Environmental Sciences Europe

RESEARCH

Open Access

Developing an ensembled machine learning model for predicting water quality index in Johor River Basin

L. M. Sidek[1*], H. A. Mohiyaden[2], M. Marufuzzaman[1], N. S. M. Noh[3], Salim Heddam[4], Mohammad Ehteram[5], Ozgur Kisi[6,7*] and Saad Sh. Sammen[8]

Abstract

Currently, the Water Quality Index (WQI) model becomes a widely used tool to evaluate surface water quality for agriculture, domestic and industrial. WQI is one of the simplest mathematical tools that can assist water operator in decision making in assessing the quality of water and it is widely used in the last years. The water quality analysis and prediction is conducted for Johor River Basin incorporating the upstream to downstream water quality monitoring station data of the river. In this research, the numerical method is first used to calculate the WQI and identify the classes for validating the prediction results. Then, two ensemble and optimized machine learning models including gradient boosting regression (GB) and random forest regression (RF) are employed to predict the WQI. The study area selected is the Johor River basin located in Johor, Peninsular Malaysia. The initial phase of this study involves analyzing all available data on parameters concerning the river, aiming to gain a comprehensive understanding of the overall water quality within the river basin. Through temporal analysis, it was determined that Mg, *E. coli*, SS, and DS emerge as critical factors affecting water quality in this river basin. Then, in terms of WQI calculation, feature importance method is used to identify the most important parameters that can be used to predict the WQI. Finally, an ensemble-based machine learning model is designed to predict the WQI using three parameters. Two ensemble ML approaches are chosen to predict the WQI in the study area and achieved a R^2 of 0.86 for RF-based regression and 0.85 for GB-based ML technique. Finally, this research proves that using only the biochemical oxygen demand (BOD), the chemical oxygen demand (COD) and percentage of dissolved oxygen (DO%), the WQI can be predicted accurately and almost 96 times out of 100 sample, the water class can be predicted using GB ensembled ML algorithm. Moving forward, stakeholders may opt to integrate this research into their analyses, potentially yielding economic reliability and time savings.

Keywords Water quality index, Gradient boosting regression, Random forest, Johor River

*Correspondence:
L. M. Sidek
Lariyah@uniten.edu.my
Ozgur Kisi
ozgur.kisi@th-luebeck.de; ozgur.kisi@iliauni.edu.ge
Full list of author information is available at the end of the article

Introduction

Water is classified as a renewable resource due to its continual circulation through the hydrologic cycle, and the fact that approximately 70% of the Earth's surface is covered by water [1]. River water plays a significant role as primary sources for drinking, water supply, hydropower generation, agricultural, industrial activities, livestock production and other economic sectors [2]. In various developing countries, water quality is tremendously deteriorating due to natural effect and human activities that lead to pollution. Pollution to rivers or any controlled waters can be classified into three distinct categories i.e., isolated pollution incidents, non-point (diffused) sources of pollution and point sources pollution. Isolated pollution incidental spillage, illegal dumping of pollution or failure of treatment processes or plants lead to bad quality effluent being discharged to the river. The natural factors that deteriorate water quality are hydrological cycle, atmospheric, climatic, and topographic changes over time. Meanwhile, the example of environmental pollution caused by human activities is from industrial, municipal and agricultural production and disposal of waste, animal husbandry, mining, sedimentation or soil erosion due to rapid land use changes and contribution of heavy metals into the water body [3].

Emergence of threats or pollution to the water resources, urges the related parties to conduct a water quality monitoring as a part of measures to improve in managing this valuable resources. Currently, there are recommendations to apply integrated river basin management as an integration in managing water resources. This approach must meet several listed criteria to ensure the river basin comply with water quality standards and these also considering protection of aquatic ecology and its habitat [4]. For instance, Malaysia is currently developing and already completed numbers of Integrated River Basin Management (IRBM) frameworks and documentation. This approach also aims in managing water resources in terms of its quality, quantity, sufficiency, and to improve the environment. Water quality assessment consists of three main aspects which are chemical, biological, and physical characteristics of water. All aspects are being determined by a set of standards known as National Water Quality Standards (NWQS) and classified based on the water beneficial uses as endorsed by Department of Environment (DOE). Currently, the Water Quality Index (WQI) model becomes a widely used tool to evaluate surface water quality for agriculture, domestic and industrial [3, 5]. WQI is accepted to be used in assessing the quality of water because it is one of the simplest mathematical tools that can assist water operator in decision making.

WQI calculation is based on NWQS focus on the specific parameter and climate condition [6] and has certain uncertainty and risk of faulty results. According to Bui et al. [7], there are many disadvantages of conventional WQI Equation including requirement of lengthy, complex and inconsistent techniques. One need to have complete details parameter results to calculate the correct WQI. If one parameter is missing, the WQI Equation cannot be computed. At present, assessing water quality involves costly and time-intensive procedures involving laboratory testing and statistical analysis. This process entails the collection of water samples, their transportation to laboratories, and a substantial amount of time and computational work. This approach is not very efficient, especially considering that water can easily transmit contaminants, and prompt action is crucial if the water is contaminated with disease-causing waste. Consequently, the dire consequences of water pollution underscore the need for a more rapid and cost-effective alternative. In many decades, researchers and water operators have used different approaches to determine the quality of surface water. Numerous types of water quality models have been applied to improve the accuracy of water quality predictions to ease in decision making. Development of model for surface water quality related parameters becomes a crucial issue in terms of its efficiency and reliability to produce a good result. The reliable water quality model may result in significant reduction of costing since it is indirectly able to determine the values of water quality-related parameters [8]. Apart from that, the usage of WQI calculation is time consuming and unintentionally associated with errors during data collection and sub-index calculation [9]. In terms of classification principles, WQI categorization is classified as supervised pattern recognition method [10].

Nowadays, many studies opted artificial intelligence (AI) to predict WQI method [11–17]. The application of Artificial Neural Network (ANN) as part of machine learning is the powerful computational method as it can represent a complex mathematical model and it has the capability to learn based on the concepts, patterns, observations or any series of data by the process called training the model [18, 19]. This method implements a model structure of neural network to capture intricate non-linear connections, especially in situations where the relationships between variables are not well understood. There are many water-quality studies have been conducted the implement the application of machine learning. Rankovic et al. modeled dissolved oxygen (DO) in the Gruza reservoir, Serbia using ANN and obtained promising results [20]. A study conducted at Cheongpyeong Dam, Korea has applied ANN to predict eight

water quality parameters which are temperature, DO, pH, conductivity, TN, TP turbidity and Chlorophyll-a using three years observed daily data and found that 7 parameters except for turbidity have an R^2 higher than 0.85, meanwhile 5 parameters; temperature, DO, pH, TN and TP shows have an RMSE of 1.0 [21]. Another previous study carried out in Gomti River, India has tested ANN to compute two parameters DO and biochemical oxygen demand (BOD) levels. It employs eleven water quality parameters as inputs that measured monthly for 10 years. It has been identified that the optimal networks are able to analyze long-term trends observed data for the sensitive parameters like DO and BOD. The study also proposes that neural networks model is suitable to be used to compute and predict river water quality and can determine the pollution trends [22]. Sakizadeh employed an ANN with Bayesian regularization to predict the WQI using 16 water quality parameters as inputs [23]. The study achieved correlation coefficients of 0.94 and 0.77, demonstrating a strong predictive performance. Abyaneh, on the other hand, focused on predicting the Chemical Oxygen Demand (COD) and BOD using conventional methods, including ANN and multi-linear regression [24]. They utilized four-parameter, namely pH, total suspended solids (TSS), temperature, and total suspended (TS), to forecast COD and BOD. An unsupervised technique (average linkage hierarchical clustering) was adopted by Ali and Qamar to classify water samples into distinct water quality groups. However, their approach omitted essential parameters related to WQI and did not employ a standardized water quality index for assessment [25]. However, their approach omitted essential parameters related to WQI and did not employ a standardized water quality index for assessment. Gazzaz et al. applied ANN to predict WQI and achieved a model that explained nearly 99.5% of the data's variation. Their model utilized 23 parameters, which could be costly when implementing it in an Internet of Things (IoT) system due to sensor prices [9]. Single feed-forward ANN was employed by Ahmad et al. to predict WQI, utilizing 25 WQ input parameters [26]. Through a combination of backward elimination and forward selective combination approaches, *R*-squared values of 0.9270 and 0.9390 and MSE values of 0.1200 and 0.1158 were achieved. However, the use of 25 parameters might be impractical for a cost-effective real-time system, considering the expense of parameter sensors. It's worth noting that many studies either relied on manual lab analysis without estimating a standardized water quality index or utilized an excessive number of parameters, which could hinder efficiency. Moreover, the WQI of Malaysia is calculated using numerical methods and use 6 parameters. It is obvious that, the above mentioned 6 parameters might not be available in all areas. For example, the AN and SS parameters need to be sampled in lab after collecting for the study area which make the research expensive and time consuming. Therefore, based on the historical data, a machine learning method can predict the WQI of a river using less parameters and thus can reduce the cost of research. In this study, two ensembled and optimized machine learning models are being proposed to predict the WQI of Johor River. Developing an ensembled machine learning model involves combining multiple individual machine learning models to improve predictive performance. These models could be of different types or variations of the same type, trained on the same data or different subsets. Ensemble methods leverage the diversity of these models to reduce overfitting and improve generalization, often resulting in more robust and accurate predictions compared to single models. Common ensemble techniques include bagging, boosting, and stacking. There are various types of machine learning models that can be used in ensemble methods. Some common types include Decision Trees, Random Forest, Neural Networks, Logistic Regression and Gradient Boosting Machines. This research first analyzes all the parameters available in that river. Then feature importance method will identify the most important parameters. Finally, ensemble-based machine learning models are designed to predict the WQI using three parameters.

Materials and methods

Study area

In this research, the study area selected is the Johor River basin located in Johor, Peninsular Malaysia. The catchment area is around 2636 km^2 and the mainstream length is around 122.7 km as shown in Fig. 1.

