



ARTICLE

Water Quality Index Using Modified Random Forest Technique: Assessing Novel Input Features

Wen Yee Wong¹, Ayman Khallel Ibrahim Al-Ani¹, Khairunnisa Hasikin^{1,*},
Anis Salwa Mohd Khairuddin², Sarah Abdul Razak³, Hanee Farzana Hizaddin⁴,
Mohd Istajib Mokhtar⁵ and Muhammad Mokhzaini Azizan⁶

¹Department of Biomedical Engineering, Faculty of Engineering, Universiti Malaya, Kuala Lumpur, 50603, Malaysia

²Department of Electrical Engineering, Faculty of Engineering, Universiti Malaya, Kuala Lumpur, 50603, Malaysia

³Institute of Biological Sciences, Faculty of Science, Universiti Malaya, Kuala Lumpur, 50603, Malaysia

⁴Department of Chemical Engineering, Faculty of Engineering, Universiti Malaya, Kuala Lumpur, 50603, Malaysia

⁵Department of Science and Technology Studies, Faculty of Science, Universiti Malaya, Kuala Lumpur, 50603, Malaysia

⁶Department of Electrical and Electronic Engineering, Faculty of Engineering and Built Environment, Universiti Sains Islam Malaysia, Bandar Baru Nilai, Nilai, Negeri Sembilan, 71800, Malaysia

*Corresponding Author: Khairunnisa Hasikin. Email: khairunnisa@um.edu.my

Received: 11 September 2021 Accepted: 27 January 2022

ABSTRACT

Water quality analysis is essential to understand the ecological status of aquatic life. Conventional water quality index (WQI) assessment methods are limited to features such as water acidic or basicity (pH), dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), ammoniacal nitrogen (NH₃-N), and suspended solids (SS). These features are often insufficient to represent the water quality of a heavy metal-polluted river. Therefore, this paper aims to explore and analyze novel input features in order to formulate an improved WQI. In this work, prospective insights on the feasibility of alternative water quality input variables as new discriminant features are discussed. The new discriminant features are a step toward formulating adaptive water quality parameters according to the land use activities surrounding the river. The results and analysis obtained from this study have proven the possibility of predicting WQI using new input features. This work analyzes 17 new input features, namely conductivity (COND), salinity (SAL), turbidity (TUR), dissolved solids (DS), nitrate (NO₃), chloride (Cl), phosphate (PO₄), arsenic (As), chromium (Cr), zinc (Zn), calcium (Ca), iron (Fe), potassium (K), magnesium (Mg), sodium (Na), *E. coli*, and total coliform, in predicting WQI using machine learning techniques. Five regression algorithms—random forest (RF), AdaBoost, support vector regression (SVR), decision tree regression (DTR), and multilayer perception (MLP)—are applied for preliminary model selection. The results show that the RF algorithm exhibits better prediction performance, with R² of 0.974. Then, this work proposes a modified RF by incorporating the synthetic minority oversampling technique (SMOTE) into the conventional RF method. The proposed modified RF method is shown to achieve 77.68%, 74%, 69%, and 71% accuracy, precision, recall, and F1-score, respectively. In addition, the sensitivity analysis is included to highlight the importance of the turbidity variable in WQI prediction. The results of sensitivity analysis highlight the importance of certain water quality variables that are not present in the conventional WQI formulation.



KEYWORDS

Artificial intelligence; random forest; environmental modeling; alternative inputs; SMOTE

1 Introduction

Water security is a rising global predicament at present. The scarcity of clean water could create a calamity of waterborne diseases around the world. Water pollution could conflict with agricultural and industrial output, which can lead to environmental degradation [1]. Factors that influence water quality include natural occurrences from the climate as well as anthropogenic inputs coming from untreated municipal waste, mining activities, industrial effluent discharges, sediment runoff, and land use changes [2,3].

Developed and developing nations across the world are constantly enforcing laws and reforming water quality management to improve sanitation and water quality status. The water quality index (WQI) has been commonly used as a universal indicator to convert several quantitative and intensive parameters into a single qualitative variable. Due to its simplicity, WQI is commonly utilized to describe the physical, chemical, and biological properties of water bodies. This representation of WQI is effective for water quality assessment and resource management [4].

In Malaysia, 90% of the nation's water supply is derived from rivers and reservoirs. However, the *per capita* demand for water availability is growing at a rapid pace to support heavy manufacturing and large-scale crop cultivation. Population growth, improvement in living standards, and rapid urbanization have also imposed additional pressure on uncontaminated water resources [5]. According to statistics from the Malaysian Department of Environment (DOE) in 2017 [6], among the 477 rivers monitored in Malaysia, 47.0% were clean, 43.4% were semi-polluted, and 9.64% were polluted. In 1974, DOE adopted an 'Opinion Poll WQI (OP-WQI)' to rank the level of water quality. The panel of experts identified dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), pH, ammoniacal nitrogen (NH₃-N), and suspended solids (SS) as the utmost priorities [7].

In other countries, WQI models with various structures and weightings are developed by using sub-indexing and aggregation based on local opinions. The considered variables are inconsistent, since water pollutants such as heavy metals, pesticides, toxic compounds, and radioactive constituents are not included when evaluating water quality [8,9]. Most of the WQI models developed are region-specific, based on expert views and local guidelines, and many researchers have also highlighted the uncertainty of the WQI model [10–12].

Determination of water quality should be site-specific, with weightings decided according to water usage, taking into account such important parameters as phosphorus, nitrogen, trace metals, and fecal coliform. These variables are missing from the current Malaysian WQI assessment, making it less effective [13]. Malaysia's WQI index only considers the use of common water quality parameters, hence leading to eclipsing problems. Eclipsing is where the true nature of water quality is not reflected due to inappropriate sub-indexing and aggregation functions [9].

The number of WQI parameters adopted in Malaysia is not flexible for the user and is independent of the natural and anthropogenic factors. Thus, this study aims to investigate the potential of other water quality parameters as input variables. The primary motivation of this study is to explore the possibilities of predicting water quality under the absence of six primary water quality input

variables (DO, BOD, COD, NH₃-N, SS, pH). To the best of the author’s knowledge, there is limited technical literature or research that focuses on predicting the water quality index using other water quality parameters. Hence, this study serves to fill the research gap of understanding the abilities of different water quality parameters in the assessment of water quality using a machine learning approach. The remaining sections of this paper are arranged as follows: [Section 2](#) introduces the manual computation of the water quality index and reviews the research gap of existing modeling techniques using artificial intelligence. [Section 3](#) outlines the area of study as well as the properties of the dataset. [Section 4](#) presents the applied methodology for modeling and the proposed modeling schema. [Section 5](#) discusses the modeling results and analysis of the input parameters. [Section 6](#) ends the paper with a conclusion.

2 Literature Review

2.1 WQI Formulation

WQI is formulated from six monitoring parameters: dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), suspended solids (SS), ammoniacal nitrogen (AN), and water acidity or basicity (pH). The formula involves subindex (SI) computation given in [Table 1](#) before being fitted into the weighted formula in [Eq. \(1\)](#). Water quality is classified based on the Interim National Water Quality Standards for Malaysia (INWQS) shown in [Table 2](#) [14]. Class I is classified as unpolluted and safe for drinking, Class II is fit for recreational and aquatic use yet requires conventional treatment, Class III is considered as polluted water supply and hence requires extensive treatment, Class IV is only for irrigation or domestic use, whereas Class V cannot be utilized for any purposes [15].

$$WQI = 0.22SIDO + 0.19SIBOD + 0.16SICOD + 0.16SISS + 0.15SIAN + 0.12SIpH \tag{1}$$

Table 1: The estimation of subindex values [14]

Parameter	Estimation value	Subindex equation
DO (in % saturation)	$x \leq 8$	$SIDO = 0$
	$8 < x < 92$	$SIDO = -0.395 + 0.030x^2 - 0.00020x^3$
	$x \geq 92$	$SIDO = 100$
BOD	$x \leq 5$	$SIBOD = 100.4 - 4.23x$
	$x > 5$	$SIBOD = 108e^{-0.055x} - 0.1x$
COD	$x \leq 20$	$SICOD = -1.33x + 99.1$
	$x > 20$	$SICOD = 103e^{-0.0157x} - 0.04x$
NH ₃ -N	$x \leq 0.03$	$SIAN = 100.5 - 105x$
	$0.3 < x < 4$	$SIAN = 94e^{-0.573x} - 5 x - 2 $
	$x \geq 4$	$SIAN = 0$
SS	$x \leq 100$	$SISS = 97.5e^{-0.00676x} + 0.05x$
	$100 < x < 1000$	$SISS = 71e^{-0.0061x} + 0.015x$
	$x \geq 1000$	$SISS = 0$
pH	$x < 5.5$	$SIpH = 17.2 - 17.2x + 5.02x^2$
	$5.5 \leq x < 7$	$SIpH = -242 + 95.5x - 6.67x^2$
	$7 \leq x < 8.75$	$SIpH = -181 + 82.4x - 6.05x^2$
	$x \geq 8.75$	$SIpH = 536 - 77.0x + 2.76x^2$

Table 2: DOE water quality index classification [14]

Parameter	Unit	Class				
		I	II	III	IV	V
NH ₃ -N	mg/l	<0.1	0.1–0.3	0.3–0.9	0.9–2.7	>2.7
BOD	mg/l	<1	1–3	3–6	6–12	>12
COD	mg/l	<10	10–25	25–50	50–100	>100
DO	mg/l	>7	5–7	3–5	1–3	<1
pH	–	>7	6–7	5–6	<5	>5
SS	mg/l	<25	25–50	50–150	150–300	>300
WQI	–	<92.7	76.5–92.7	51.9–76.5	31.0–51.9	<31.0

Manual water quality calculation is lengthy, as it requires the process of converting raw data into subindex. Therefore, alternative techniques have been explored recently to provide a direct and simplified approach compared to the time-consuming computation process.

2.2 Machine Learning in WQI

River engineering to predict the ecological variables of water quality using artificial intelligence (AI) and machine learning (ML) is a widely recognized method to simplify the computation process of water quality assessment. Water quality parameters contain complex interactions and possess several nonlinear variables. AI techniques are adopted to understand the most suitable and efficient approach for modeling nonlinear inputs, evolving as a black-box model which brings environmental insights for strategic and planning purposes [16]. Various research works and technical literature are published based on the application of ML models such as neural networks, adaptive neuro-fuzzy inference systems, regression methods, ensemble techniques, and other hybrid models in water resource management [17–20]. These publications strive to understand the correlation of each water quality parameter, investigate the minimal number of inputs required for water quality index prediction, and spatially assess the point source and distribution of water pollution.

