | 64.1 (48.3) | 69.9 (69.4) |
| Median [Min, Max] | 51.2 [9.60, 816] | 55.1 [0, 451] | 52.8 [0, 816] |
| $NH_4$ | | | |
| Mean (SD) | 4.77 (4.66) | 7.78 (8.96) | 6.29 (7.32) |
| Median [Min, Max] | 3.42 [0.220, 35.5] | 4.31 [0, 46.8] | 3.86 [0, 46.8] |
| $PO_4$ | | | |
| Mean (SD) | 0.580 (1.04) | 0.589 (1.15) | 0.585 (1.09) |
| Median [Min, Max] | 0.200 [0.010, 7.32] | 0.0600 [0, 8.24] | 0.140 [0, 8.24] |
| *Turbidity* | | | |
| Mean (SD) | 36.2 (30.9) | 23.3 (13.6) | 29.6 (24.6) |
| Median [Min, Max] | 31.1 [10.1, 383] | 21.2 [8.21, 99.0] | 23.3 [8.21, 383] |
| *TSS* | | | |
| Mean (SD) | 45.6 (30.5) | 30.8 (16.5) | 38.1 (25.5) |
| Median [Min, Max] | 37.7 [10.5, 409] | 26.6 [2.01, 133] | 32.5 [2.01, 409] |
| *Coliform* | | | |
| Mean (SD) | 179,000 (1,260,000) | 865,000 (3,210,000) | 527,000 (2,470,000) |
| Median [Min, Max] | 14,000 [200, 16000000] | 24,000 [0, 16000000] | 19,000 [0, 16000000] |
| *DO* | | | |
| Mean (SD) | 3.46 (1.95) | 2.89 (1.56) | 3.17 (1.78) |
| Median [Min, Max] | 3.30 [0.0100, 11.3] | 2.70 [0, 7.36] | 3.00 [0, 11.3] |
| *Water temperature* | | | |
| Mean (SD) | 25.3 (4.82) | 24.4 (4.95) | 24.9 (4.90) |
| Median [Min, Max] | 24.8 [16.2, 35.8] | 24.6 [0, 33.4] | 24.7 [0, 35.8] |
| pH | | | |
| Mean (SD) | 7.32 (0.327) | 7.32 (0.505) | 7.32 (0.426) |
| Median [Min, Max] | 7.30 [5.69, 8.70] | 7.33 [0, 8.68] | 7.30 [0, 8.70] |
| *WQI* | | | |
| Mean (SD) | 24.1 (21.5) | 25.8 (24.7) | 25.0 (23.2) |
| Median [Min, Max] | 12.6 [0, 80.8] | 13.4 [0.780, 87.5] | 12.9 [0, 87.5] |

## Research on calculating the surface water quantity index by machine learning and deep learning methods

### *Calculation results of the surface water quality index*

Based on the results, select parameters for machine learning model building and deep research ($BOD_5$, $NH_4^+$, $PO_4^{3-}$, turbidity, TSS, coliform and DO). The study built a model from the above parameters to forecast only surface water quality (WQI) according to 4 models, namely, gradient boosting (GB), extreme gradient boosting (XGBoost), recurrent neural networks (RNN) and long short-term memory (LSTM). The results of hyperparameter tuning are shown in Table 3.

The results of the WQI report and comparison charts between the predicted and measured WQI values for the experimental dataset according to these 4 models are presented in Fig. 11.

### *Evaluation results for models*

Table 4 shows the results of the evaluation of the machine learning and deep learning models (based on 4 criteria) to predict the surface water quality index in the study area.

According to the calculation results, machine learning models are more accurate than deep learning models. The gradient boosting model has the most accurate