The tributaries of the Johor River include the Seluyut River, Sengi River, Redan River, Temon River, and Tiram River. According to the hydrological data, the average flow rate of the Johor River is 37.5 m^3/s. Meanwhile, the annual mean rainfall intensity in this region is about 2360 mm, with mean temperature around 27 °C. In nature, river is categorized geographically into three zones: headwater (Upstream), transition (or middle stream) and depositional zone (or downstream) [27]. Based on the land-use activities and the pollution points, the study area is mapped into three streams as well shown in Figs. 1 and 2. In the Johor River, the upstream comprises the mountains with areas forest, very steep gradients with high ridges and deep valleys. The rivers originate in this zone within a channel network. The middle stream zone comprises the lower mountains and hills and has steep slopes but with mixed vegetation with alternate human activities such as oil palm plantation.

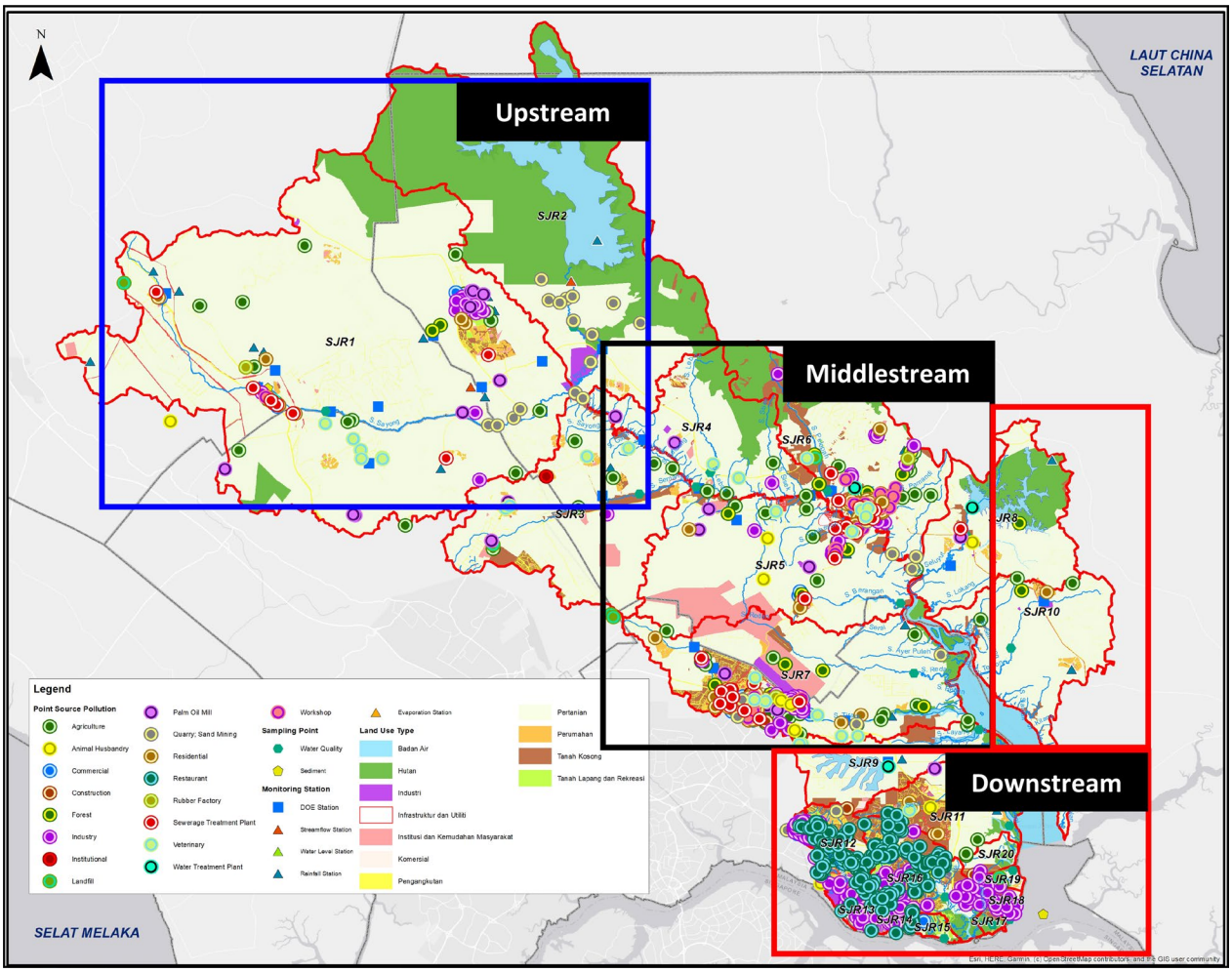

Fig. 1 Johor River Basin consists of all the land use and point source of pollution

The downstream zone begins when the river leaves the hills. The Johor River starts to meander due to low level of gradient. The upstream of the river systems is important for covering downstream ecosystems because they are closely linked. However, the terms 'upstream' and 'downstream' also relatively depend on the upstream and downstream relationships occur at different locations and scales, and the magnitude and nature of river, land use change and runoff generation [23]. The upstream area is located at the upper part of the Johor River dominated by of palm-oil plantation followed by forest, sand mining industrial area and residential. Meanwhile, the middle of the river covered by palm oil plantation, forest at the eastern part of the catchment and the mixed developed area (amenities, residential, commercial, and vacant land). Downstream area still dominated by the green area followed by vacant land and other mixed developed area (including high and low-density development area). Water quality for various parameters is decreasing from upstream to the downstream of the river stretch due to anthropogenic and land use activities in the catchment [28–30].

Therefore, this paper will investigate the relationship of water quality located in these three different segments with the effect of land use activities using temporal analysis at the beginning of analysis. Then the correlation of water quality parameters with the WQI and an ML-based feature importance analysis will be done to identify the most important parameters for WQI prediction. The water quality for Sungai Johor and its tributaries has been subjected to periodic monitoring and assessment by the Department of Environment (DOE), Malaysia. DOE currently provides a total of 43 stations in the catchment of Johor River. The water quality at these stations is monitored five to six times a year depending on the stations. The stations are usually situated downstream of known point pollution sources to monitor the river quality. Other stations are located to provide baseline or ambient

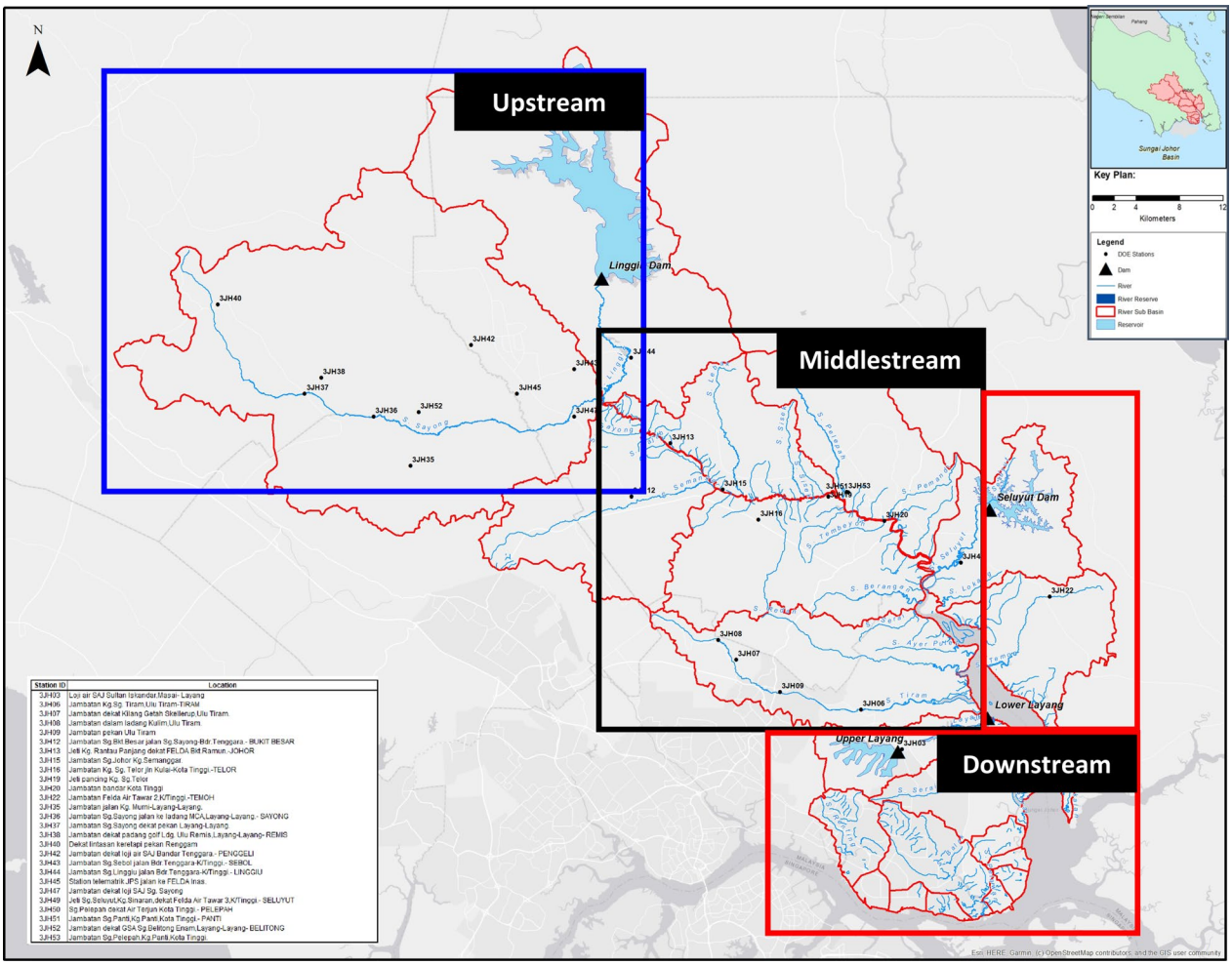

Fig. 2 The Department of Environment monitoring stations in Johor River basin

water quality data. The locations of DOE monitoring stations in Johor River basin are within the three segments as shown in Fig. 2.