The most prominent tool for water quality modeling is artificial neural networks (ANNs), as the outcomes delivered from ANNs are consistently high in accuracy. ANNs are remarkable at forecasting and identifying problem variables from a set of complex or inaccurate data. When it comes to estimating WQI using ANN, the use of backpropagation neural networks (BPNN) and radial basis function neural networks (RBFNN) is equally effective for simplifying the computation process of WQI in the Langat and Klang Rivers, Malaysia [21]. The parameters used for the work in [21] are based on DO, BOD, COD, NH₃-N, SS, and pH, with the best coefficient of determination (R^2) evaluated using RBFNN at 0.9807 when BOD is excluded from the input. The authors ranked DO as having the highest correlation with WQI. Othman et al. [15] constructed an ANN model of 98.78% accuracy as the basis of water quality assessment. The publication benchmarked that the best performance was attained when only five input parameters were considered, eliminating pH as an input variable. Nevertheless, ANNs are prone to overfitting and facing local minima problems [22], though these shortcomings could be overcome with optimization and generalization techniques.

Tiwari et al. [23] explored the capability of an adaptive neuro-fuzzy inference system (ANFIS) by predicting WQI based on two clustering techniques, fuzzy C-means (FCM) and subtractive clustering (SC). The model's performance was significantly close to the experimental value, and it suggested ammonia, chlorides, and fecal coliform as the most sensitive parameters for water quality assessment in India. Multistep modeling performed by Elkiran et al. [24] compared several AI models, i.e., BPNN, ANFIS, support vector machine (SVM), and a linear autoregressive integrated moving average (ARIMA) model, in DO prediction. The authors explained that ANFIS might have better verification in two of the three stations evaluated, while SVM has an edge in prediction performance in one of the stations. Abobakr Yahya et al. [25] developed an SVM approach for ungauged catchment water quality prediction; the presented model proved to be efficient, with high precision, yet sensitive to the error level.

2.3 Research Gap in WQI Modeling

In summary, based on the above literature, it is evident that water quality prediction models based on machine learning are reliable and capable. It is also apparent that the determinant features of water quality vary according to the area of study, since different rivers have different land use purposes, and hence, certain water quality parameters may seem significant for a specific river but not be universally applicable. Although the application of AI models to water quality assessment has significantly increased, many models merely consider input parameters from the water quality index formula. Meanwhile, other water quality parameters and their role in water quality assessment have yet to be explored. There are approximately 30 water quality parameters, which include biological, chemical, and physical parameters, as well as trace metals of water bodies that interact with each other and affect water quality. However, there is a lack of critical parameters in most WQI models [9]; for example, the Malaysian WQI only considers common WQ inputs, neglecting biological indicators.

Moreover, the influence of land use, human activities, and socioeconomic behavior have different effects on the variability of water quality. The hydrological process of rivers is dynamic and complex; hence, it is mandatory to understand other significantly influential variables that have nonlinear processes in water quality assessment. This information is essential as a way toward formulating water quality in terms of adaptive parameters that suit the land use activities around the river.

The existing development of water quality modeling had the following research gaps: (1) Existing publications are focused on comparative analysis of different AI techniques and approaches, aiming at simplifying the WQI calculation process; the current WQI was limited to just six water quality parameters, so other water quality parameters were not taken into consideration. Models should be updated when new data or evidence is found about what is polluting the water source. (2) Sensitivity analysis from research articles [15,21,26] has continuously concluded that among the six water quality input variables, certain variables can be omitted, thus suggesting the possibility to investigate more water quality aspects for water quality assessment and management. (3) Change in anthropogenic activities in the 21st century has had an impact on water quality exposure, hence the rising need to redefine important water quality variables as input parameters.

3 Materials and Methods

3.1 Study Area

The two rivers considered in this research study are the Klang River and the Langat River (Fig. 1); most of the rivers in both river basins are in the semi-polluted and polluted categories within the Federal Territory of Kuala Lumpur and state of Selangor, Malaysia. The two main rivers supply

hundreds of thousands of people in the Klang Valley. Therefore, consumers were highly affected when wastewater treatment plants were forced to shut down due to water pollution. Hence, there is an arising need to monitor the health of both rivers, as they are important water sources for domestic use.

The Klang River is approximately 120 km in length with a basin area of 1288 km², located at 2°55'N–3°25'N latitude and 101°15'E–101°55'E longitude. The upstream begins in the mountains of Peninsular Malaysia at 1200 m altitude, comprising the Gombak and Hulu Langat districts, running through the heart of Kuala Lumpur, then crosses part of Selangor state before stemming with Port Klang to the Straits of Malacca. Most of the river basin is surrounded by urban infrastructures and industrial, recreational, and residential utilities. The river lies in a tropical climate area with heavy rainfall during the Southwest and Northeast Monsoons, average monthly humidity of 80%–85%, and uniform temperature between 27°C and 31°C. The area receives daily sunshine for 4.5 to 7.0 h per day, with an evaporation rate ranging between 3.0 and 5.0 mm daily [26,27]. The water quality of the Klang River basin is deteriorating due to the combination of rapid urbanization and high occurrences of pollution. Based on Sharif et al. [28], among the sources of pollution are residential and industrial runoff from construction sites and sewage discharge from treatment plants. It is estimated that 77,000 metric tons of garbage are dumped into the Klang River every year.

Meanwhile, the Langat River stretches across a total catchment area of 1815 km² across Kajang, Cheras, Bangi, and Putrajaya, situated between latitudes 2°40'N 152"N and 3°16'N 15"N and longitudes 101°19'N 20"E and 102°1'N 10"E. The main river is 141 km long, located 40 km to the east of Kuala Lumpur, featuring large tributaries such as the Semenyih River, the Lui River, and the Beranang River. The river flows from a highland altitude of 1500 m at the Pahang–Selangor border, draining towards the Straits of Malacca. The Langat and Semenyih Reservoirs are built for domestic and industrial water supplies and to generate power for Langat Valley. The Langat River also supports numerous activities such as effluent discharge, irrigation, fishing, and sand mining along its streams. The catchment also serves as an important water source facility to 1.2 million people [25,29].

3.2 Data Collection and Monitoring Sites

The dataset used for this study was obtained from the Department of Environment (DOE), Malaysia at 94 monitoring stations (shown in Fig. 1) along the Klang and Langat Rivers during the period between 2014 and 2019. The water quality parameters that are included in this study are DO, BOD, COD, NH₃-N, SS, pH, conductivity (COND), salinity (SAL), turbidity (TUR), dissolved solids (DS), nitrate (NO₃), chloride (Cl), phosphate (PO₄), Arsenic (As), Chromium (Cr), Zinc (Zn), Calcium (Ca), Iron (Fe), Potassium (K), Magnesium (Mg), Sodium (Na), *E. coli*, and total coliform. The type of pollution for both the Klang and Langat Rivers is similar, as both areas are located in densely populated industrial and residential districts; given the average similarity of the data analyzed, both datasets were merged, and the statistical properties of each river are compiled in Table 3 below.