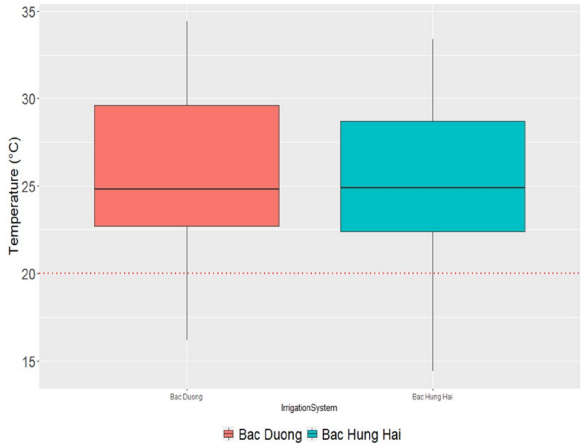


**Fig. 3** Temperature chart of Bac Duong and Bac Hung Hai irrigation system

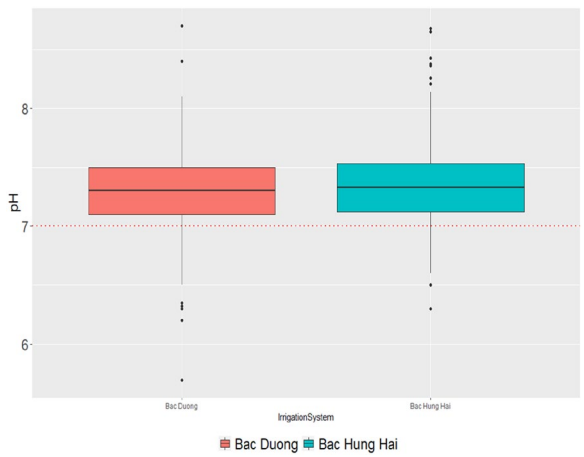


**Fig. 4** pH chart of Bac Duong and Bac Hung Hai irrigation system

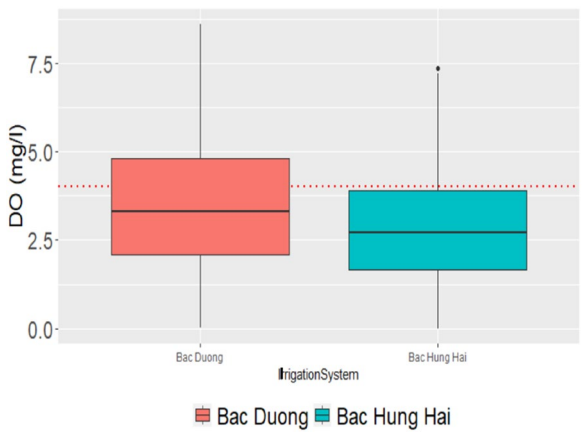


**Fig. 5** DO chart of Bac Duong and Bac Hung Hai irrigation system

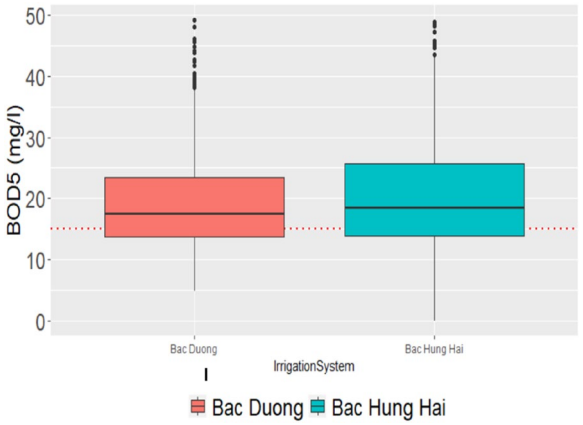


**Fig. 6** $BOD_5$ chart of Bac Duong and Bac Hung Hai irrigation system

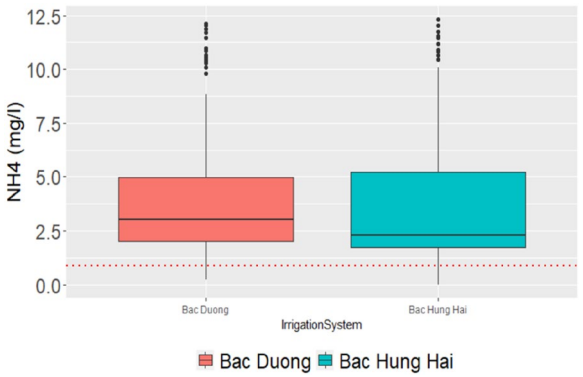


**Fig. 7** $NH_4^+$chart of Bac Duong and Bac Hung Hai irrigation system

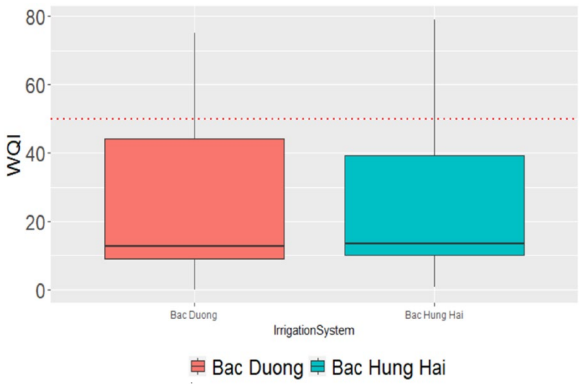


**Fig. 8** WQI chart of Bac Duong and Bac Hung Hai irrigation system

prediction results because it has the highest coefficient of determination ($R^2$ of 0.96) and the lowest values of errors (MAE, MSE, and RMSE) are 2.61, 19.9 and 4.46, respectively. Next is the XGBoost model with an $R^2$ of 0.89 and corresponding error values (3.70; 51.6; 7.18). The