The monitoring system comprises both automatic monitoring stations as well as the manual mode of monitoring. The automatic stations are located upstream of water intake points and provides water quality data with respect to pH, DO, temperature, conductivity (& salinity), turbidity and ammoniacal nitrogen. Threshold values are established that act as a trigger to activate appropriate actions when breached. In the manual component of the program, sampling activities and in situ measurements are undertaken at the designated sampling stations on a pre-determined monitoring frequency. Reports are submitted monthly while a summary yearly report is presented to the DOE within the first quarter of the following year to be incorporated into the DOE's Annual Environmental Quality Report. In addition, pollution events observed during sampling activities and at the automatic stations are also reported. For this study, 11 stations are categorized in the upstream segment, 22 stations in the midstream segment and the rest 11 stations in the downstream segment. The selected stations of this category are based on the geographic location on the river stretch.

Water quality index

The usual approach of the water quality indices is to process the water quality data into a single numerical value that able to represent overall status of water quality with a score ranging from 0 to 100. Typically, WQI consists of four general processes: (1) selecting desired water quality parameters for analysis, (2) reading of water quality data and convert the concentration of each parameter into a single-value dimensionless sub-index, (3) determining the weighting factor for each water quality parameter and (4) calculation of final single value of water quality index using the

calculated sub-indices with the weighting factors [4, 31, 32]. The Malaysian Water Quality Index (MWQI) that was developed by DOE Malaysia is used to assess and classify the surface water quality, then categorized them based on the beneficial uses of water locally. The framework for WQI was developed according to the four-common process of WQI models as mentioned above. The first process of MWQI model to determine the water quality and its classification is the parameter selection. There are six typical physicochemical water quality parameters used; BOD, COD, DO, ammoniacal nitrogen (NH_3-N), SS and pH value [3]. Second process is the calculation of sub-index value for each selected parameter where specific best fitted equations were developed to transform the measured water quality value to a non-dimensional sub-index value [4, 33, 34]. The existing WQI equation developed by Department of Environment (DOE) is as shown in Eq. 1.

$$\mathrm{WQI} = (0.22 * \mathrm{SlDO}) + (0.19 * \mathrm{SIBOD}) + (0.16 * \mathrm{SICOD}) + (0.15 * \mathrm{SIAN}) + (0.16 * \mathrm{SISS}) + (0.12 * \mathrm{SIpH}) \quad (1)$$

where SIDO = 0, for $x \leq 8$; SIDO = 100, for $x \geq 92$ and SID O = $-0.395 + 0.030x^2 - 0.00020x^3$, for $8 < x < 92$.

SIBOD = $100.4 - 4.23x$, for $x \leq 5$ and SIBOD = $108e^{-0.055x} - 0.1x$, for $x > 5$.

SICOD = $-1.33x + 99.1$; for $x \leq 20$ and SICOD = $103e^{-0.0157x} - 0.04x$, for $x > 20$.

SIAN = $100.5 - 105x$ for $x \leq 0.3$; SIAN = $94e - 0.573x - 5|x - 2|$, for $0.3 < x < 4$ and SIAN = 0, for $x \geq 4$.

SISS = $97.5e - 0.00676x + 0.05x$, for $x \leq 100$; SISS = $71e^{-0.0061x} - 0.015x$, for $100 < x < 1000$ and SISS = 0, for $x \geq 1000$.

SIPH = $17.2 - 17.2x + 5.02 \times 2$, for $x < 5.5$; SIPH = $-242 + 95.5x - 6.67 \times 2$, for $5.5 \leq x < 7$; SIPH = $-181 + 82.4x - 6.05 \times 2$, for $7 \leq x < 8.75$ and SIPH = $536 - 77.0x + 2.76x^2$, for $x \geq 8.75$.

The third process is to identify parameters weighting factor where each parameter is assigned to have different weight value as per expert panel opinions. According to the WQI Eq. (Eq. 1), the sum of weight values in six parameters is equal to 1. Among six parameters, the weighting factor for DO and BOD are the highest with value of 0.22 and 0.19, respectively. There is also same weight value (0.16) was used for COD and SS, while 0.15 was assigned for NH_3–N and the lowest value was determined for pH with value 0.12. The final process is for WQI evaluation where DOE has classified the index into 5 classes and categorized them to three categories to evaluate the surface water quality as shown in Eq. 2. Different class will serve in different purpose as stated in the National Water Quality Standards of Malaysia

$$\mathrm{Class} = f(\mathrm{WQI}) = \begin{cases} \mathrm{I}, & \mathrm{WQI} > 92.7 \\ \mathrm{II}, & 92.7 \geq \mathrm{WQI} > 76.5 \\ \mathrm{III}, & 76.5 \geq \mathrm{WQI} > 51.9 \\ \mathrm{IV}, & 51.9 \geq \mathrm{WQI} > 31.0 \\ \mathrm{V}, & \mathrm{WQI} \leq 31.0 \end{cases} \quad (2)$$

In this research, at first the numerical method is used to calculate the WQI and identify the classes for validating the prediction results. It is obvious that WQI calculation methods are quite lengthy and required more cost to get results from sampling, then only can be calculated for WQI. This research collected the DOE data from the year 2008 until 2018 and the data distribution based on three major streams and the class of water are shown in Fig. 3.

The research collected other parameters too and the analysis is shown in Fig. 4. Some parameters are distributed while others are too insignificant to show any pattern.

Machine learning models

Gradient Boosting Regression (GB)

Gradient Boosting Regression (GB) belongs to the category of supervised machine learning (ML) regression tree models using the concept of ensemble learning. As a regression model, the goal of GB is to build a function that guaranteed a robust link between an ensemble of inputs and one output variables. By developing an ensemble of weak learners at successive steps, a weighted strategy for all generated weak learners is adopted for providing a final strong model. The combination of weak learner can achieve better generalization and good predictive accuracies. From a mathematical point of view, the GBR can be formulated as follows [35, 36]:

$$\vartheta_n(x) = \vartheta_{n-1}(x) + \delta_n(x, \sigma_n) \quad (3)$$

where n corresponds to the number of iterations, σ_n the parameters of the regression tree model, $\delta_n(x, \sigma_n)$ is the regression tree function, and $\vartheta_n(x)$ is the regression model. The model parameters are estimated by reducing a loss function (Fig. 5).

Random Forest Regression (RF)

Random forest regression (RF) is a machine-learning model that uses several trees to train a regression model. The RF is a bagging method and it is defined as an

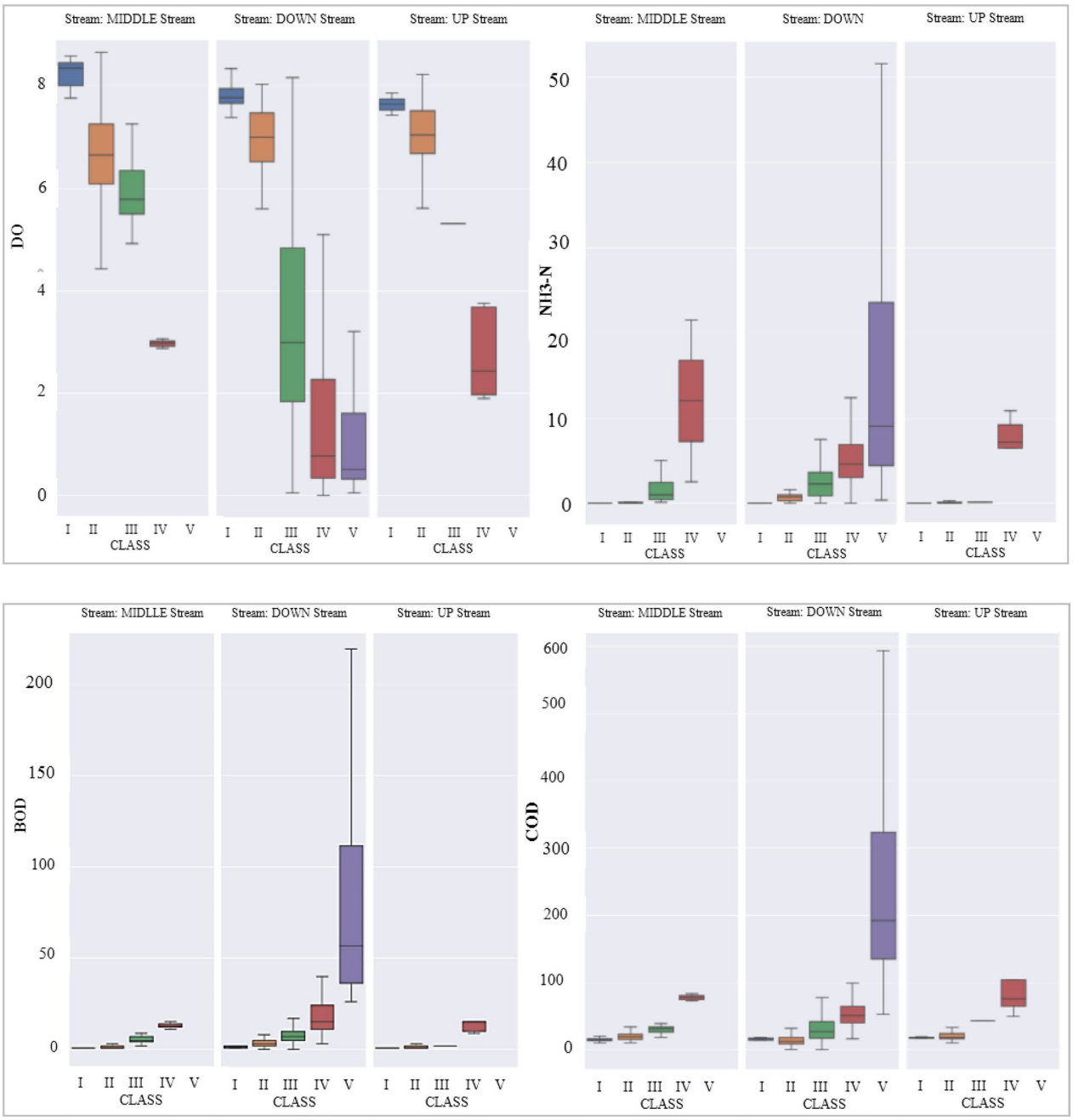

Fig. 3 Distribution of WQI parameters among the three streams and sub classes

ensemble of classification and regression tree (CART). The RF model can be developed by generating a large number of trees using a bootstrap strategy. Each single tree, that is a weak learner, is trained using a subset of predictors and the final model is provided by a voting strategy. During the training process, the out-of-bag is used during the fitting of the trees for performing a cross-validation strategy [37]. For developing a RF model, we need to determine three user-defined parameters, which are: (*i*) the number of variables used at each tree, (*ii*) the number of trees in the forest, and (*iii*) the minimum number of terminal nodes [37] (Fig. 6).