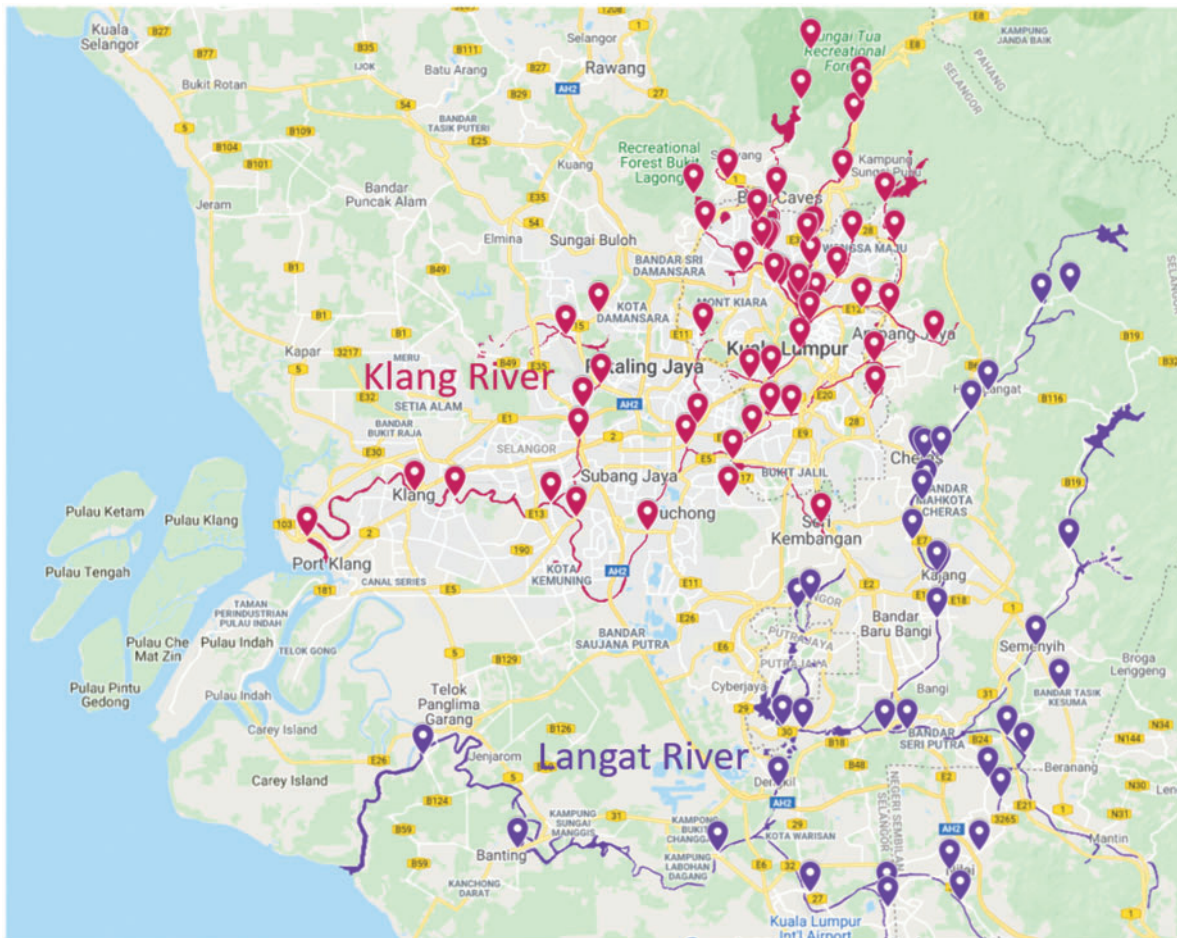


Figure 1: Klang river and Langkat river at Selangor, Malaysia

Table 3: Data set statistical properties

Parameters	Min	Max	Mean	SD
pH	4.15	9.83	7.34	0.38
DO (mg/l)	0.01	9.13	5.49	1.75
BOD (mg/l)	1.00	89.00	10.21	7.69
COD (mg/l)	4.50	263.40	31.48	23.06
SS (mg/l)	0.00	4400.0	46.93	165.18
NH ₃ -N (mg/l)	0.009	30.90	3.94	3.69
COND (μS/cm)	3.50	38577.82	373.82	1963.73
SAL (ppt g/l)	0.00	24.43	0.20	1.2
TUR (NTU)	0.00	3529.2	43.19	111.68
DS (mg/l)	1.00	24800.0	249.33	1341.23
NO ₃ (mg/l)	0.01	280.0	1.24	6.39
Cl (mg/l)	0.948	15500.0	68.94	682.64
PO ₄ (mg/l)	0.01	9.13	0.3811	0.5004
As (mg/l)	0.0009	0.604	0.0176	0.0295

(Continued)

Table 3 (continued)

Parameters	Min	Max	Mean	SD
Cr (mg/l)	0.0009	0.241	0.0012	0.0035
Zn (mg/l)	0.001	3.78	0.0292	0.0855
Ca (mg/l)	0.1	298	15.14	16.92
Fe (mg/l)	0.0097	22.6	0.2715	0.5874
K (mg/l)	0.10	500	6.119	16.96
Mg (mg/l)	0.10	2600	5.067	53.72
Na (mg/l)	0.10	8280	43.15	363.25
<i>E. coli</i> (cfu/100 ml)	0.00	6.6e9	28.4e6	205.9e6
Total coliform (cfu/100 ml)	0.00	1500e9	4.95e9	30.9e9

The distribution of DO in the Klang River ranged between 0.50 mg/l and 11.51 mg/l, with an average of 5.4 mg/l; a significant fraction of DO values fall in Class III. As compared to the Langat River, which has an average DO of 6.0 mg/l, the Klang River is much more deprived of dissolved oxygen to sustain aquatic life. Both rivers have a mean DO value of 5.49. Pollution in the Klang River is much more evident, as BOD and COD levels are in Class IV (10.59 mg/l) and Class III (32.47 mg/l). The Klang River is at serious ecological risk, as the mean value of NH₃-N is categorized in Class V at 4.28 mg/l.

The pH value, on the other hand, is in Class I, similar for both the Klang and Langat Rivers at an average of 7.34 and 7.31, respectively. The concentration of SS in the Klang River is under control in Class II, with a mean of 35.05 mg/l, whereas the SS concentration in the Langat River demands more attention, as it is in Class III (115.22 mg/l). The BOD and COD concentration of the Langat River are in the same class as the Klang River; ammonia levels, averaging 2.0 mg/l, fall into Class IV.

Throughout the monitoring period, the Klang River was classified as Class III 56.7% of the time, and 62.2% of the Langat River's data are categorized as polluted in Class III according to the WQI index. The distribution of WQI for both rivers is illustrated in Fig. 2. The WQI classification of the entire dataset has the highest percentage in Class III (57.42%). Overall, only a fraction (28.75%) of WQI data are in Class I (2.72%) and Class II (26.02%). There are a total of 759 instances where the water was unfit for any purpose.

The variability of samples is largely due to the geographical variation of pollution sources dispersed into the river basin. Both rivers flow through urban areas and townships where wastewater or sewage effluent is poured into the river channels. For example, significant variation can be observed in turbidity, as the maximum value is 3529 NTU and the lowest is 0.00 NTU. The large variation could be the impact of industrial effluents and urban runoff mixing from water treatment processes.

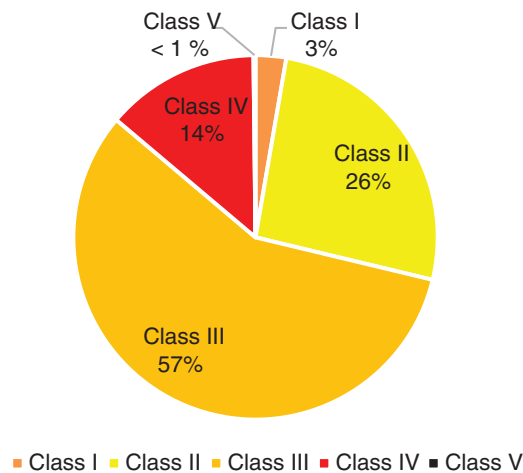


Figure 2: Class distribution of water quality in Klang and Langat River for $n = 5,483$ samples

4 Methodology

The major goal of this study is to propose a prediction model that predicts WQI using a new set of features instead of the conventional features that are usually applied in the WQI formula. Therefore, the six common WQ variables (DO, BOD, COD, $\text{NH}_3\text{-N}$, SS, and pH) were not included in the modeling process. Instead, new water quality features are applied in this work, including COND, SAL, TUR, DS, NO_3 , Cl, PO_4 , As, Cr, Zn, Ca, Fe, K, Mg, Na, *E. coli*, and total coliform. The methodology of this study is illustrated in Fig. 3 below, which involves data preprocessing of the dataset, followed by preliminary model selection using five machine learning regression algorithms. This step acts as the data analysis stage to select a suitable model capable of predicting WQI with the 17 input features. Next, the best-fitting algorithm is chosen for classification training. In the training phase, two sets of training data are applied; the first set is the unmodified training data, which contain imbalanced classes, while the next set of training data is modified using the data augmentation technique. The synthetic minority oversampling technique (SMOTE) is an oversampling approach to recreate synthetic examples of minority classes [30]. The results are then analyzed based on the accuracy, precision, recall, and confusion matrix. Lastly, sensitivity analysis was carried out to evaluate the relative importance of the 17 new input parameters among all 23 input parameters.

4.1 Preliminary Model Selection

Regression algorithms are suggested in this study as a data analysis step for choosing an algorithm that could predict the continuous values of WQI. The regression technique is a supervised machine learning algorithm that allows the prediction of a continuous outcome from one or more predictors. The basis of this technique often involves modeling the relationship and dependencies between the target output and input features to predict the value of a given new datapoint. The following five machine-learning algorithms are employed in this study.

4.1.1 Multilayer Perception

Multilayer perception (MLP) is a variant of the classical ANN model, popular for both classification and regression. MLP models operate on a network of sigmoid activation neurons connected

by links of several weights. The three basic layers of the MLP model consist of the input layer going through hidden layers to reach the output layer [31].

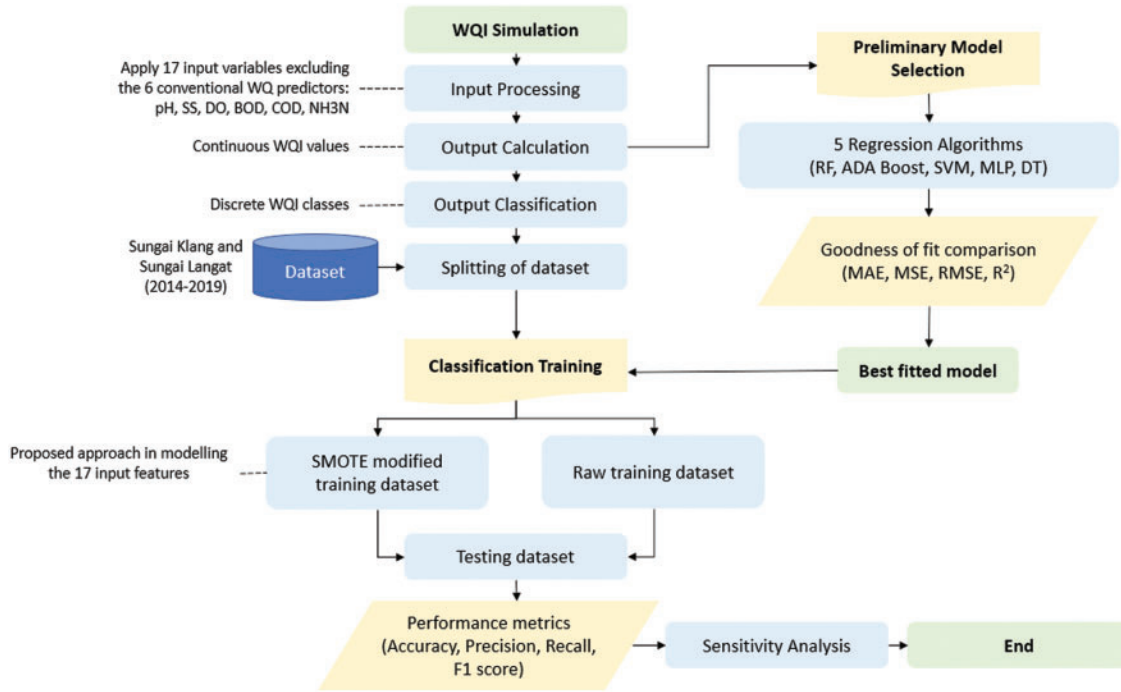


Figure 3: The proposed methodology

The net input Net_i is the addition of the threshold b_i with the summation of input values x (x_i to x_j) multiplied by assigned weights w (w_i to w_{ij}). The weights represent the strength of each neuron, which is optimized throughout the whole training process.

$$Net_i = b_i + \sum_{j=1}^n w_{ij}x_j \tag{2}$$

The activation function then accepts the input from the previous layer and transfers the output to the next layer [29].

$$f(Net_i) = \frac{1}{1 + e^{-Net_i}} \tag{3}$$

4.1.2 AdaBoost

Adaptive boosting (AdaBoost) is an ensemble learning-based model that combines multiple base learners, which generally outperform a single learner. The AdaBoost regression can automatically adjust the weightage of the model based on estimation errors by generating multiple regressors [32]. It has the potential to enhance the generalization capability of nonlinear and complicated regression problems [33].