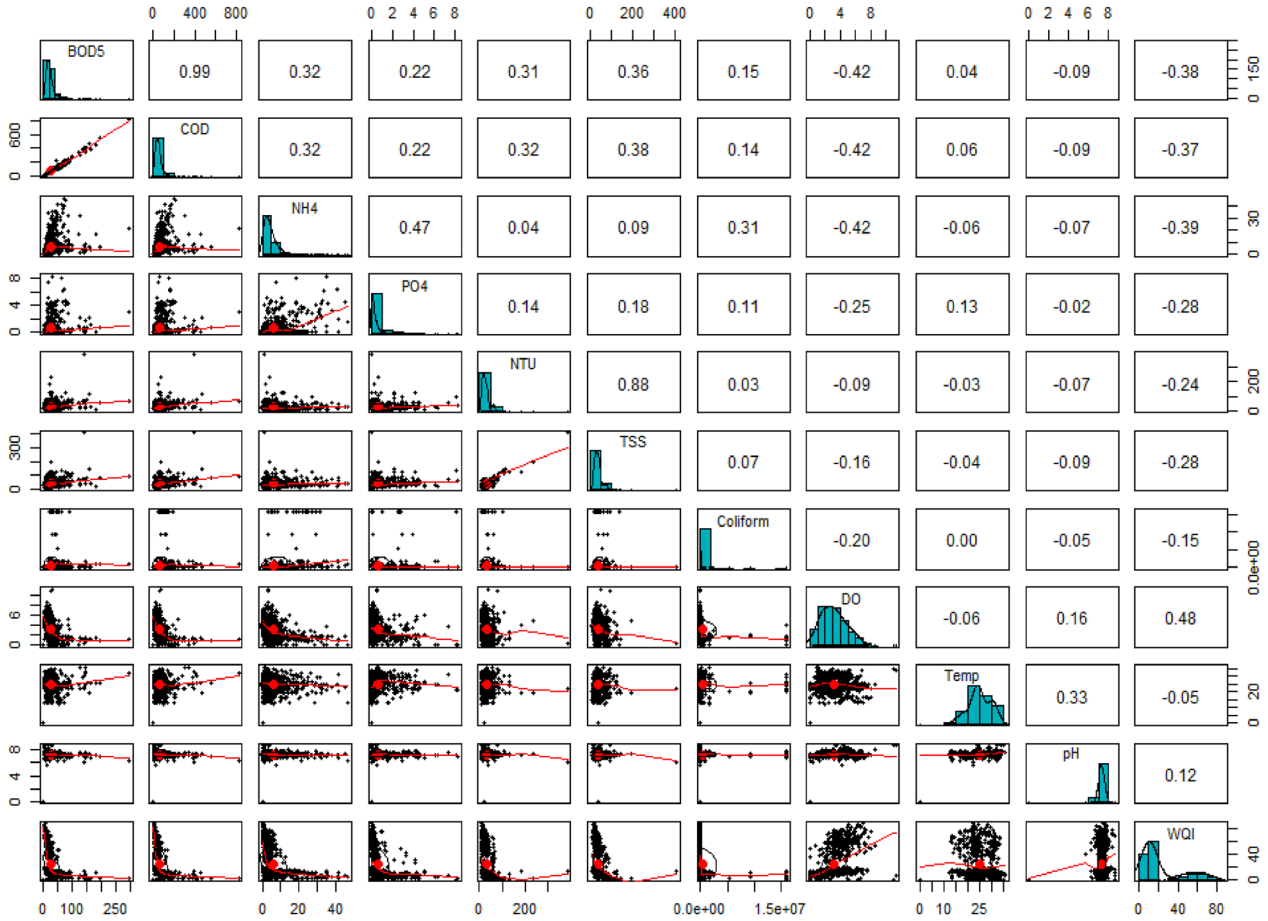


**Fig. 9** Correlation chart of water quality parameters

**Table 2** Summary of results of selected models by BMA method

| | p! = 0 | EV | SD | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|---|---|---|
| Intercept | 100.0 | 3.317e+01 | 1.404e+01 | 4.104e+01 | 3.783e+01 | 3.906e+01 | 8.806e+00 | 1.066e+01 |
| BOD5 | 92.1 | −5.236e−01 | 1.822e−01 | −5.718e−01 | −5.954e−01 | −5.784e−01 | −5.986e−01 | −5.823e−01 |
| COD | 15.2 | −2.276e−02 | 6.155e−02 | | | | | |
| NH4 | 100.0 | −1.403e+00 | 2.596e−01 | −1.404e+00 | −1.411e+00 | −1.400e+00 | −1.368e+00 | −1.358e+00 |
| PO4 | 92.5 | −6.661e+00 | 2.828e+00 | −7.106e+00 | −7.320e+00 | −7.047e+00 | −7.557e+00 | 7.288e+00 |
| NTU | 71.9 | −1.625e−01 | 1.212e−01 | −1.872e−01 | −2.603e−01 | | −2.535e−01 | |
| TSS | 62.8 | −1.255e−01 | 1.123e−01 | −1.712e−01 | | −2.385e−01 | | −2.309e−01 |
| Coliform | 96.6 | −7.360e−06 | 2.597e−06 | −7.576e−06 | −7.522e−06 | −7.994e−06 | −7.565e−06 | −8.025e−06 |
| DO | 100.0 | 4.149e+00 | 4.468e−01 | 4.101e+00 | 4.141e+00 | 4.092e+00 | 4.054e+00 | 4.009e+00 |
| Temp | 0.0 | 0.000e+00 | 0.000e+00 | | | | | |
| pH | 19.3 | 7.896e−01 | 1.841e+00 | | | | 3.973e+00 | 3.877e+00 |
| nVar | | | | 7 | 6 | 6 | 7 | 7 |
| $r^2$ | | | | 0.387 | 0.381 | 0.381 | 0.384 | 0.384 |
| BIC | | | | −3.34e+02 | −3.34e+02 | −3.34e+02 | −3.31e+02 | −3.31e+02 |
| post prob | | | | 0.209 | 0.193 | 0.190 | 0.054 | 0.047 |