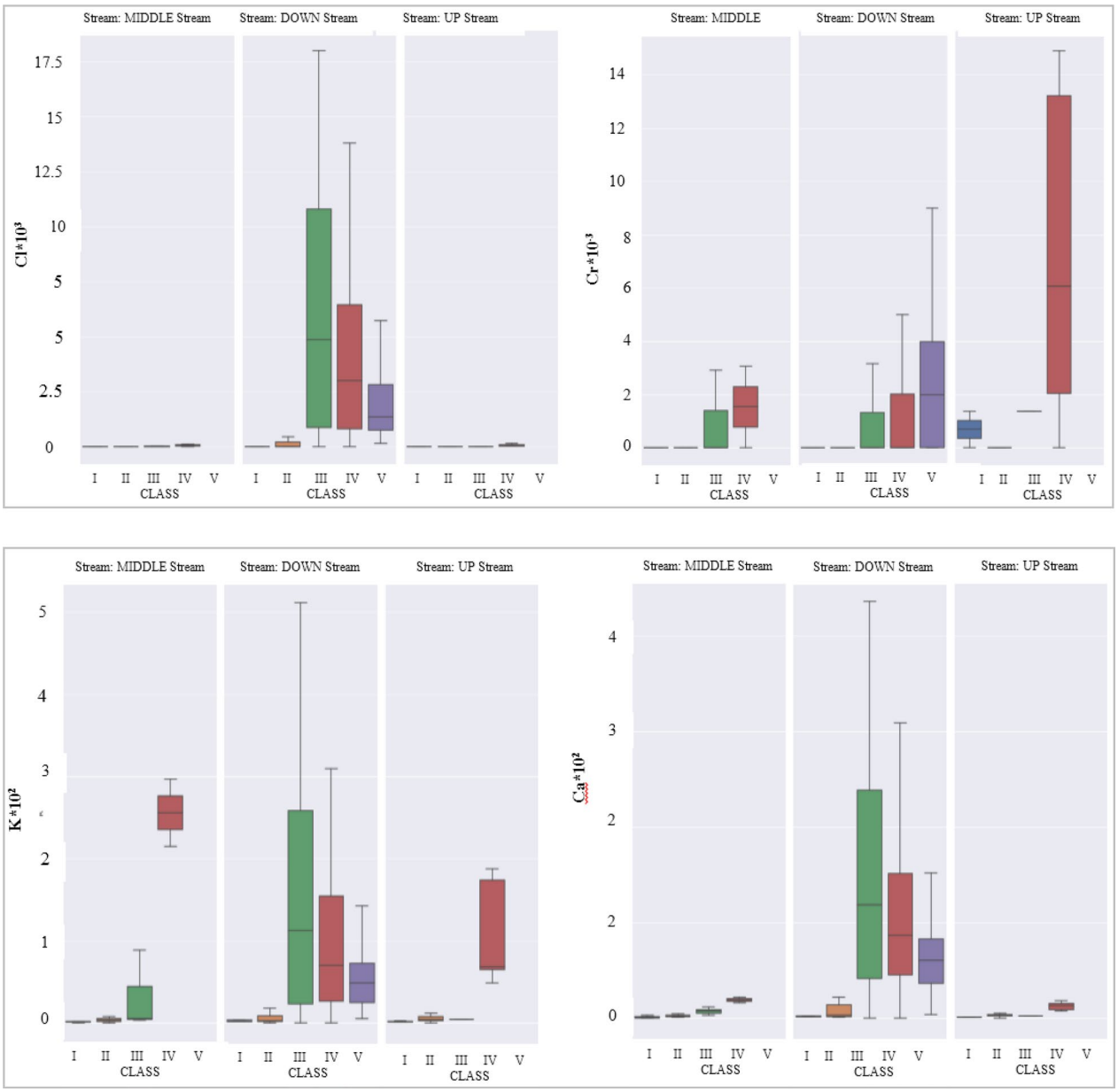

Fig. 4 Pollution materials (Cl, Cr, K and Ca) in Johor River basin, based on three streams and sub classes

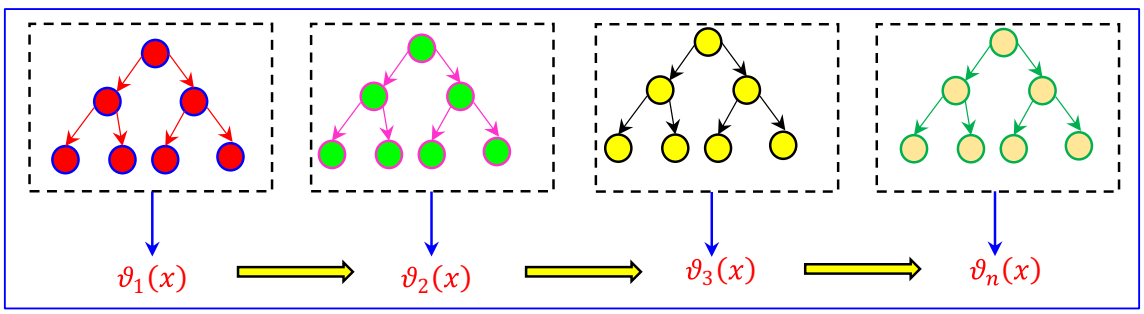

Fig. 5 Architecture of the Gradient Boosting Regression (GB)

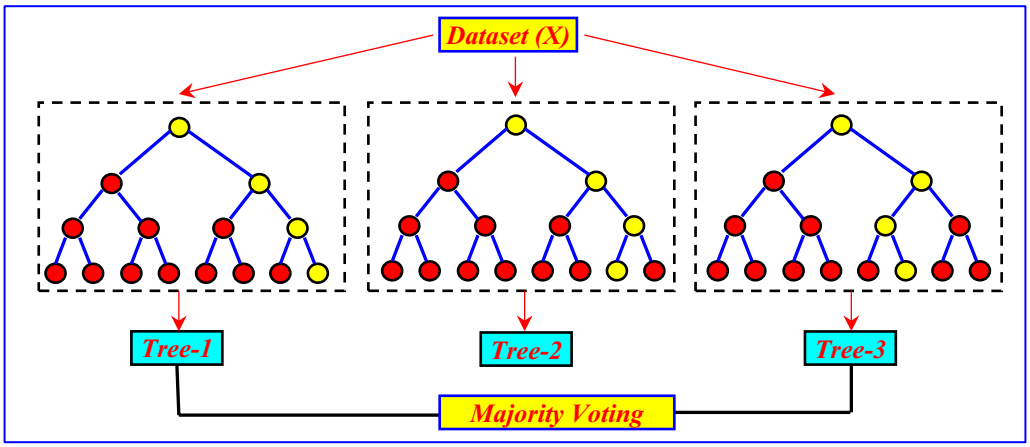

Fig. 6 Architecture of the random forest regression (RF)

Models development

The prediction of WQI will largely depend on the water quality parameters. In this research the correlation among the parameters is identified then ML-based feature importance method is applied. After identifying the most predictive variables, two ensembled ML approach was developed to predict the WQI. We have implemented GB and RF methods considering the scores of feature importance. These algorithms, particularly tree-based regression techniques, were selected due to their ability to handle data with diverse measurement scales. Moreover, they are robust against missing values and outliers while simplifying predictions for individual cases and intricate relationships between variables [38]. Given the dataset's varying dimensions, we opted for the RF and GB methods. To assess model accuracy, we performed a random split of the dataset for both model training and testing. Additionally, we fine-tuned the machine learning models' hyperparameters using the 'Randomized Search CV' method based on cross-validation (CV) scores. Four hyperparameters such as "n_estimators", "min_samples_split", "min_samples_leaf" and "max_depth" of RF and GB algorithms have been optimized as maximum depth to 10 to prevent overfitting of the data [39]. Our implementation of machine learning was carried out using Python's scikit-learn tools and the prediction accuracy is measured using coefficient of regression i.e., R^2 values. Also, the water class can be calculated from the WQI and the results of ML-based prediction and the observed data using Eq. 4.

$$\text{Accuracy} = \frac{\sum \left|(\text{Obs} - \text{Pred})\right|}{D} \times 100 \tag{4}$$

where D is the Total number of test data, Obs is the Observed water class and Pred is the Water class based on Predicted WQI.

Results and discussion

The WQI is calculated using all the parameters and the dataset is prepared to validate the ML-based prediction results. Figure 7 shows the WQI distribution among all the streams and sub classes. After completed the preliminary analysis, we recommend focusing on the examination of 13 selected parameters in a temporal context. For this analysis, we've identified seven water quality stations within the Johor River basin. These stations include one station upstream, two stations in the middle stream and four stations at downstream part of the basin. The distribution plot of those selected 13 parameters is shown in Fig. 8. The concentration of each parameter was compared with Class II permissible value which adopted from NWQS.