4.1.3 Support Vector Regression Machine

Support vector regression machines (SVRMs) are a generalization of support vector machines to estimate the continuous output value. SVRMs are widely known in pattern recognition and regression

problems due to their generalization principle adopted from the structural risk minimization theory (SRM), which minimizes the empirical risks of overfitting in statistical learning theory [34]. The development of SVR models is aimed at reducing the upper bound error, where the four main kernel functions of the SVR model are the linear, polynomial, radial basis, and sigmoid functions [35].

The value of the kernel function $K(x_i, x_j)$ is equal to the inner product of two vectors, x_i and x_j . The common kernel function types are shown in Table 4, where γ , r , and d are kernel parameters.

Table 4: Kernel function [25]

Linear	$K(x_i, x_j) = x_i^T \times x_j$
Polynomial	$K(x_i, x_j) = (\gamma x_i^T \times x_j + r)^d$
Radial basis	$K(x_i, x_j) = \exp(-\ x_i - x_j\ ^2)$
Sigmoid	$K(x_i, x_j) = \tanh(x_i^T \times x_j + r)$

4.1.4 Decision Tree Regression

Decision tree regression (DTR) is a popular machine learning tool used in event outcome prediction, investment risks, and decision-making. Decision trees classify instances based on feature values, starting from the root node. The instances are represented with nodes, and each branch in the decision tree holds a value that the node can assume [36]. The decision tree model applies a top-down, recursive divide-and-conquer mode. After the feature of the root node is selected, the branch of each feature is given possible values, and this process is repeated recursively until all instances contain the same class. The decision tree approach is useful in predicting unpredictable outcomes by adding new scenarios into the complex dataset until a behavioral pattern is found [26]. The DTR algorithm extracts the features from the dataset, organizes the features into a tree-shaped structure, and combines a series of basic rules in the prediction stage [17].

4.1.5 Random Forest

Random forest (RF) is an ensemble learning algorithm that can be both a classification and a regression method. The RF creates several decision tree predictors that were initially uncorrelated from the sample dataset and assembles these weak learners into a strong learner [37]. The training data are randomized into a few samples and formed into several tree predictors. The RF algorithm shown in Fig. 4 generates results based on the average result of each tree [38,39].

4.2 Regression Evaluation

In order to evaluate the goodness of fit of the trained algorithms, four statistical measurements are applied to analyze the performance of the RF, AdaBoost, DTR, SVR, and MLP-NN models. The performance metrics applied are the mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and the coefficient of determination (R^2). MAE is often used to represent the discrepancy of error expected from the forecast on average, it is the average absolute difference between observed and predicted outcomes. Meanwhile, MSE and RMSE represent the squaring or square root of the computed error. Both measures are disproportionate to the actual increase in error. The R^2 metric measures the proportion variation of the outcome estimated by predictor variables. In regression models, R^2 is the squared correlation between observed and predicted values. A higher

R^2 and lower MAE, MSE, and RMSE values present better fitness and smaller discrepancy. These statistical measures are calculated based on the following formulas:

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n} \quad (4)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (6)$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (7)$$

Begin

Create n number of samples from total of N training data

In each single sample:

1. Randomly select m features from total features M
2. Calculate node using best split point
3. Divide into daughter nodes at optimal split
4. Build single tree iteratively from step 1-3

When n number of samples have formed T number of trees:

1. Test data is sent to each individual tree
2. Generate k number of results

Generate final results based on the average of results k

Select best tree predictors on the basis of majority votes

End

Figure 4: The random forest pseudocode applied in this work

4.3 Classification Training

Classification training refers to models' learning to categorize class labels based on the input data. The classification models assign label values and separate new observations into a specific class. In this study, classes are generated from the WQI ranges provided in [Table 2](#). Discrete categorical data are converted from the continuous WQI values based on the index range. There are a total of five classes with Class III as the majority class, while Class I and Class V have a limited number of observations. This section of the study includes the methodology proposed for addressing the case of an imbalanced dataset.

4.3.1 Synthetic Minority Oversampling Technique (SMOTE)

An imbalanced dataset occurs when the dataset contains classification categories that are not equally represented. Imbalanced datasets are commonly handled by using oversampling or under-sampling techniques. SMOTE is a feasible approach for oversampling the dataset with recreation of synthetic samples for the minority class [30]. The new data are sampled along the line segments joining k minority class samples based on interpolation ([Fig. 5](#)).

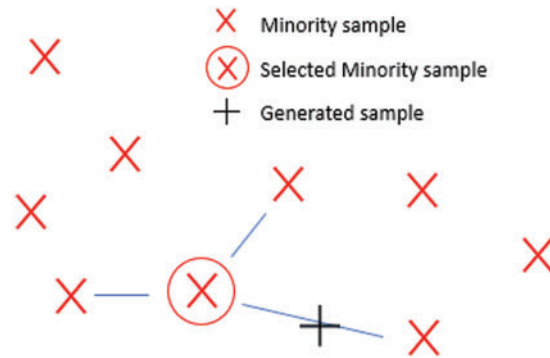


Figure 5: An illustration of the SMOTE interpolation of a selected minority sample, extracted from Douzas et al. [40]

4.3.2 Classification Evaluation

Several performance metrics are applied in the evaluation of the classification models. The confusion matrix is used to show information between the actual and predicted values performed by the classification system by means of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). A common measure of classification performance is prediction accuracy, which shows the proportion of the total number of correct predictions against the total number of the testing dataset.

$$Accuracy = \frac{TP + TN}{N} \times 100\% \quad (8)$$

where N is the total number of testing data.

Precision, recall, and F1-score are also calculated to evaluate the effectiveness of the classifiers. Precision measures the ratio of correct positive class predictions to the total predicted positives. Recall or sensitivity quantifies the ratio of correct positive class predictions to the total positive observations. The F-measure is a score balancing between precision and recall. Ideally, measures closer to one show the best performance.

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (11)$$

4.4 Sensitivity Analysis

Sensitivity analysis is the assessment of the importance of each input variable in the fitted model. The analysis is based on the 'leave-one-out' method, where changes in network error are recorded. The analysis performs parameter ranking according to the ratio of the leave-one-out model with the reference model. The results of sensitivity analysis often provide useful information on the effects of each predictor, which is insightful for pruning input parameters [7].

In this study, two sensitivity analyses are carried out to compare the rankings of all 23 parameters with the 17 new parameters. The method computes the misclassification rate when each variable is removed. The ratio of each variable is then calculated based on the error of the reduced model against the full model. A higher ratio returns higher significance of the parameter in the system.

5 Results and Discussion

5.1 Input Variables and Data Processing

A total of 5,483 samples of water quality parameters were obtained after pre-processing of the dataset received from the Department of Environment. Features such as *E. coli* and dissolved solids have very wide scales and distributions, which may lead to inaccuracy and degradation of the predictive performance, specifically algorithms that incorporate dimensional or distance elements. Hence, data normalization is applied to the dataset by setting the minimum and maximum threshold to be between $[-1, +1]$. To ensure each water quality variable has equal bounds in the analysis, the dataset is fit into a mean of zero and a variance of one.

Next, the water quality index (WQI) is calculated based on Eq. (1) and classified into five classes shown in Table 2 and Fig. 2. Split validation is used in this study by splitting the data into two groups: 80% for training and 20% for testing.

The correlation coefficient between input variables for the Langat and Klang Rivers is depicted in Table 5. It is apparent that salinity, DS, Cl, Ca, K, Mg, and Na are significantly correlated with conductivity at $r > 0.8$. A high positive correlation is also observed between Na and K ($r = 0.92$), which in turn are at $r = 0.97$ and 0.96 with Cl, respectively. The hardness of water could also be represented by the good relation between Ca and Cl (0.80). These values show the influence of ion concentration on the salinity and conductivity of the river.

The highest negative correlation is found between NO_3 and As (-0.04), and between Fe and K (-0.04), Mg (-0.03), and Na (-0.03). The presence of *E. coli* and total coliform has no effect on other water quality parameters, while both are closely related at $r = 0.62$.

Table 5: Correlation coefficient (r) between input variables with output WQI values and class

	COND	SAL	TUR	DS	NO ₃	Cl	PO ₄	As	Cr	Zn	Ca	Fe	K	Mg	Na	<i>E. coli</i>	TC	WQI	Class
COND	1.00																		
SAL	1.00	1.00																	
TUR	0.02	0.02	1.00																
DS	0.96	0.96	0.02	1.00															
NO ₃	0.00	0.00	0.03	0.00	1.00														
Cl	0.97	0.97	0.01	0.96	-0.01	1.00													
PO ₄	0.00	0.00	-0.02	0.01	0.04	-0.02	1.00												
As	0.06	0.05	0.01	0.06	-0.04	0.05	0.31	1.00											
Cr	0.03	0.03	0.02	0.03	0.12	0.03	0.07	0.02	1.00										
Zn	0.01	0.01	0.05	0.02	0.37	0.00	0.03	0.02	0.16	1.00									
Ca	0.80	0.79	0.02	0.80	0.07	0.80	0.14	0.14	0.03	0.07	1.00								
Fe	-0.03	-0.03	0.03	-0.03	0.00	-0.03	0.01	-0.01	0.05	0.01	-0.06	1.00							
K	0.93	0.93	0.02	0.93	0.00	0.96	0.04	0.07	0.03	0.01	0.83	-0.04	1.00						
Mg	0.85	0.85	0.02	0.85	-0.01	0.88	-0.02	0.03	0.02	0.00	0.71	-0.03	0.94	1.00					
Na	0.95	0.95	0.02	0.94	-0.01	0.97	-0.02	0.05	0.03	0.01	0.78	-0.03	0.92	0.85	1.00				
<i>E. coli</i>	0.00	0.00	-0.01	0.00	-0.02	-0.01	0.07	-0.03	0.00	0.01	0.07	0.00	0.01	-0.01	-0.01	1.00			
TC	0.00	-0.01	-0.02	0.00	-0.03	-0.01	0.07	-0.02	-0.01	-0.01	0.07	0.00	0.01	-0.01	-0.01	0.62	1.00		
WQI	-0.05	-0.05	-0.18	-0.05	-0.01	-0.02	-0.49	-0.33	-0.03	-0.06	-0.36	-0.03	-0.01	-0.01	-0.02	-0.14	-0.16	1.00	
Class	0.05	0.05	0.17	0.05	0.01	0.02	0.43	0.28	0.03	0.06	0.33	0.04	0.01	0.01	0.03	0.14	0.15	-0.91	1.00

5.2 Preliminary Model Selection

In this study, the Python programming language is used for data pre-processing and the prediction model. The scikit learn packages are applied to select the appropriate model for the prediction analysis. Five machine learning models are trained to predict WQI based on 17 input parameters, and the performance is shown in Figs. 6a–6e below. Based on the simulated and observed WQI for 20 sample data, it is evident that despite excluding the conventional six input parameters used to formulate WQI, the new 17 input variables are competent in achieving close prediction for WQI in most of the prediction models.