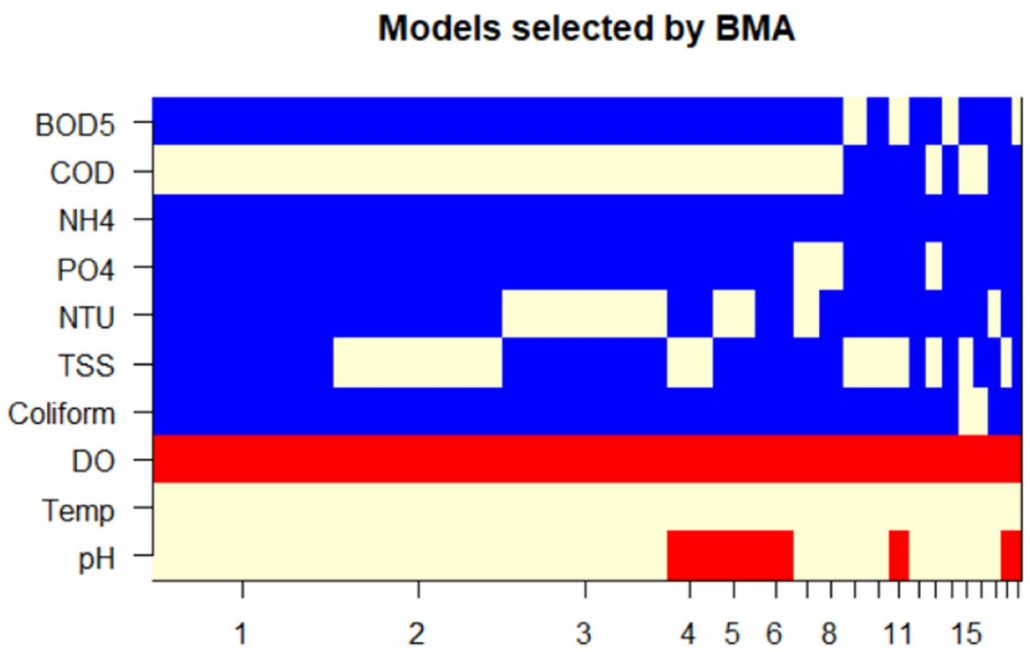


**Fig. 10** Graph of the selection of important water quality parameters

**Table 3** Table of results of hyperparameter tuning

| No | Model name | Hyperparameter tuning |
|---|---|---|
| 1 | Gradient boosting (GB) | Distribution = "Gaussian"<br>cv.folds = 10:<br>shrinkage parameter = 0.01<br>Each terminal node should have at least 10 observations: n.minobsinnode = 10<br>n.trees = 500 |
| 2 | eXtreme gradient boosting (XGBoost) | The number of trees (nround = 100);<br>The shrinkage parameter λ (eta in the params): 0.01;<br>The number of splits in each tree: max.depth = 5 |
| 3 | Recurrent neural networks (RNN) | learning_rate = 0.001<br>epochs = 500<br>batch_size = 32<br>validation_split = 0.2<br>verbose = 1 |
| 4 | Long short-term memory (LSTM) | learning_rate = 0.00001<br>epochs = 1000,<br>batch_size = 32,<br>validation_split = 0.2,<br>verbose = 1 |

RNN model has an $R^2$ of 0.84; the error values are 5.50, 76.6, and 8.75. The LSTM model has an $R^2$ of 0.85; the error values are 5.30, 71.0, and 8.42. The machine learning models applied in this study can all predict the WQI for the study area well (very high coefficient of determination, greater than 0.8). This is a solid scientific basis and an important result for being able to apply machine learning models in calculating WQI for other regions with similar conditions as the study area, especially in difficult conditions in monitoring of water quality parameters for calculation of WQI according to the traditional method.