According to Fig. 8, the DO parameter at the upstream station falls below the Class II value, but it exhibits an increasing trend over time. The midstream part, DO is better than upstream and above than 5 mg/L followed the Class II while in the downstream part, DO is mostly lower and does not exceed 5 mg/L as it is possible since the area is populated by industrial activity and possible to contribute to the pollution. Parameters BOD and COD recorded at the upstream part, are exceeding Class II and they improve by time (correlate with DO values). Midstream part shows both parameters are better than upstream but they still exceed 3 mg/L for BOD and 25 mg/L for COD which a bit higher than Class II. High BOD value is mainly contributed by human activities at the surrounding areas. However, in the downstream

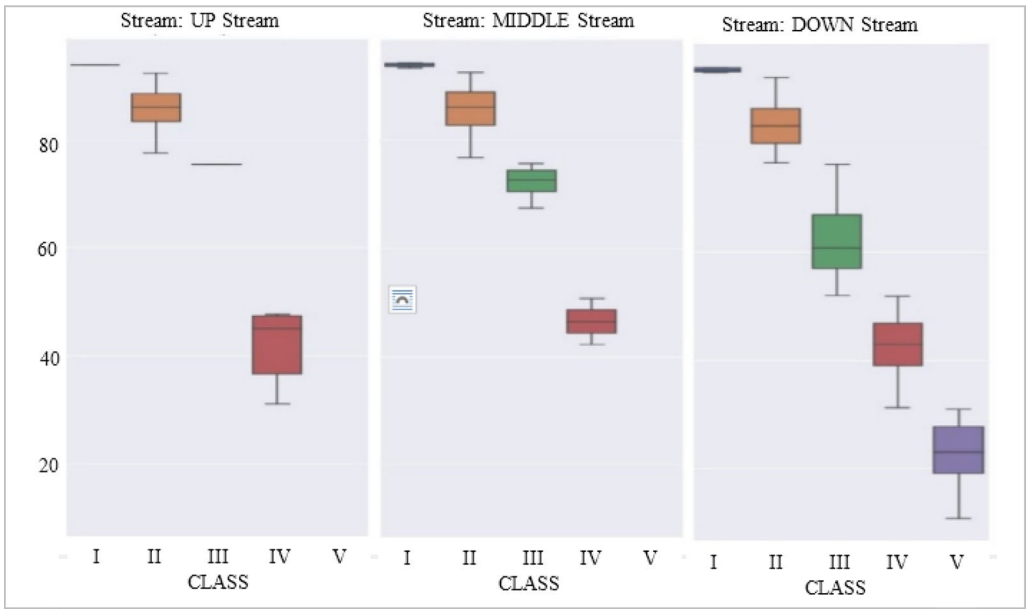

Fig. 7 The numerical WQI distribution among different class and streams

section, the levels of these parameters are generally higher and often exceed the Class II values for both BOD and COD with one exception being a station at one of the Johor River tributaries; Tiram River, that is quite low compared to other downstream station. This is also due to the low dense industrial activity at that area.

Suspended Solid (SS), Dissolved Solid (DS) and Total Solid (TS) are also crucial parameters to be considered in Johor River Basin due to their non-compliance to Class II. The temporal analysis reveals that the SS values consistently surpass the Class II limit of 50 mg/L, but there is an improvement in these values over time, particularly at the upstream station. This explanation finds roots in the distribution of mining industries in the area where more than five sand-mining industries were operating in the river basin during the study period. Other area at the upstream is experiencing current agricultural activity mainly in palm oil plantation. In the midstream section, SS values are somewhat better than those at the upstream part, although they still occasionally exceed Class II standards. The recorded levels of SS are being related to human activities and sedimentation at the surrounding areas. On top of that, downstream also recorded low value of SS below than Class II. Only on certain period, the value spike up exceeds Class II and suspected due to local seasonal changes. Meanwhile, at the downstream part, the value is high and more than Class II. Total solid can come from suspended solid and dissolved solid which can be contributed from any discharges either sewage treatment plants, industrial plant, or extensive crop irrigation.

The temporal analysis for pH shows similar trend for upstream and midstream part of Johor River Basin where its value is within the Class II. Same to downstream part in which mostly the value is within Class II value except for the fluctuated value at one of the Johor River tributaries i.e., Perembi River. The NH_3-N and PO_4 parameters present that the upstream value is above Class II and improve by times. These values exhibit a correlation with the dissolved oxygen (DO) readings and are influenced by changes in land use. They can result from various factors, including sedimentation due to site clearing or agricultural activities in the surrounding areas. NH_3-N and PO_4 are usually can be found in fertilizer, detergent, and pesticide. However, the midstream part recorded better reading than upstream even the reading spike up at certain time, but still exceed Class II (0.3 mg/L for NH_3–N). In contrast, in the downstream portion, both NH_3–N and PO_4 consistently exhibit high readings and frequently surpass the Class II standards.

Other parameters to be considered are E-Coli and Total Coliform. Both parameters recorded at the upstream and midstream part have values that exceeded the Class II. However, the value for the upstream part is improved by time. The downstream part shows certain station detected to have high E-Coli and Total Coliform as this part is monopolized by industrial activity. Both parameters are usually contributed by human and animal feces. In this case, Total Coliform is more crucial compared to E-Coli as it is not only can be found in feces but also from other sources. According to Environmental Protection Agency in United States, generally

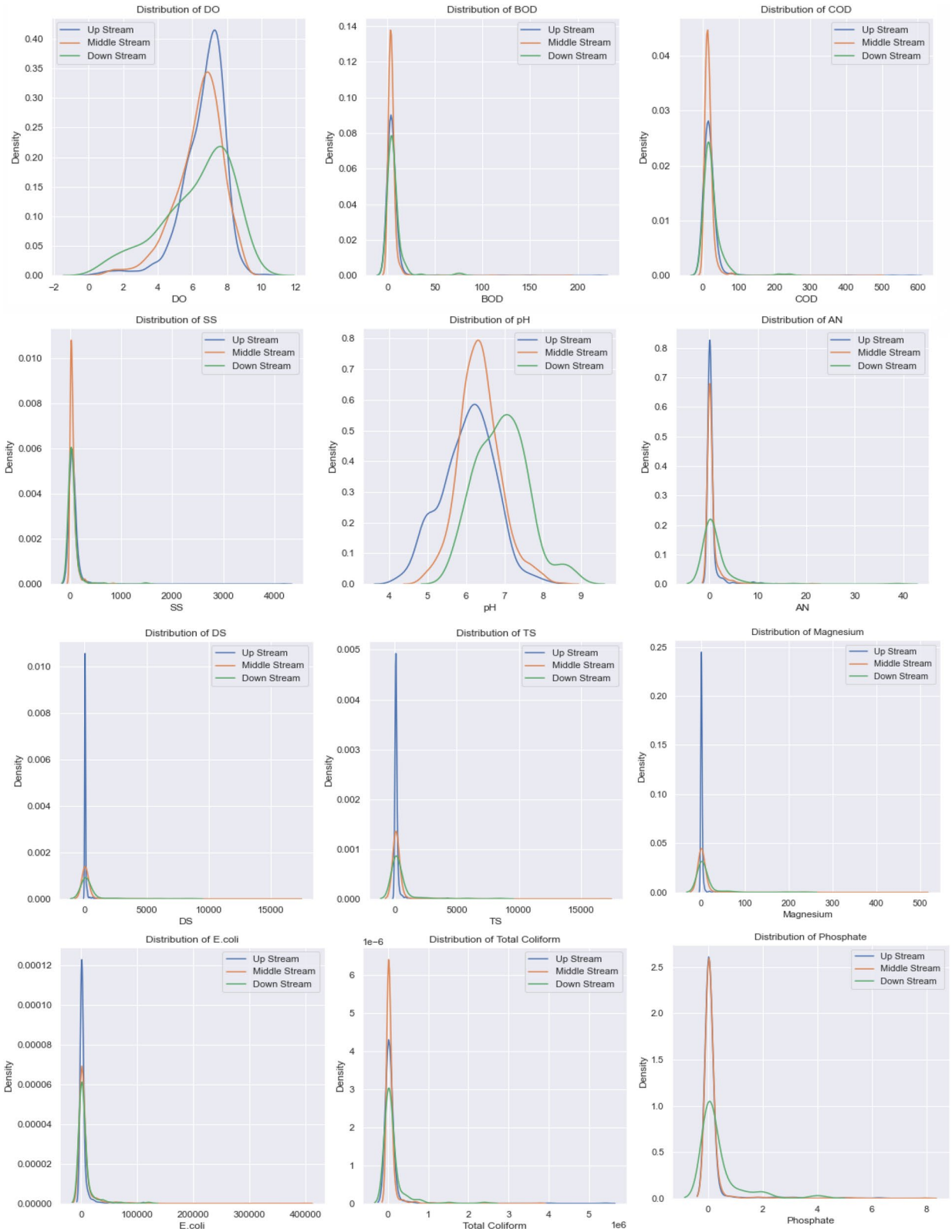

Fig. 8 Temporal analysis distribution

Coliform are bacteria that are not harmful and naturally present in environment where it also function as indicator to detect the presence of fecal bacteria like E-Coli [40]. Magnesium (Mg) and Iron (Fe) also being plot for the temporal analysis. The Mg recorded that, from the upstream to have high value of Mg downstream mostly and the highest is recorded at the downstream. Usually, Mg is either contained in sediment, location is near to seawater, ores and mining or limestone area. In terms of Fe, the upstream value presents high value in certain times and slightly above Class II value. It is possible that it is affected by plantations activities and certainly it comes from sediment. It is common that Fe is found in minerals and some industrial activities. However, in the midstream and downstream parts the value slightly exceeds Class II and it is also recorded that the value spikes up at the certain times. In this research, 28 parameters of water quality data are being collected and the Pierson correlation heatmap among those parameters are shown in Fig. 9. As mentioned in previous section explaining the WQI equation, there are only six parameters involved in the numerical calculations. However, the main objective of this article is to reduce those six parameters to 4 and predict the WQI. Those four parameters are chosen from the 28 parameters that are used as experimental.