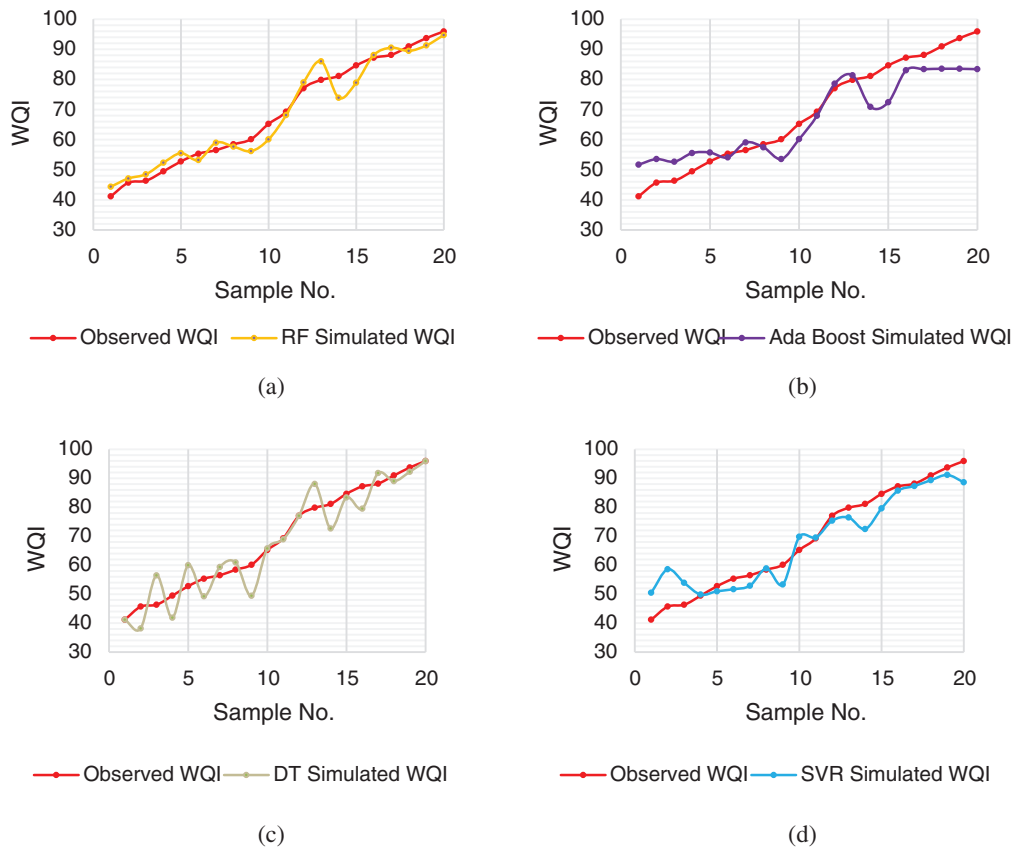


Figure 6: (Continued)

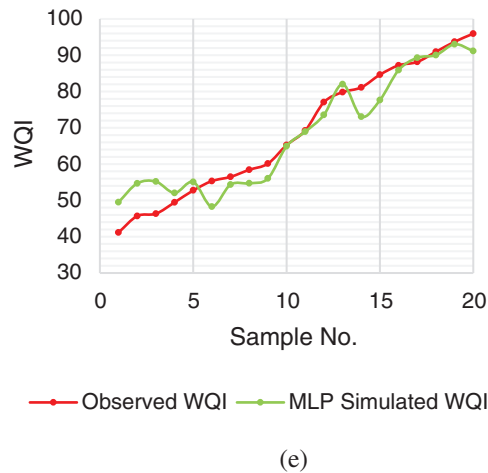


Figure 6: Comparison of the observed and simulated WQI using (a) random forest, (b) AdaBoost, (c) decision tree regression, (d) support vector regression, and (e) multilayer perception

The DTR model showed the highest fluctuation pattern and was unable to capture the pattern of the WQI accurately. Even though the AdaBoost model was able to reach exact predictions for a few of the datapoints, the model showed high deviation from the actual value for most of the samples. On the other hand, both SVR and MLP models are good at modeling the WQI, since several simulated values are in line with the observed values. Apparently, the RF model has better prediction performance, given its higher accuracy in simulating WQI compared to the other models. The RF model is better at simulating the pattern of the observed values.

Table 6 presents a comparison of performance metrics’ evaluation between the trained models. The results indicate that in most of the simulated WQIs, the RF model predominated, with the highest R^2 and a fair RMSE. The weakest performance was associated with the SVR model, for which the performance metrics have higher mean errors in both datasets. The AdaBoost, DTR, and MLP models produced average results compared to the other methods.

Table 6: Simulation results of the 5 machine learning algorithms

Parameter	Results			
	MAE	MSE	RMSE	R^2
Random forest	1.806	5.875	2.425	0.974
AdaBoost	5.791	51.507	7.173	0.753
DTR	4.759	40.733	6.382	0.807
SVR	6.012	62.709	7.915	0.690
MLP	5.484	52.479	7.241	0.742

The obtained RMSEs for RF, ADA, DTR, SVR, and MLP are 2.425, 7.173, 6.382, 7.915, and 7.241, respectively. The RF model provides the smallest RMSE, MSE, and MAE among the other algorithms, which confirms the capability of the RF model to predict WQI. The reason RF outperforms the other models is its capability of learning nonlinear relationships between the WQI and input layers. Random forest possesses higher robustness in handling outliers and unbalanced datasets, as it is efficient in decreasing bias and overfitting of data [41]. Thus, the overall results show the superior performance of RF models with the input combination COND, SAL, TUR, DS, NO₃, Cl, PO₄, As, Cr, Zn, Ca, Fe, K, Mg, Na, *E. coli*, and total coliform. All five techniques were compared to the actual outcome, plotted in Fig. 7. The results suggest that the RF regression method provides better predictive accuracy in terms of all performance evaluation parameters. The pattern of the plot obtained portrays the linearity of the algorithm, which has moderate accuracy. The substantial improvement in the predictive accuracy of the RF approach using the alternative set of input features indicates that it can be effectively used in predicting the impact of water quality.

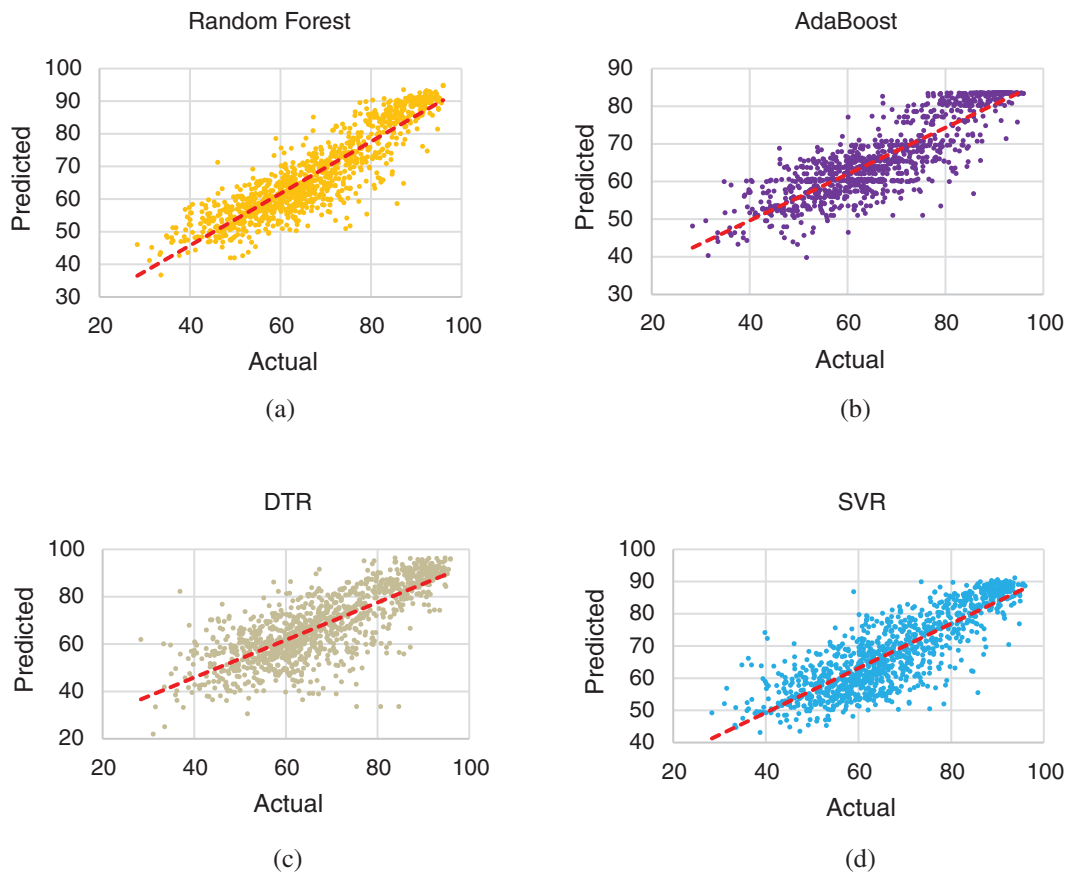


Figure 7: (Continued)

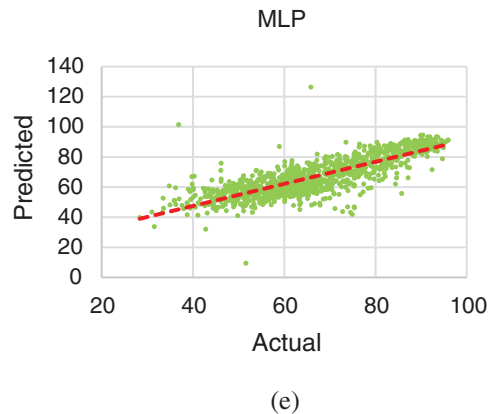


Figure 7: Correlation between the observed and simulated WQIs using (a) random forest, (b) AdaBoost, (c) decision tree regression, (d) support vector regression, and (e) multilayer perceptron

5.3 Classification Training

As per the investigation conducted in 5.2, the RF model is proven competent in modeling the behavior of the 17 water quality inputs, which correspond to the continuous WQI output. This section of the study performs classification modeling of the five WQ classes shown in Fig. 2 using a random forest classifier. The study aims to improve the efficiency of classification prediction through a modified training approach. Two RF classification models will be tested on two different datasets: Model A is trained on Training Set A, which contains an unmodified training dataset, and Model B is based on the SMOTE-modified Training Set B. Models A and B are then tested on a new testing set, and both models' performance is then compared in terms of their confusion matrix, accuracy, precision, and recall, tabulated in Table 7.