## Discussion

According to the effectiveness evaluation of four machine learning and deep learning models (Table 4), two machine learning algorithms (gradient boosting and XGBoost) and two deep learning algorithms (RNN and

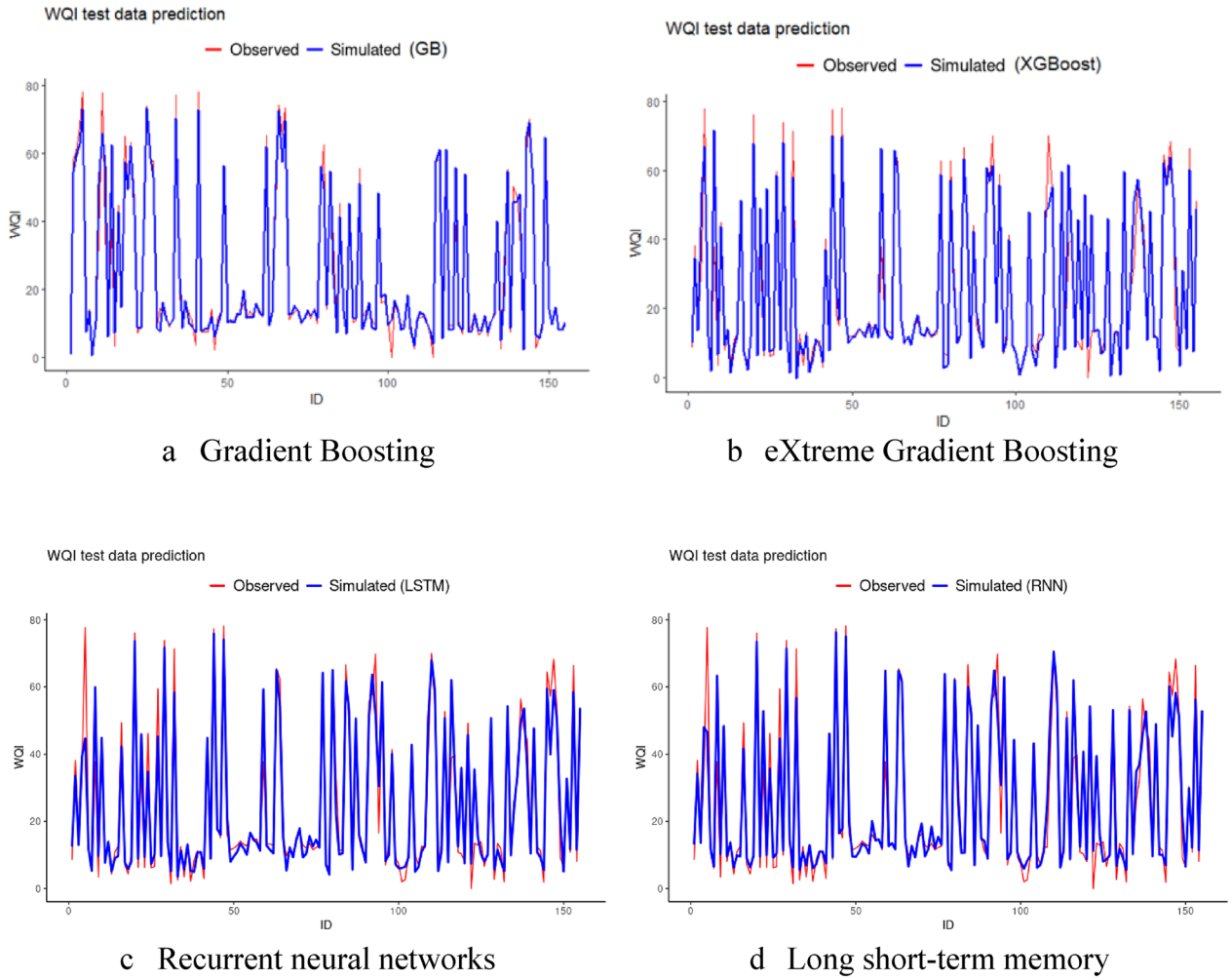


**Fig. 11** Comparison chart between forecast and actual WQI for the test data

**Table 4** Statistical table of evaluation results of models to predict the surface water quality index in the study area

| Models | Input parameters | Output | Evaluation criteria | | | |
|---|---|---|---|---|---|---|
| | | | MAE | MSE | RMSE | $R^2$ |
| Gradient Boosting | $BOD_5$, $NH_4^+$, $PO_4^{3-}$, Turbidity, TSS, Coliform and DO | WQI | 2.61 | 19.9 | 4.46 | 0.96 |
| XGBoost | $BOD_5$, $NH_4^+$, $PO_4^{3-}$, Turbidity, TSS, Coliform and DO | WQI | 3.70 | 51.6 | 7.18 | 0.89 |
| RNN | $BOD_5$, $NH_4^+$, $PO_4^{3-}$, Turbidity, TSS, Coliform and DO | WQI | 5.50 | 76.6 | 8.75 | 0.84 |
| LSTM | $BOD_5$, $NH_4^+$, $PO_4^{3-}$, Turbidity, TSS, Coliform and DO | WQI | 5.30 | 71.0 | 8.42 | 0.85 |

LSTM) performed most effectively. With coefficients of determination ranging from 0.84 (RNN) to 0.96 (gradient boosting), machine learning and deep learning models can accurately predict the WQI for the study area. This is because each algorithm reacts differently to distinct input variables and data samples (Hussain and Khan 2020; Khoi et al. 2022). According to the findings of Morton and Henderson (2008) and Yang and Moyer (2020), the distribution of the water quality data is nonlinear (Khoi et al. 2022). Consistent with the findings of Hussain and Khan, these findings indicate that the most accurate prediction depends on the model parameters for the given scenario of the input variables (Hussain and Khan 2020).