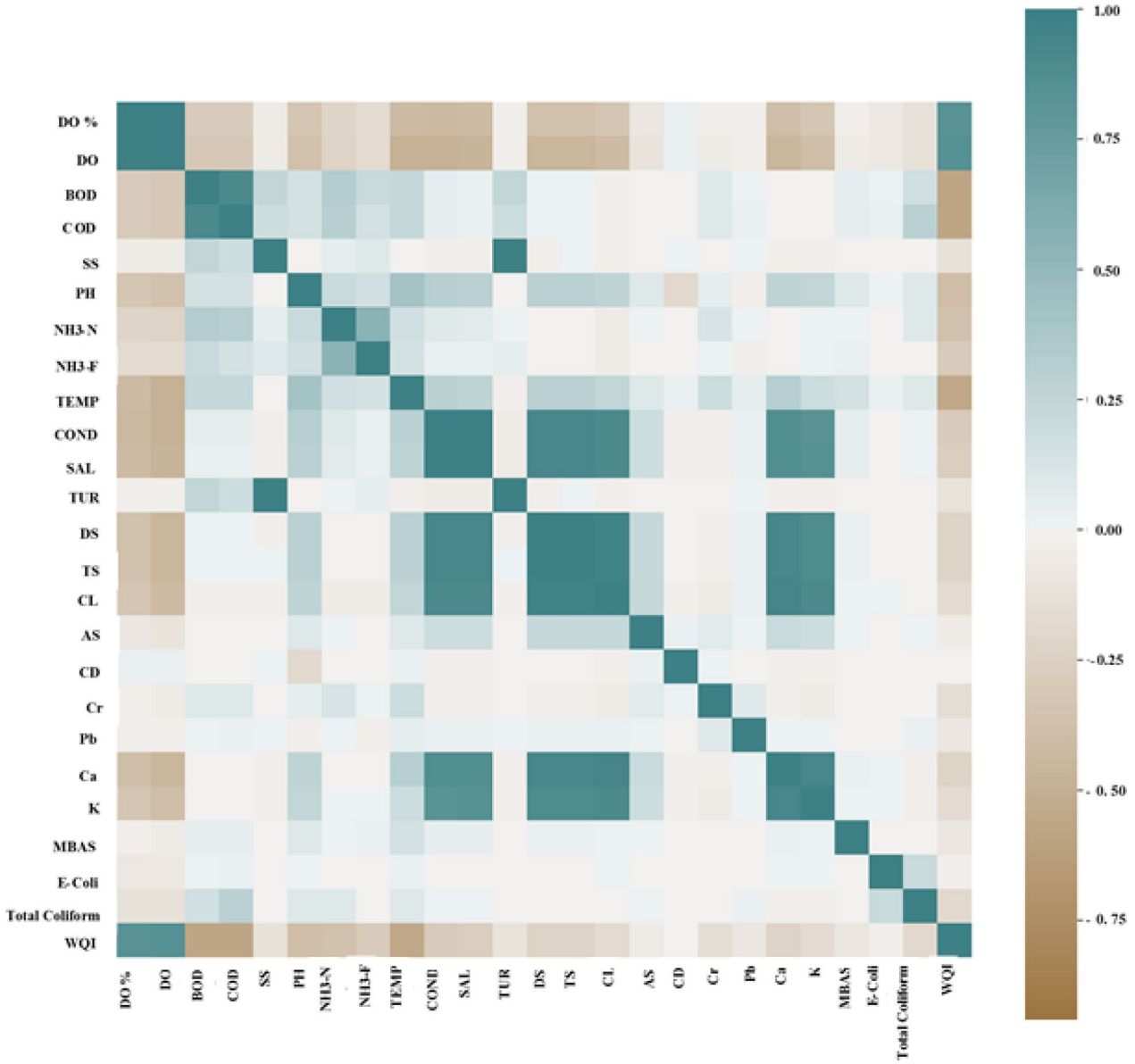

Fig. 9 Heatmap of Pierson correlation of Johor River basin WQ parameters

This research assessed the impact of water quality parameters using the Index of Pearson Correlation (PCI), which quantifies the strength of the relationship based on coefficient values (0.9–1; very high, 0.7–0.89; high, 0.5–0.69; moderate, 0.26–0.49; weak, 0–0.25; very weak). In the PCI matrix, negative values indicate an inverse relationship between parameters. When one parameter decreases, another tends to increase. For instance, in Fig. 7, we observe a moderate positive correlation among most parameters. Whereas a strong positive correlation value found among DO and DO% with WQI and negative correlation values were found for BOD, COD and AN with WQI. This negative correlation implies that if these parameters increase then the WQI could decrease of that state. Given the absence of significant correlations among most parameters, we turn to feature importance analysis for further investigation. Identifying important parameters from the correlation table can be challenging. Therefore, we incorporated feature importance analysis to enhance the efficiency and effectiveness of predictive models, particularly for generating the WQI in the Johor River basin, Malaysia. Figure 10 presents the feature importance graph of water quality parameters considered in our study.

This ML-based feature importance model assigns an importance score to each variable, with a higher score indicating greater importance [40]. Notably, DO/DO% emerges as the variable with the highest predictive power for WQI. Consequently, the ML model includes DO% as a major factor in predicting WQI within the study area. The chart presented in Fig. 8 explained one of the practices before applying machine learning model into any dataset called Feature Importance Score (FIS). This is the usual practices to determine how impactful of each independent variable to the dependent variable. The score is in the range between 0 and 1. The FIS was calculated using random forest algorithm where this model build up a decision trees. The feature importance is computed based on the following two principles called Gini Impurity and Average Decrease in Impurity. The Gini Impurity is a measure of how often a randomly chosen element would be incorrectly classified. It is used as a criterion to split the data in decision trees. For each tree in the Random Forest, the Gini Impurity is calculated for

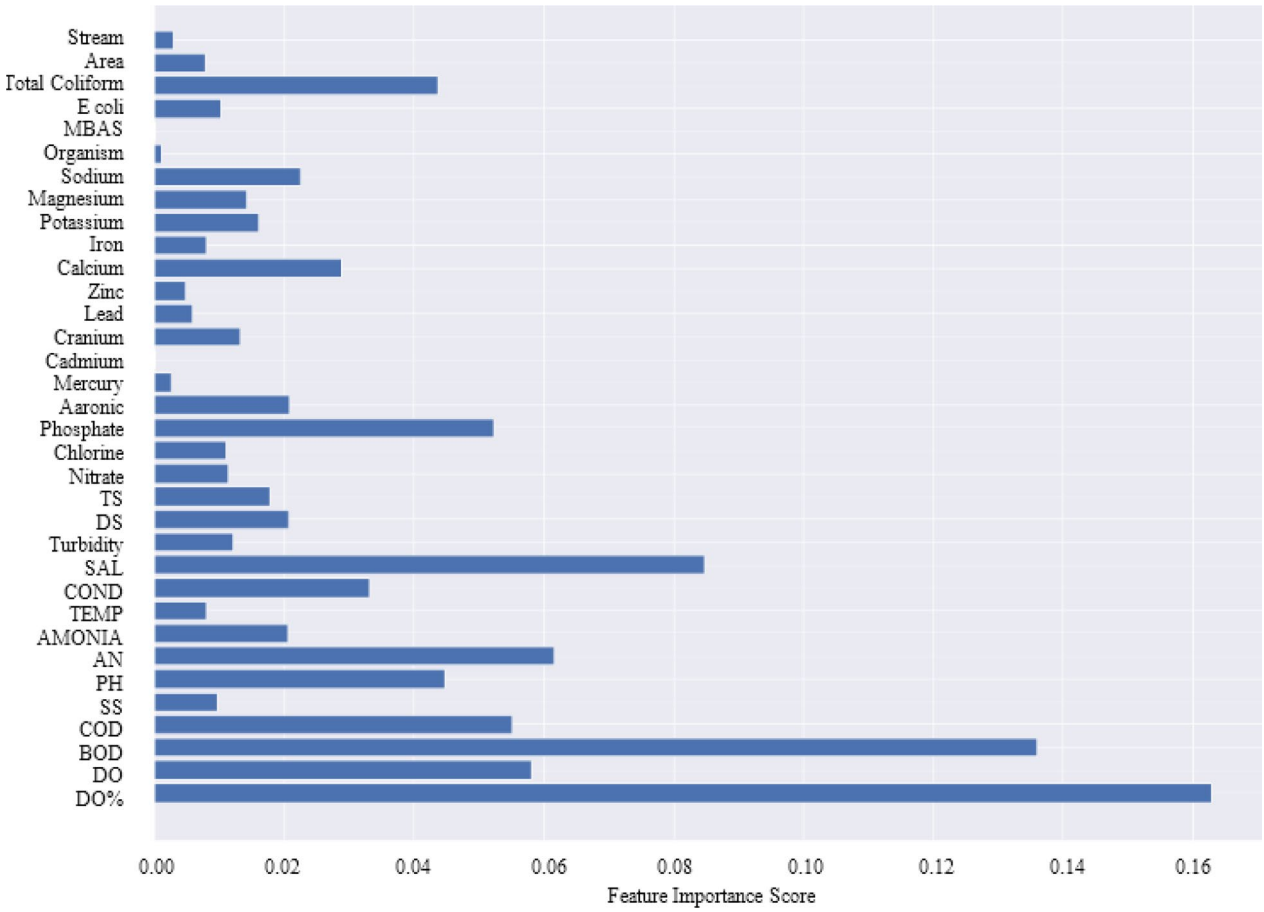

Fig. 10 Feature importance scores of Johor River basin WQ parameters

each split point based on each feature. The feature that leads to the greatest reduction in Gini Impurity is chosen as the splitting feature for that node. Followed by next principle which is Average Decrease in Impurity. Once the Random Forest is trained, the feature importance is calculated by aggregating the impurity decrease over all trees. For each feature, the total decrease in Gini Impurity across all trees is averaged. Features that result in a higher average decrease in impurity are considered more important because they contribute more to the overall performance of the Random Forest.