Table 7: Performance metrics for WQI classification of RF Model A and Model B

Class	Precision	Recall	F1-score
<i>Model A (Accuracy = 79.66%)</i>			
I	0.33	0.10	0.15
II	0.85	0.80	0.82
III	0.80	0.92	0.85
IV	0.66	0.44	0.53
V	0.00	0.00	0.00
Macro average	0.53	0.45	0.47
<i>Model B (Accuracy = 77.68%)</i>			
I	0.54	0.48	0.51
II	0.80	0.78	0.79
III	0.83	0.83	0.83
IV	0.52	0.59	0.56

(Continued)

Table 7 (continued)

Class	Precision	Recall	F1-score
V	1.00	0.75	0.86
Macro average	0.74	0.69	0.71

The recall rate (Table 7) shows that the sensitivity of Model A towards Class V is the worst. This is critical, as dangerous water that is incorrectly classified could be consumed. An improvement in recall is seen in Model B, with a 75% sensitivity rate. Moreover, the precision and recall rates of Class I also improve from 33% to 54% and 10% to 48% when Model B is trained with the SMOTE modified training dataset. However, the accuracy of Model B is much lower, as Model B’s performance in predicting Classes II and III shows a decline. This is possibly due to the training process of SMOTE—the presence of indistinguishable boundaries between the two classes introduces noise to the model [40]. Nonetheless, Model B has the best performance in average F1-score as compared to the unmodified Model A. In the case of water classification, precision and sensitivity are the most important estimates. In Fig. 8, class-wise accuracy is represented in the form of a confusion matrix. Fig. 9 provides a visual overview of the classification performance of Model A and Model B. Significant improvements are observed for predictions of Class V and Class I. Overall, the accuracy of both models is substantially similar in agreement, at 79.66% for Model A and 77.68% for Model B.

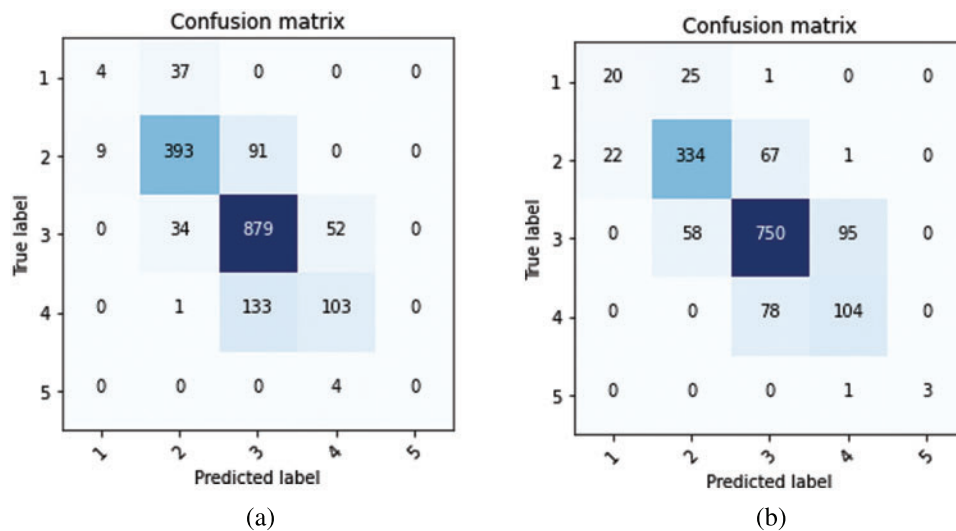


Figure 8: Confusion matrix for WQI classification of random forest (a) Model A, (b) Model B

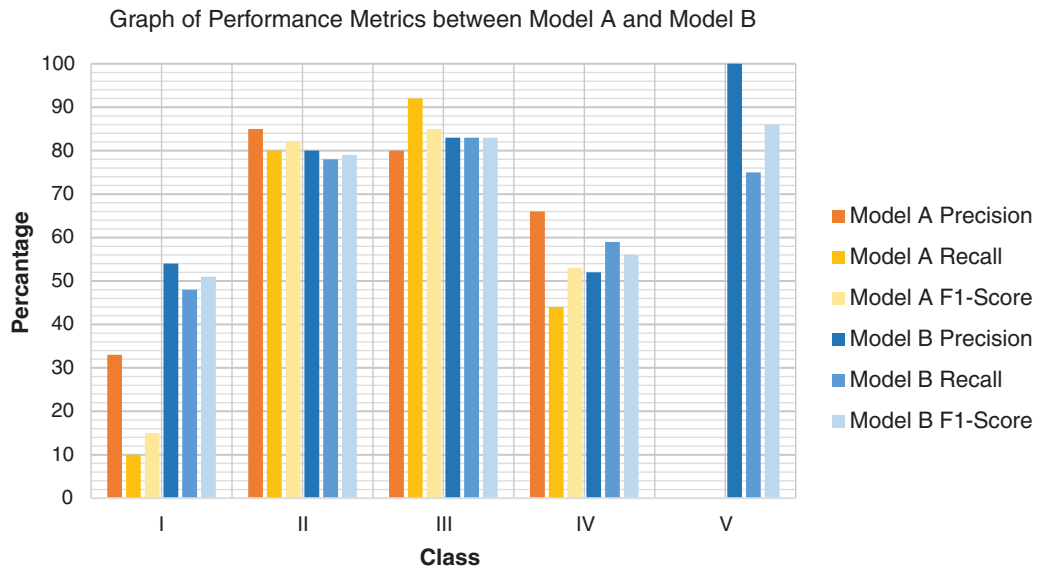


Figure 9: Visual comparison of the precision, recall, and F1-score between Model A and Model B

5.4 Sensitivity Analysis

The results of sensitivity analysis among the 17 and 23 input variables (Table 8) highlight the importance of certain water quality variables that are not present in the current WQI formulation. The rankings from Table 8 show turbidity as the first determining factor for predicting WQI with unconventional features. The firm ground of turbidity’s importance was also proven when the 23 predictors were analyzed. Turbidity ranked as the sixth parameter, while the first five parameters were the conventional input features. The common input feature pH shows less significance in this study as the 13th variable. Moreover, sensitivity analysis also shows that the variables sodium, chloride, and conductivity were less influential in both WQI predictions. These variables could be removed from the future experimental phase to provide better generalization and less noise for the model.

Table 8: Sensitivity analysis among 17 and 23 predictors

Ratio (17 predictors)	Input variables	Ranking	Input variables	Ratio (23 predictors)
1.104	TUR	1	DO	1.962
1.061	Total coliform	2	NH ₃ N	1.482
1.049	NO ₃	3	BOD	1.352
1.046	DS	4	SS	1.209
1.044	Mg	5	COD	1.177
1.043	As	6	TUR	1.004
1.043	K	7	Ca	1.001
1.041	Ca	8	SAL	0.999
1.035	Fe	9	Zn	0.998
1.029	SAL	10	PO ₄	0.986

(Continued)

Table 8 (continued)

Ratio (17 predictors)	Input variables	Ranking	Input variables	Ratio (23 predictors)
1.025	PO ₄	11	K	0.985
1.020	Cr	12	NO ₃	0.983
1.019	Zn	13	pH	0.977
1.015	<i>E. coli</i>	14	<i>E. coli</i>	0.977
1.007	Cl	15	Total coliform	0.973
1.001	COND	16	Fe	0.970
0.993	Na	17	Mg	0.969
–	–	18	As	0.965
–	–	19	Cr	0.963
–	–	20	COND	0.961
–	–	21	DS	0.951
–	–	22	Cl	0.940
–	–	23	Na	0.925

5.5 Discussion

The significance of unconventional water quality variables has been constantly published in research materials that highlight the absence of COND, DS, Cl, Na, Cl, As, Zn, Fe, and other heavy metal concentrations such as aluminum (Al), cadmium (Cd), chromium (Cr), and lead (Pb) from drinking water quality standards [42]. Results from Wagh et al. [43] also further showed that groundwater is constantly saturated with pollutants such as COND, Na, Cl, and DS which exceeded the desirable limits. Areas that are affected by urban land use activities often have COND, NO₃, PO₄, Cl, DO, BOD, Pb, Cd, and total coliform values above the stipulated standards [44].

Therefore, it is evident that alternative water quality parameters should be included in assessing water quality. The hypothesis is proven in this study that these inputs will contribute towards accurate prediction of water quality. Although none of the inputs used were part of the conventional WQI subindex's formula, the proposed modified RF model managed to achieve a coefficient of determination of 0.974 and error below 6%. The proposed model indicates that 97.4% of the variability of WQI can be explained by the 17 water quality inputs. Nonetheless, a modified approach is required when handling the classification of the dataset. From Section 5.3, it is evident that the proposed integration of SMOTE with RF could improve the real positives predicted.

The findings from this study correspond to previous works tabulated in Table 9, which shows that the proposed modified RF model is efficient in WQ modeling. The adoption of SVMs in [45] suggested that, if the least relevant parameters are included, the SVM model has less advantage in prediction. This is also proven in this study, as the SVR model showed R² of only 0.69. On the other hand, Ho et al. [26] conducted a leave-one-out analysis on all six common parameters and concluded that exclusion of DO, BOD, and COD had a large impact on accuracy. Sensitivity analysis by Gazzaz et al. [7] pointed out that, for the Perak River, DO, BOD, NH₃-N, pH, COD, and turbidity should receive priority in

WQI computation instead of using SS. The research ranked SS concentration 18th in importance, while the WQI ranked SS concentration third.