Comparing each of the four machine learning and deep learning models demonstrates that the gradient boosting model outperforms the others in the research domain. Compared to other studies, XGBoost is the most appropriate (machine learning) algorithm for the La Buong River basin (Vietnam) (Khoi et al. 2022); the random forest model has the highest predictive accuracy for WQI values in the An Kim Hai irrigation system (Vietnam) (Lap et al. 2023). DFNN outperformed XGBoost, MLP, and RF in India's Mahanadi River Basin (Singha et al. 2021). Asadollah et al. found that ExT performed better than DT and supported vector regression (SVR) in Hong Kong's Lam Tseen River. Furthermore, DT performed better than MLP in Pakistan's Rawal Dam Lake (Ahmed et al. 2021). In general, the performance of various machine learning and deep learning algorithms will vary when applied to different regions. Consequently, finding and developing a generalized deep learning and machine learning model for water quality assessment applications is an ongoing challenge (Khoi et al. 2022).

The absence of consideration of the cross-effects between the explanatory variables, specifically the cross-correlation between land use types and climate conditions, was a significant gap in previous research affecting the water quality in irrigation systems (Kouadri et al. 2021; Kung and Wu 2021; Kung and Mu 2019; Amanullah et al. 2020). Changes in land use (Ahmad et al. 2021) and climate change affect hydrological components and, consequently, river discharge and the transport of pollutants (Khoi et al. 2019). In addition, the operation of irrigation facilities and the decreased quantity of water supply to irrigation systems both contribute to pollution (Sulaeman et al. 2018). To improve the accuracy of machine learning and deep learning models, it is essential to consider land use, operating modes, water depletion, and climate change.

## Conclusion

The findings of the research not only offered a way to calculate the surface water quality index using artificial intelligence (machine learning and deep learning) but also offered a scientific basis for doing so. In the parts of the research field where the machine learning approach is implemented, it performs quite admirably. In this work, the Bayes technique, also known as BMA, was utilized to choose (optimal) parameters for the purpose of developing the WQI computer learning model. A total of seven parameters were chosen for inclusion in the model, including DO, $BOD_5$, $NH_4^+$, $PO_4^{3-}$, turbidity, TSS and coliform (fewer than with the traditional method).

The results of the WQI calculations for the two types of machine learning models indicate that the machine learning model provides more accurate predictions than the deep learning model. The gradient boosting model produces the most accurate predictions of the available models. After that comes the eXtreme Gradient Boosting model, also known as XGBoost, followed by the RNN model and the LSTM model. The accuracy of each of these models is very high, ranging from 84 to 96%.

The outcomes demonstrate that applying machine learning and deep learning algorithms can significantly reduce the number of water quality parameters without compromising model accuracy. Therefore, machine learning and deep learning models are both capable of calculating (predicting) the WQI for the area under study with a high level of precision and may be applied to other regions that have characteristics that are comparable (especially for developing countries such as Vietnam). This will help developing countries, which are still struggling in surface water quality monitoring, improve their assessment and management of surface water quality.

Our study obviously suffers from some limitations that should be addressed in future work: there might be multicollinearity causing overfitting problems because the water quality variables used in this study are closely related to each other. Thus, further investigations should be performed to overcome these limitations. Possibly applicable methods are to utilize regularization techniques such as ridge, lasso, and elastic net to solve overfitting problems.

#### Acknowledgements

The author would like to thank the steering committee of the project "Study on the Impact of Water Supply on Water Pollution in Irrigation Systems in the Red River Delta and Propose Solutions" of Assoc. Dr. Vu Thi Thanh Huong provided data on the water quality of the Red River Delta Irrigation System for this study.

#### Author contributions

NDP and HHD contributed equally to this study. TNT and NMT conducted the statistical analysis and prepared the figures. All authors read and approved the final manuscript.

#### Funding

This study was not funded by any organization.

#### Availability of data and materials

The data and materials used in this study are available upon request. Please contact Nguyen Duc Phong (Email Address: phongndtv@gmail.com for further information.

### Declarations

#### Ethics approval and consent to participate

This study was conducted following the ethical guidelines set.

#### Consent for publication

All participants have given written consent for their data to be used in this research study and for the results to be published in a scientific journal.