Furthermore, from Fig. 10 it is evident that, although AN has a moderate correlation value but in feature importance the predictive power became too low. Also, few parameters which are used in numerical WQI calculation such as SS and pH have low predictive power as the feature importance score is too low. It is found that, only the DO%, BOD and COD have moderate predictive power. Thus, the ML model will predict the WQI based on these three parameters only. After identifying the most important parameters the dataset is being split into training and validation for implementing the ML model. Two ensembled machine learning models, GB and RF are employed to predict the WQI for Johor River basin of Malaysia. Total 1637 data points are used in this research, among those data 328 data points i.e., 20% is used for validating the ML model and the rest used for training the ML model. The samples are divided randomly for training and validation. After train those 2 optimized ensembled ML models, the predicted WQI values are compared with the observed (numerically calculated) WQI in Johor River basin. The comparison graph is shown in Fig. 11. The blue lines indicated the observed values whereas the red lines are the GB-based predicted values and the green line is RF-based predicted values.

Figure 11 clearly shows that the ML model predictions are very close to the numerical WQI calculations. Although in numerical method six parameters are required but in ML-based prediction, we are using only three parameters. Thus, we can say that the ML-based prediction surely reduces the cost of analysis as well as the WQI calculation. Moreover, after analyzing the testing dataset the water class is generated and the summarized results are shown in Table 1. The R^2 values obtained in predicting the WQI are 0.86 and 0.85 for RF and GB models, respectively. According to the table we also observed that the ensembled results have more than 95% accuracy in predicting the water class, which is very significant based on our previous temporal analysis. In this research, we are not only predicting the WQI using ML model but also determining the most important parameter in the river using temporal analysis. The ML results sometimes depended on the study area and may show different accuracies in different zones. Therefore, temporal analysis is done to determine the continuity of ML analysis which identifies the most significant parameter in the river based on three different zones (upstream,

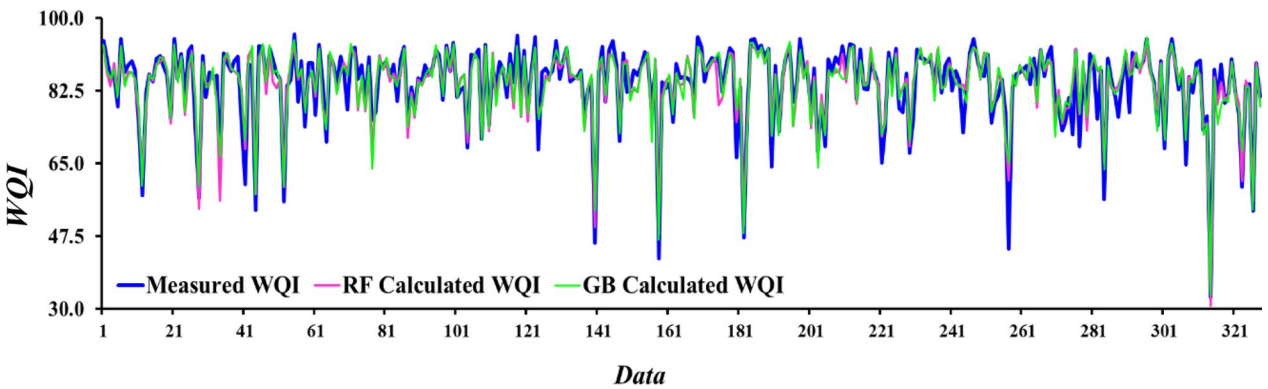

Fig. 11 Comparison between ML-based WQI prediction results

Table 1 Comparison of ML models

No of test data	No of true prediction		No of false prediction		Accuracy (%) in water class		R^2 value in WQI prediction	
	RF	GB	RF	GB	RF	GB	RF	GB
328	278	282	50	46	95.73	96.35	0.86	0.85

middle stream and downstream) that need to be tackle by manually observed.

Further comparison between the models using graphical representation is done using scatterplot as shown in Fig. 12. While the RF seems to be slightly more accurate compared to the GB, the two ML models were characterized by less scattered data and high fitting capabilities. Finally, based on the Boxplot and Violin plot reported in Fig. 13, it is clear that the two machine learning models were able to correctly predict the WQI with slightly superiority in favor to the RF model especially in correctly handling the outliers.

As previously discussed, generating the Water Quality Index (WQI) necessitates both in-situ and laboratory analyses, which yield results for each parameter as required by the WQI formula. This process entails significant costs, particularly when frequent monitoring and analysis are necessary to generate time series data. From an economic standpoint, this study proposes alternative methods aimed at reducing the number of parameters analyzed while still accurately obtaining the WQI using three predicted parameters derived from machine learning analysis. Over the long term, acquiring more data enables further analysis of water quality, facilitating the determination of WQI at specific locations.

Conclusions

Water, being an indispensable resource for survival, relies on the WQI to assess its quality. Traditionally, evaluating water quality has required researchers to engage in costly and intricate laboratory analyses. However, this research ventured into an alternative approach, harnessing machine learning to forecast water quality using just three readily available water

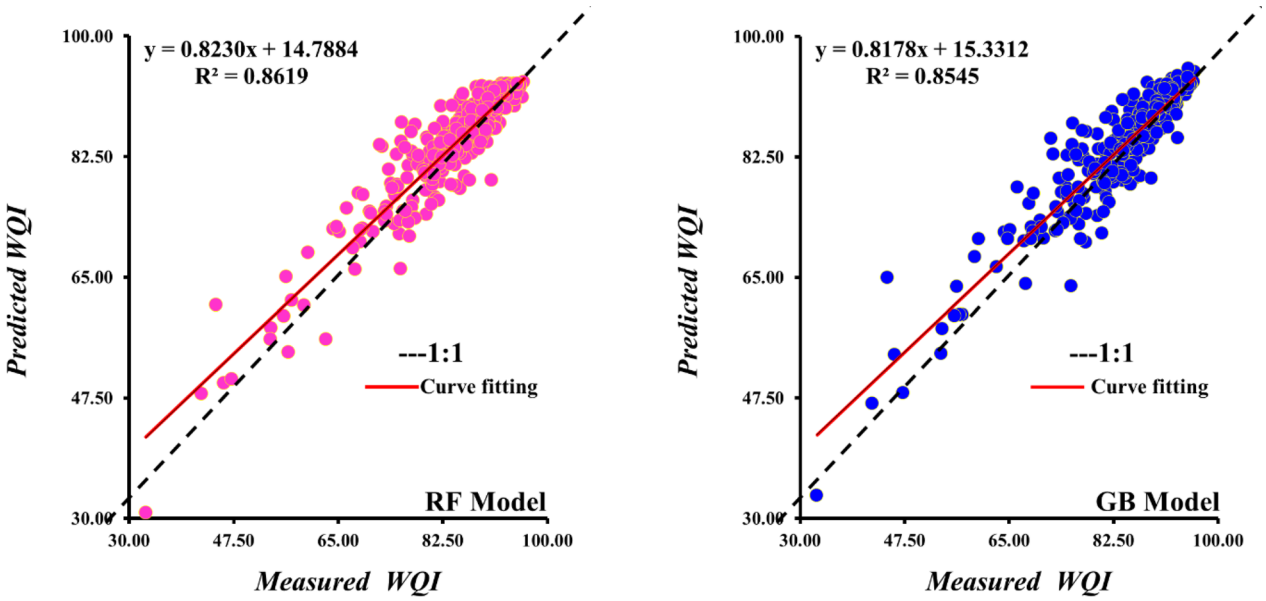

Fig. 12 Scatterplot of numerical WQI versus ML prediction

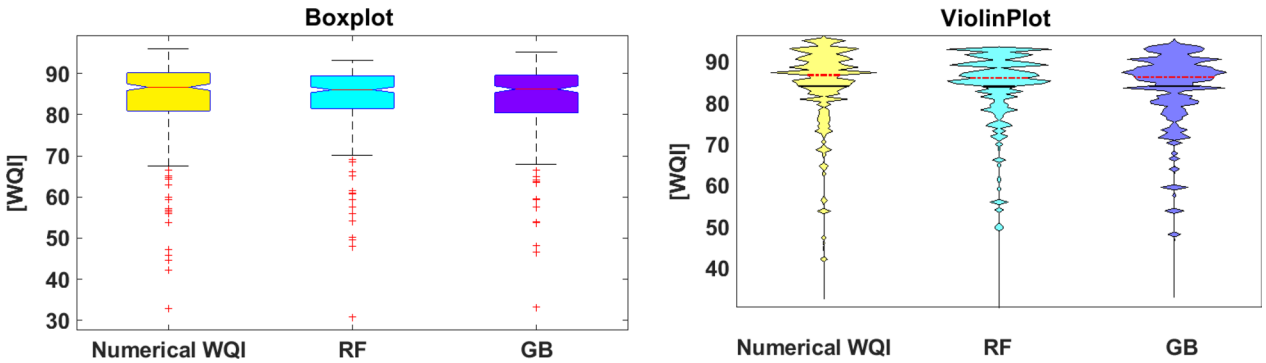

Fig. 13 Boxplot and violin plot comparison of numerical WQI and ML prediction

quality parameters. The dataset for this study was sourced from DID Malaysia and comprised 1637 samples collected from 44 distinct water quality stations located within the Johor River basin in Malaysia. This research first analyzes all the parameters data available for the river to acquire the overall picture of water quality condition in the river basin. From temporal analysis, Mg, E-Coli, SS, and DS are identified as the critical parameters in this river basin that possibly degrade the water quality. ML-based feature importance method was applied to identify the parameters with the most predictive powers. Finally, two ensembled ML approaches were developed to predict the WQI in the study area and achieved a R^2 of 0.86 was achieved for RF-based regression and 0.85 for GB-based ML technique in validation dataset (328 samples). The ensembled GB approach outperformed and achieved to identify water class with more than 96% accuracy as well. Therefore, this research proves that using only BOD, COD and DO% the WQI can be accurately predicted and almost 96 times out of 100 sample, the water class can be predicted using GB ensembled ML algorithm. In the future, researchers or decision-makers may choose to include this research as one of the methods to consider in their analyses. These findings could offer benefits in terms of economic reliability and time savings.