Table 9: Comparison of results in this study with other research works

References	Location	Input data	Prediction algorithm/Performance evaluation	Remarks
This study	Klang and Langat Rivers, Selangor	COND, SAL, TUR, DS, NO ₃ , Cl, PO ₄ , As, Cr, Zn, Ca, Fe, K, Mg, Na, <i>E. coli</i> , total coliform	RF // R ² = 0.974, Accuracy = 77.68%, Precision = 74%, Recall = 69%, F1-score = 71%	Although the inputs used do not contain relative weights in the WQI formula, the output generated is accurate with the use of random forest in nonlinear data. Classification training using modified SMOTE training data could improve the efficiency and reliability of the classifier.
Leong et al. [45]	Perak River, Perak	DO, BOD, COD, SS, NH ₃ -N, pH	SVM // R ² = 0.8796 LS-SVM // R ² = 0.9227	In this study, using support vector machines, the model in the least squares method is much more stable and accurate. The authors also trained the model using 31 predictors but concluded that too many predictors lowered the accuracy, as most of the variables had low correlation with WQI.

(Continued)

Table 9 (continued)

References	Location	Input data	Prediction algorithm/Performance evaluation	Remarks
Ho et al. [26]	Klang River, Selangor	DO, BOD, COD, SS, NH ₃ -N, pH	Decision Tree // Accuracy w/o NH ₃ -N: 84.09% w/o NH ₃ -N, pH: 81.82% w/o NH ₃ -N, pH, SS: 77.27%	The research concluded that WQI can be predicted with decision trees without much loss of information if certain variables such as NH ₃ -N are excluded from the input. The most important parameters corresponding to the relative weights are DO, BOD, and COD.
Gazzaz et al. [7]	Kinta River, Perak	DO, BOD, COD, SS, NH ₃ -N, pH, Turbidity, Mg, Ca, K, Cl, WT, <i>E. coli</i> , Zn, DS, Fe, TS, Na, SS, EC, NO ₃ -N, As, PO ₄ -P, Total coliform	ANN // R ² = 0.954	The study applied principal factor analysis (PFA) and identified 23 out of the 31 water quality variables that are responsible for the variation of water data. The inputs used contained indicators of nutrients, microbes, and heavy metals that are not included in the current WQI formula. The ANN approach is capable of calculating and forecasting water quality in a more robust manner.

In summary, the use of other WQ parameters should not be entirely neglected when assessing WQI. The WQI formulation should be tailored according to the river's needs. Besides that, a suitable machine learning model as suggested in this work is important to model the nonlinear inputs.

The use of the proposed new water quality variables is beneficial because most of the water quality parameters are interrelated. In the published work by Joarder et al. [46], electrical conductivity was analyzed as one of the most appropriate variables to justify most of the dependent variables for WQI. Electrical conductivity is also directly related to the content of DS and salinity, where a high electrical

conductivity signifies a significant number of impurities present in the water source. Commonly, the impurities consist of ions such as chloride, phosphate, and nitrate from sewage runoff and agricultural waste. Chloride content dominant in water content could cause corrosion of iron plates or pipes [47]. Excess nitrate and phosphate concentration is toxic to human consumption and could stimulate algal blooms that would cause oxygen depletion or eutrophication [48]. However, previous works only considered DO in the WQI formulation to represent these reactions, which can be considered inadequate and less robust.

Moreover, the measure of turbidity is also influential in predicting WQI yet neglected by the conventional WQI formula. Turbidity describes the clarity of water, where high turbidity represents a dense mixture of clay and organic matter. The presence of these foreign compounds directly affects light transmission to aquatic plants; hence, lowered photosynthesis would result in lower dissolved oxygen and unfavorable water quality [49]. In addition, the potential risk of heavy metal contamination originating from water sources was also found to affect the mangrove forests of Kuala Selangor [50] and the Klang estuary [51]. Since these metal concentrations are not present in the conventional WQI measurement, the estuary ecosystem is unknowingly still being fed from heavy metal-contaminated water sources even though WQI is continuously monitored.

Nevertheless, despite the high compliance of the RF model in modeling the nonlinear relation of these inputs, the simulation model could be improved through optimization algorithms to reduce the errors. The high average magnitude of errors is because the target WQI does not include any of the proposed new input variables. Furthermore, with the contribution of the SMOTE resampling technique, the accuracy of the imbalanced dataset has been improved for minority classes. Model accuracy, on the other hand, has been improved by implementing k-means clustering during oversampling to remove unnecessary noise generated during the interpolation phase [40]. The adoption of novel input water quality parameters can be further investigated through the use of deep learning predictors that are proven effective and stable [52,53]. Subsequently, in future sustainability water resources management, reassessment of WQI formulation can be carried out to fit more water quality parameters. This work serves as a step forward toward utilizing more biological predictors in water quality monitoring.

6 Conclusion

This study explored the prospective use of unconventional input parameters for water quality index (WQI) prediction using a random forest (RF) regression and classification model. The new input features investigated are able to simulate WQI fairly. Comprehensive comparisons between the performance of five machine learning models were carried out. The findings revealed that the RF model exhibits better prediction performance given its ability to capture the nonlinear characteristics of input variables to the water quality index. Therefore, it can be concluded that the RF model is suitable for forecasting the impact of pollutants on water quality based on the inputs as investigated in this study. This study further proposed a modified RF model by incorporating the SMOTE technique to address the imbalance in minority classes. The findings showed drastic improvement in the proposed model's sensitivity, from 45% to 69%, and precision, from 53% to 74%. Sensitivity analysis provided insights on the influential parameter of turbidity on water quality prediction. Although the current findings are insufficient to formulate a new water quality index, this paper demonstrates the importance and capability of other water quality parameters as a step toward better water quality representation. The results aligned with other research that highlighted the influence of these parameters on water quality. Hence, it is crucial that the input parameters used in formulating WQI should include inorganic compounds and heavy metal constituents. Future works

can be extended from this study to include socioeconomic factors for source identification of these pollutants to offer flexibility in water quality assessment. The model can also be extended to more datasets before it is widely adopted for new WQI computation.

Funding Statement: This study is supported by the Ministry of Higher Education through MRUN Young Researchers Grant Scheme (MY-RGS), MR001-2019, entitled “Climate Change Mitigation: Artificial Intelligence-Based Integrated Environmental System for Mangrove Forest Conservation” and UM-RU Grant, ST065-2021, entitled “Climate-Smart Mitigation and Adaptation: Integrated Climate Resilience Strategy for Tropical Marine Ecosystem.” The authors are grateful to the Department of Environment Malaysia (DOE) for the provided data.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Bakker, K. (2012). Water security: Research challenges and opportunities. *Science*, 337(6097), 914–915. DOI 10.1126/science.1226337.
2. Das Kangabam, R., Govindaraju, M. (2019). Anthropogenic activity-induced water quality degradation in the Loktak Lake, a Ramsar site in the Indo-Burma biodiversity hotspot. *Environmental Technology*, 40(17), 2232–2241. DOI 10.1080/09593330.2017.1378267.
3. Ramos, M. A. G., de Oliveira, E. S. B., Pião, A. C. S., de Oliveira Leite, D. A. N., de Angelis, D. F. (2016). Water quality index (WQI) of Jaguari and Atibaia Rivers in the region of Paulínia, São Paulo. *Brazil Environmental Monitoring and Assessment*, 188(5), 263. DOI 10.1007/s10661-016-5261-z.
4. Tan Pei Jian, B., Ul Mustafa, M. R., Hasnain Isa, M., Yaqub, A., Ho, Y. C. Y. (2020). Study of the water quality index and polycyclic aromatic hydrocarbon for a river receiving treated landfill leachate. *Water*, 12(10), 2877. DOI 10.3390/w12102877.
5. Ahmed, F., Siwar, C., Begum, R. A. (2014). Water resources in Malaysia: Issues and challenges. *Journal of Food, Agriculture and Environment*, 12(2), 1100–1104.
6. DOE, Environmental Quality Report (EQR) (2017). *Department of environment*. Malaysia: Ministry of Energy, Science, Technology, Environment and Climate Change.
7. Gazzaz, N. M., Yusoff, M. K., Aris, A. Z., Juahir, H., Ramli, M. F. (2012). Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. *Marine Pollution Bulletin*, 64(11), 2409–2420. DOI 10.1016/j.marpolbul.2012.08.005.
8. Kachroud, M., Trolard, F., Kefi, M., Jebari, S., Bourrié, G. (2019). Water quality indices: Challenges and application limits in the literature. *Water*, 11(2), 361. DOI 10.3390/w11020361.
9. Uddin, M. G., Nash, S., Olbert, A. I. (2021). A review of water quality index models and their use for assessing surface water quality. *Ecological Indicators*, 122, 107218. DOI 10.1016/j.ecolind.2020.107218.
10. Juwana, I., Muttill, N., Perera, B. J. C. (2016). Uncertainty and sensitivity analysis of West Java Water Sustainability Index—A case study on Citarum catchment in Indonesia. *Ecological Indicators*, 61(1), 170–178. DOI 10.1016/j.ecolind.2015.08.034.
11. Seifi, A., Dehghani, M., Singh, V. P. (2020). Uncertainty analysis of water quality index (WQI) for groundwater quality evaluation: Application of Monte-Carlo method for weight allocation. *Ecological Indicators*, 117(2–4), 106653. DOI 10.1016/j.ecolind.2020.106653.
12. Wu, Z., Wang, X., Chen, Y., Cai, Y., Deng, J. (2018). Assessing river water quality using water quality index in Lake Taihu Basin, China. *Science of the Total Environment*, 612(2), 914–922. DOI 10.1016/j.scitotenv.2017.08.293.