**Competing interests**
The authors declare that they have no competing interests.

Received: 1 May 2023 Accepted: 12 June 2023
Published online: 04 July 2023

**References**

Abba SI, Pham QB, Saini G, Linh NTN, Ahmed AN, Mohajane M, Khaledian M, Abdulkadir RA, Bach Q-V (2020) Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index. Environ Sci Pollut Res 27:41524–41539

Abu El-Magd SA, Ismael IS, El-Sabri MAS, Abdo MS, Farhat HI (2023) Integrated machine learning–based model and WQI for groundwater quality assessment: ML, geospatial, and hydroindex approaches. Environ Sci Pollut Res 30(18):53862–53875. https://doi.org/10.1007/s11356-023-25938-1

Ahmad W, Iqbal J, Nasir MJ, Ahmad B, Khan MT, Khan SN, Adnan S (2021) Impact of land use/land cover changes on water quality and human health in district Peshawar Pakistan. Sci Rep 11(1):16526. https://doi.org/10.1038/s41598-021-96075-3

Ahmed M, Mumtaz R, Hassan Zaidi SM (2021) Analysis of water quality indices and machine learning techniques for rating water pollution: a case study of Rawal Dam, Pakistan. Water Supply 21(6):3225–3250. https://doi.org/10.2166/ws.2021.082

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

Aldhyani THH, Al-Yaari M, Alkahtani H, Maashi M (2020) Water quality prediction using artificial intelligence algorithms. Appl Bionics Biomech 2020:1–12. https://doi.org/10.1155/2020/6659314

Amanullah, Khalid, S., Imran, Khan, H. A., Arif, M., Altawaha, A. R., . . . Parmar, B. (2020). Effects of Climate Change on Irrigation Water Quality. In S. Fahad, M. Hasanuzzaman, M. Alam, H. Ullah, M. Saeed, I. Ali Khan, & M. Adnan (Eds.), Environment, Climate, Plant and Vegetation Growth (pp. 123-132). Cham: Springer International Publishing.

Aminu II (2022) A novel approach to predict water quality index using machine learning models: a review of the methods employed and future possibilities. Global J Eng Technol Adv 13(2):026–037. https://doi.org/10.30574/gjeta.2022.13.2.0184

Babbar R, Babbar S (2017) Predicting river water quality index using data mining techniques. Environ Earth Sci. https://doi.org/10.1007/s12665-017-6845-9

Bentéjac C, Csörgő A, Martínez-Muñoz G (2021) A comparative analysis of gradient boosting algorithms. Artif Intell Rev 54(3):1937–1967. https://doi.org/10.1007/s10462-020-09896-5

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

Chinh VQ (2019) Monitoring and forecasting water quality in the irrigation system of Bac Duong for agricultural production in 2019. Retrieved from Ha Noi

Dimple D, Rajput J, Al-Ansari N, Elbeltagi A (2022) Predicting irrigation water quality indices based on data-driven algorithms: case study in semiarid environment. J Chem 2022:1–17. https://doi.org/10.1155/2022/4488446

Fernandez del Castillo A, Yebra-Montes C, Verduzco Garibay M, de Anda JE, Garcia-Gonzalez A, Gradilla-Hernández MSA (2022) Simple prediction of an ecosystem-specific water quality index and the water quality classification of a highly polluted river through supervised machine learning. Water 14(8):1235. https://doi.org/10.3390/w14081235

Hameed M, Sharqi SS, Yaseen ZM, Afan HA, Hussain A, Elshafie A (2016) Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia. Neural Comput Appl 28(S1):893–905. https://doi.org/10.1007/s00521-016-2404-7

Hinne M, Gronau QF, van den Bergh D, Wagenmakers E-J (2020) A conceptual introduction to Bayesian model averaging. Adv Methods Pract Psychol Sci 3(2):200–215. https://doi.org/10.1177/2515245919898657

Huong VTT (2018) Research and propose solutions to reduce water pollution in Bac Hung Hai irrigation system. Retrieved from Ha Noi

Hussain D, Khan AA (2020) Machine learning techniques for monthly river flow forecasting of Hunza River, Pakistan. Earth Sci Inform 13(3):939–949. https://doi.org/10.1007/s12145-020-00450-z

Ibrahim H, Yaseen ZM, Scholz M, Ali M, Gad M, Elsayed S, Khadr M, Hussein H, Ibrahim HH, Eid MH, Kovács A, Péter S, Khalifa MM (2023) Evaluation and prediction of groundwater quality for irrigation using an integrated water quality indices, machine learning models and GIS approaches: a representative case study. Water 15(4):694. https://doi.org/10.3390/w15040694

Joseph VR (2022) Optimal ratio for data splitting. Stat Anal Data Min ASA Data Sci J 15(4):531–538. https://doi.org/10.1002/sam.11583

Khoi DN, Nguyen VT, Sam TT, Nhi PT (2019) Evaluation on effects of climate and land-use changes on streamflow and water quality in the La Buong River Basin, Southern Vietnam. Sustainability. https://doi.org/10.3390/su11247221

Khoi DN, Quan NT, Linh DQ, Nhi PTT, Thuy NTD (2022) Using machine learning models for predicting the water quality index in the La Buong River, Vietnam. Water 14(10):1552. https://doi.org/10.3390/w14101552

Kouadri S, Elbeltagi A, Islam ARMT, Kateb S (2021) Performance of machine learning methods in predicting water quality index based on irregular dataset: application on Illizi region (Algerian southeast). Appl Water Sci 11(12):190. https://doi.org/10.1007/s13201-021-01528-9