Acknowledgement
Special acknowledgement to Drainage and Irrigation Department and Department of Environment Malaysia for sharing the data.

Author contributions
All authors (LMS, HAM, MM, NSMN, SH, ME, OK, SSS) contributed to the manuscript. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. This research was supported by the Ministry of Higher Education (MoHE), Malaysia, through the Trans Disciplinary Research Grant Scheme, under project code of TRGS/1/2020/UNITEN/01/1/1.

Availability of data and materials
The data of this study are available from the authors upon request.

Code availability
Not applicable.

Declarations

Competing interests
The authors declare no conflict of interest.

Author details
[1]Institute of Energy and Infrastructure (IEI), Universiti Tenaga Nasional (UNITEN), 43000 Kajang, Selangor Darul Ehsan, Malaysia. [2]Civil Engineering and Geoinformatics Unit, TNB Research Sdn Bhd, 43000 Kajang, Selangor Darul Ehsan, Malaysia. [3]Centre for Dam Safety & Sustainability Intelligence, Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional (UNITEN), 43000 Kajang, Selangor Darul Ehsan, Malaysia. [4]Faculty of Science, Agronomy Department, Hydraulics Division University, 20 Août 1955, Route El Hadaik, BP 26, 21024 Skikda, Algeria. [5]Department of Water Engineering and Hydraulic Structures, Faculty of Civil Engineering, Semnan, Iran. [6]Department of Civil Engineering, Lübeck University of Applied Sciences, 23562 Lübeck, Germany. [7]Department of Civil Engineering, Ilia State University, 0162 Tbilisi, Georgia. [8]Department of Civil Engineering, College of Engineering, Diyala University, Diyala 32001, Iraq.

Received: 4 October 2023 Accepted: 24 March 2024
Published online: 01 April 2024

References

1. Awang H, Daud Z, Hatta MZM (2015) Hydrology properties and water quality assessment of the Sembrong Dam, Johor, Malaysia. Procedia Soc Behav Sci 195:2868–2873
2. Effendi H (2016) River water quality preliminary rapid assessment using pollution Index. Procedia Environ Sci 33:562–567
3. Uddin MG, Nash S, Olbert AI (2021) A review of water quality index models and their use for assessing surface water quality. Ecol Ind 122:107218
4. Pak HY, Chuah CJ, Tan ML, Yong EL, Snyder SA (2021) A framework for assessing the adequacy of Water Quality Index—quantifying parameter sensitivity and uncertainties in missing values distribution. Sci Total Environ 751:141982
5. Noh NM, Sidek LM, Haron SH, Puad AHM, Selamat Z, Razad AZA, Fai CM (2019) Analysis of urban water quality trends for effective reservoir sedimentation management in Cameron Highland. Int J Environ Technol Manage 22:276–290
6. Al-Mamun A, Zainuddin Z (2013) Sustainable river water quality Management in Malaysia. IIUM Eng J. https://doi.org/10.31436/iiumej.v14i1.266
7. 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
8. Rajaee T, Khani S, Ravansalar M (2020) Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: a review. Chemom Intell Lab Syst 200:103978
9. Gazzaz NM, Yusoff MK, Aris AZ, Juahir H, Ramli MF (2012) Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. Mar Pollut Bull 64:2409–2420
10. Azhar SC, Aris AZ, Yusoff MK, Ramli MF, Juahir H (2015) Classification of river water quality using multivariate analysis. Procedia Environ Sci 30:79–84
11. Sihag P, Kumar M, Sammen SS (2021) Predicting the infiltration characteristics for semi-arid regions using regression trees. Water Supply 21(6):2583–2595. https://doi.org/10.2166/ws.2021.047
12. Sihag P, Dursun OF, Sammen SS, Malik A, Chauhan A (2021) Prediction of aeration efficiency of Parshall and Modified Venturi flumes: application of soft computing versus regression models. Water Supply 21(8):4068–4085. https://doi.org/10.2166/ws.2021.161
13. Almohammed F, Sihag P, Sammen SS, Ostrowski KA, Singh K, Prasad CVSR, Zajdel P (2022) Assessment of soft computing techniques for the prediction of compressive strength of bacterial concrete. Materials 15:489. https://doi.org/10.3390/ma15020489
14. Pham QB, Sammen SS, Abba SI et al (2021) A new hybrid model based on relevance vector machine with flower pollination algorithm for phycocyanin pigment concentration estimation. Environ Sci Pollut Res 28:32564–32579. https://doi.org/10.1007/s11356-021-12792-2
15. Ehteram M, Sammen SS, Panahi F et al (2021) A hybrid novel SVM model for predicting CO2 emissions using multiobjective seagull optimization. Environ Sci Pollut Res 28:66171–66192. https://doi.org/10.1007/s11356-021-15223-4
16. Pham QB, Mohammadpour R, Linh NTT et al (2021) Application of soft computing to predict water quality in wetland. Environ Sci Pollut Res 28:185–200. https://doi.org/10.1007/s11356-020-10344-8
17. Abba SI, Hadi SJ, Sammen SS, Salih SQ, Abdulkadir RA, Pham QB, Yaseen ZM (2020) Evolutionary computational intelligence algorithm coupled with self-tuning predictive model for water quality index determination. J Hydrol 587:124974. https://doi.org/10.1016/j.jhydrol.2020.124974
18. Rahman LF, Marufuzzaman M, Alam L, Bari MA, Sumaila UR, Sidek LM (2021) Developing an ensembled machine learning prediction model for marine fish and aquaculture production. Sustainability 13:9124

19. Marufuzzaman M, Bin Ibne Reaz M, Rahman LF, Farayez A (2017) A location based sequence prediction algorithm for determining next activity in smart home. J Eng Sci Technol Rev 10:161–165
20. Ranković V, Radulović J, Radojević I, Ostojić A, Čomić L (2010) Neural network modeling of dissolved oxygen in the Gruža reservoir, Serbia. Ecol Model 221:1239–1244
21. Won Seo I, Yun SH, Choi SY (2016) Forecasting water quality parameters by ANN model using pre-processing technique at the downstream of Cheongpyeong Dam. Procedia Eng 154:1110–1115
22. Singh KP, Basant A, Malik A, Jain G (2009) Artificial neural network modeling of the river water quality-a case study. Ecol Model 220:888–895
23. Sakizadeh M (2016) Artificial intelligence for the prediction of water quality index in groundwater systems. Model Earth Syst Environ 2:8
24. Abyaneh HZ (2014) Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. J Environ Health Sci Eng 12:1–8
25. Ali M, Qamar AM. Data analysis, quality indexing and prediction of water quality for the management of Rawal watershed in Pakistan. In: Eighth international conference on digital information management (ICDIM 2013); 2013.
26. Ahmad Z, Rahim NA, Bahadori A, Zhang J (2017) Improving water quality index prediction in Perak River basin Malaysia through a combination of multiple neural networks. Int J River Basin Manag 15:79–87
27. Berhanu B, Seleshi Y, Amare M, Melesse AM (2016) Upstream-downstream linkages of hydrological processes in the Nile River basin. Landscape dynamics, soils and hydrological processes in varied climates. Springer, pp 207–223
28. Suratman S, Mohd Sailan MI, Hee YY, Bedurus EA, Latif MT (2015) A preliminary study of water quality index in Terengganu River basin, Malaysia. Sains Malays 44:67–73
29. Ismail WR, Ibrahim MN, Najib SA (2018) Longitudinal changes in suspended sediment loading and sediment budget of Merbok River Catchment, Kedah, Malaysia. Pertanika J Sci Technol 26:1899–1991
30. Zhao MM, Chen YP, Xue LG, Fan TT (2020) Three kinds of ammonia oxidizing microorganisms play an important role in ammonia nitrogen self-purification in the Yellow River. Chemosphere 243:125405
31. Gupta S, Gupta SK (2021) A critical review on water quality index tool: genesis, evolution and future directions. Eco Inform 63:101299
32. Sim SF, Tai SE (2018) Assessment of a physicochemical indexing method for evaluation of tropical river Water Quality. J Chem 2018:1–13
33. Mohiyaden HA, Sidek LM, Hayder G, Noh MN (2018) Water Quality Assessment Klang River water treatment plants. Int J Eng Technol 7:639–642
34. Noh NSM, Sidek LM, Haron SH, Puad AHM, Selamat Z (2018) Pollutant loading analysis of suspended solid, nitrogen and phosphorus at Bertam Catchment, Cameron Highlands using MUSIC. Int J Eng Technol 7:743–748
35. Li K, Chang F, Shi S, Jiang C, Bai Y, Dong H et al (2023) A new method of Ionic fragment contribution-gradient boosting regressor for predicting the infinite dilution activity coefficient of dichloromethane in ionic liquids. Fluid Phase Equilib 564:113622. https://doi.org/10.1016/j.fluid.2022.113622
36. Xu N, Wang Z, Dai Y, Li Q, Zhu W, Wang R, Finkelman RB (2023) Prediction of higher heating value of coal based on gradient boosting regression tree model. Int J Coal Geol. https://doi.org/10.1016/j.coal.2023.104293
37. Breiman L (2001) Random forests. Mach Learn 45:5–32
38. Tehrany MS, Pradhan B, Jebur MN (2013) Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. J Hydrol 504:69–79
39. Bramer M (2007) Avoiding overfitting of decision trees. Principles of data mining. Springer, pp 119–134
40. Serwecińska L, Kiedrzyńska E, Kiedrzyński M (2021) A catchment-scale assessment of the sanitary condition of treated wastewater and river water based on fecal indicators and carbapenem-resistant *Acinetobacter* spp. Sci Total Environ 750:142266

Publisher's Note