Kung C-C, Mu JE (2019) Prospect of China's renewable energy development from pyrolysis and biochar applications under climate change. Renew Sustain Energy Rev 114:109343. https://doi.org/10.1016/j.rser.2019.109343

Kung C-C, Wu T (2021) Influence of water allocation on bioenergy production under climate change: a stochastic mathematical programming approach. Energy 231:120955. https://doi.org/10.1016/j.energy.2021.120955

Lap BQ, Phan T-T-H, Nguyen HD, Quang LX, Hang PT, Phi NQ, Hoang VT, Linh PG, Hang BTT (2023) Predicting Water Quality Index (WQI) by feature selection and machine learning: a case study of An Kim Hai irrigation system. Ecol Inform 74:101991. https://doi.org/10.1016/j.ecoinf.2023.101991

Mohd Zebaral Hoque J, Ab. Aziz NA, Alelyani S, Mohana M, Hosain M (2022) Improving water quality index prediction using regression learning models. Int J Environ Res Public Health 19(20):13702. https://doi.org/10.3390/ijerph192013702

Mokhtar A, Elbeltagi A, Gyasi-Agyei Y, Al-Ansari N, Abdel-Fattah MK (2022) Prediction of irrigation water quality indices based on machine learning and regression models. Appl Water Sci. https://doi.org/10.1007/s13201-022-01590-x

Morton R, Henderson BL (2008) Estimation of nonlinear trends in water quality: an improved approach using generalized additive models. Water Resour Res. https://doi.org/10.1029/2007wr006191

Ni L, Wang D, Wu J, Wang Y, Tao Y, Zhang J, Liu J-F (2020) Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model. J Hydrol 586:124901

Osman AAA, Ahmed AN, Chow MF, Huang YF, El-Shafie A (2021) Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Eng J 12:1545–1556

Schapire RE (2003) The boosting approach to machine learning: an overview. In: Denison DD, Hansen MH, Holmes CC, Mallick B, Yu B (eds) Nonlinear estimation and classification. Springer, New York, pp 149–171

Shamsuddin IIS, Othman Z, Sani NS (2022) Water quality index classification based on machine learning: a case from the Langat River Basin Model. Water 14(19):2939. https://doi.org/10.3390/w14192939

Shin Y, Kim T, Hong S, Lee S, Lee E, Hong S, Lee C, Kim T, Park MS, Park J, Heo T-Y (2020) Prediction of chlorophyll-a concentrations in the Nakdong River using machine learning methods. Water 12(6):1822. https://doi.org/10.3390/w12061822

Singha S, Pasupuleti S, Singha SS, Singh R, Kumar S (2021) Prediction of groundwater quality using efficient machine learning technique. Chemosphere 276:130265. https://doi.org/10.1016/j.chemosphere.2021.130265

Sulaeman D, Arif S, Sudarmadji S (2018) Trash-polluted irrigation: characteristics and impact on agriculture. IOP Conf Ser Earth Environ Sci 148:012028. https://doi.org/10.1088/1755-1315/148/1/012028

Than NH, Ly CD, Tat PV, Thanh NN (2016) Application of a neural network technique for prediction of the water quality index in the Dong Nai River,

Vietnam. J Environ Sci Eng B 5:7. https://doi.org/10.17265/2162-5263/2016.07.007
Tiyasha T, T. M., & Yaseen, Z. M. (2021) Deep learning for prediction of water quality index classification: tropical catchment environmental assessment. Nat Resour Res 30(6):4235–4254. https://doi.org/10.1007/s11053-021-09922-5
Tuan NV (2020) Regression modelling and scientific discovery. General Publishing House, Ho Chi Minh City
Wu T, Zhang W, Jiao X, Guo W, Hamoud YA (2020) Comparison of five boosting-based models for estimating daily reference evapotranspiration with limited meteorological variables. PLoS ONE 15(6):e0235324. https://doi.org/10.1371/journal.pone.0235324
R. K. Yadav, A. Jha and A. Choudhary, "IoT based prediction of water quality index for farm irrigation," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 1443-1448, doi: 10.1109/ICAIS50930.2021.9395921.
Yang G, Moyer DL (2020) Estimation of nonlinear water-quality trends in high-frequency monitoring data. Sci Total Environ 715:136686. https://doi.org/10.1016/j.scitotenv.2020.136686
Ye Q, Yang X, Chen C, Wang J (2019) River water quality parameters prediction method based on LSTM-RNN model. In: 2019 Chinese control and decision conference (CCDC), pp 3024–3028

## Publisher's Note

**Nguyen Duc Phong** is a PhD student in Land and water Environment at Vietnam Academy for Water Resources.

**Ha Hai Duong** has a PhD in water resources and modelling at the Institute for Water and Environment.

**Trinh Ngoc Thang** is a research fellow in water resources and the environment at the Institute for Water and Environment.

**Nguyen Minh Tu** is a research fellow in water resources at the Institute for Water and Environment.