13. Naubi, I., Zardari, N. H., Shirazi, S. M., Ibrahim, N. F. B., Baloo, L. (2016). Effectiveness of water quality index for monitoring Malaysian river water quality. *Polish Journal of Environmental Studies*, 25(1), 231–239. DOI 10.15244/pjoes/60109.
14. DOE, Environmental Quality Report (EQR) 2008 (2008). Department of Environment. Kuala Lumpur: Ministry of Natural Resources and Environment Malaysia.
15. Othman, F., Alaaeldin, M. E., Seyam, M., Ahmed, A. N., Teo, F. Y. et al. (2020). Efficient river water quality index prediction considering minimal number of inputs variables. *Engineering Applications of Computational Fluid Mechanics*, 14(1), 751–763. DOI 10.1080/19942060.2020.1760942.
16. Tung, T. M., Yaseen, Z. M. (2020). A survey on river water quality modelling using artificial intelligence models: 2000–2020. *Journal of Hydrology*, 585(3731), 124670. DOI 10.1016/j.jhydrol.2020.124670.
17. Asadollah, S. B. H. S., Sharafati, A., Motta, D., Yaseen, Z. M. (2021). River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *Journal of Environmental Chemical Engineering*, 9(1), 104599. DOI 10.1016/j.jece.2020.104599.
18. Dashti Latif, S., Najah Ahmed, A., Sherif, M., Sefelnasr, A., El-Shafie, A. (2021). Reservoir water balance simulation model utilizing machine learning algorithm. *Alexandria Engineering Journal*, 60(1), 1365–1378. DOI 10.1016/j.aej.2020.10.057.
19. Ibrahim, K. S. M. H., Huang, Y. F., Ahmed, A. N., Koo, C. H., El-Shafie, A. (2022). A review of the hybrid artificial intelligence and optimization modelling of hydrological streamflow forecasting. *Alexandria Engineering Journal*, 61(1), 279–303. DOI 10.1016/j.aej.2021.04.100.
20. Nur Adli Zakaria, M., Abdul Malek, M., Zolkepli, M., Najah Ahmed, A. (2021). Application of artificial intelligence algorithms for hourly river level forecast: A case study of Muda River, Malaysia. *Alexandria Engineering Journal*, 60(4), 4015–4028. DOI 10.1016/j.aej.2021.02.046.
21. Hameed, M., Sharqi, S. S., Yaseen, Z. M., Afan, H. A., Hussain, A. et al. (2017). Application of artificial intelligence (AI) techniques in water quality index prediction: A case study in tropical region, Malaysia. *Neural Computing and Applications*, 28(1), 893–905. DOI 10.1007/s00521-016-2404-7.
22. Perea, R. G., Poyato, E. C., Montesinos, P., Díaz, J. A. R. (2019). Optimisation of water demand forecasting by artificial intelligence with short data sets. *Biosystems Engineering*, 177(2), 59–66. DOI 10.1016/j.biosystemseng.2018.03.011.
23. Tiwari, S., Babbar, R., Kaur, G. (2018). Performance evaluation of two ANFIS models for predicting water quality index of River Satluj (India). *Advances in Civil Engineering*, 2018(3), 8971079. DOI 10.1155/2018/8971079.
24. Elkiran, G., Nourani, V., Abba, S. I. (2019). Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach. *Journal of Hydrology*, 577(2–4), 123962. DOI 10.1016/j.jhydrol.2019.123962.
25. Abobakr Yahya, A. S., Ahmed, A. N., Othman, F. B., Ibrahim, R. K., Afan, H. A. et al. (2019). Water quality prediction model based support vector machine model for ungauged river catchment under dual scenarios. *Water*, 11(6), 1231. DOI 10.3390/w11061231.
26. Ho, J. Y., Afan, H. A., El-Shafie, A. H., Koting, S. B., Mohd, N. S. et al. (2019). Towards a time and cost effective approach to water quality index class prediction. *Journal of Hydrology*, 575(1), 148–165. DOI 10.1016/j.jhydrol.2019.05.016.
27. Mohamed, I., Othman, F., Ibrahim, A. I., Alaa-Eldin, M., Yunus, R. M. (2015). Assessment of water quality parameters using multivariate analysis for Klang River basin, Malaysia. *Environmental Monitoring and Assessment*, 187(1), 1–12. DOI 10.1007/s10661-014-4182-y.
28. Sharif, S. M., Kusin, F. M., Asha'ari, Z. H., Aris, A. Z. (2015). Characterization of water quality conditions in the Klang River Basin, Malaysia using self organizing map and K-means algorithm. *Procedia Environmental Sciences*, 30, 73–78. DOI 10.1016/j.proenv.2015.10.013.
29. Raheli, B., Aalami, M. T., El-Shafie, A., Ghorbani, M. A., Deo, R. C. (2017). Uncertainty assessment of the multilayer perceptron (MLP) neural network model with implementation of the novel hybrid MLP-FFA

- method for prediction of biochemical oxygen demand and dissolved oxygen: A case study of Langat River. *Environmental Earth Sciences*, 76(14), 1–16. DOI 10.1007/s12665-017-6842-z.
30. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. DOI 10.1613/jair.953.
 31. Ay, M., Kisi, O. (2012). Modeling of dissolved oxygen concentration using different neural network techniques in Foundation Creek, El Paso County, Colorado. *Journal of Environmental Engineering*, 138(6), 654–662. DOI 10.1061/(ASCE)EE.1943-7870.0000511.
 32. Zhu, X., Zhang, P., Xie, M. (2021). A joint long short-term memory and AdaBoost regression approach with application to remaining useful life estimation. *Measurement*, 170(4), 108707. DOI 10.1016/j.measurement.2020.108707.
 33. Zhang, P., Yang, Z. (2018). A novel AdaBoost framework with robust threshold and structural optimization. *IEEE Transactions on Cybernetics*, 48(1), 64–76. DOI 10.1109/TCYB.2016.2623900.
 34. Nayak, J., Naik, B., Behera, D. H. (2015). A comprehensive survey on support vector machine in data mining tasks: Applications & challenges. *International Journal of Database Theory and Application*, 8(1), 169–186. DOI 10.14257/ijtda.2015.8.1.18.
 35. Li, J., Abdulmohsin, H. A., Hasan, S. S., Kaiming, L., Al-Khateeb, B. et al. (2019). Hybrid soft computing approach for determining water quality indicator: Euphrates River. *Neural Computing and Applications*, 31(3), 827–837. DOI 10.1007/s00521-017-3112-7.
 36. Kotsiantis, S. B., Zaharakis, I. D., Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, 26(3), 159–190. DOI 10.1007/s10462-007-9052-3.
 37. Singh, B., Sihag, P., Singh, K. (2017). Modelling of impact of water quality on infiltration rate of soil by random forest regression. *Modeling Earth Systems and Environment*, 3(3), 999–1004. DOI 10.1007/s40808-017-0347-3.
 38. Rodriguez-Galiano, V., Mendes, M. P., Garcia-Soldado, M. J., Chica-Olmo, M., Ribeiro, L. (2014). Predictive modeling of groundwater nitrate pollution using random forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). *Science of the Total Environment*, 476–477, 189–206. DOI 10.1016/j.scitotenv.2014.01.001.
 39. Wang, J., Yan, W., Wan, Z., Wang, Y., Lv, J. et al. (2020). Prediction of permeability using random forest and genetic algorithm model. *Computer Modeling in Engineering & Sciences*, 125(3), 1135–1157. DOI 10.32604/cmescs.2020.014313.
 40. Douzas, G., Bacao, F., Last, F. (2018). Improving imbalanced learning through a heuristic over-sampling method based on K-means and SMOTE. *Information Sciences*, 465(1), 1–20. DOI 10.1016/j.ins.2018.06.056.
 41. Hidayat, F., Astsauri, T. M. S. (2021). Applied random forest for parameter sensitivity of low salinity water injection (LSWI) implementation on carbonate reservoir. *Alexandria Engineering Journal*, 61(3), 2408–2417. DOI 10.1016/j.aej.2021.06.096.
 42. Ahmed, M., Mokhtar, M., Abd Majid, N. (2021). Household water filtration technology to ensure safe drinking water supply in the Langat River Basin. *Malaysia Water*, 13(8), 1032. DOI 10.3390/w13081032.
 43. Wagh, V. M., Panaskar, D. B., Varade, A. M., Mukate, S. V., Gaikwad, S. K. et al. (2016). Major ion chemistry and quality assessment of the groundwater resources of Nanded Tehsil, a part of south-east Deccan Volcanic Province, Maharashtra, India. *Environmental Earth Sciences*, 75(21), 1–26. DOI 10.1007/s12665-016-6212-2.
 44. Okeke, P. N., Adinna, E. N. (2013). Water quality study of Ontamiri River in Owerri, Nigeria. *Universal Journal of Environmental Research & Technology*, 3(6), 641–649.
 45. Leong, W. C., Bahadori, A., Zhang, J., Ahmad, Z. (2021). Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM). *International Journal of River Basin Management*, 19(2), 149–156. DOI 10.1080/15715124.2019.1628030.

46. Joarder, M., Raihan, F., Alam, J., Hasanuzzaman, S. (2008). Regression analysis of ground water quality data of Sunamganj District, Bangladesh. *International Journal of Environmental Research*, 2(3), 291–296.
47. Jothivenkatachalam, K., Nithya, A., Mohan, S. C. (2010). Correlation analysis of drinking water quality in and around Perur block of Coimbatore District, Tamil Nadu, India. *Rasayan Journal of Chemistry*, 3(4), 649–654.
48. Kihampa, C., Wenaty, A. (2013). Impact of mining and farming activities on water and sediment quality of the Mara River Basin, Tanzania. *Research Journal of Chemical Sciences*, 3(7), 15–24.
49. Berry, W., Rubinstein, N., Melzian, B., Hill, B. (2003). The biological effects of suspended and bedded sediment (SABS) in aquatic systems: A review. In: *Internal report*, vol. 32, no. 1, pp. 54–55. United States Environmental Protection Agency, Duluth. <https://archive.epa.gov/epa/sites/production/files/2015-10/documents/sediment-appendix1.pdf>.
50. Nyangon, L., Zainal, A. N. S., Pazi, A. M. M., Gandaseca, S. (2019). Heavy metals in mangrove sediments along the Selangor River, Malaysia. *Forest and Society*, 3(2), 278–288. DOI 10.24259/fs.v3i2.6345.
51. Elturk, M., Abdullah, R., Mohamad Zakaria, R., Abu Bakar, N. K. (2019). Heavy metal contamination in mangrove sediments in Klang estuary, Malaysia: Implication of risk assessment. *Estuarine, Coastal and Shelf Science*, 226(1–3), 106266. DOI 10.1016/j.ecss.2019.106266.
52. Jin, X. B., Zheng, W. Z., Kong, J. L., Wang, X. Y., Zuo, M. et al. (2021). Deep-learning temporal predictor via bidirectional self-attentive encoder-decoder framework for IOT-based environmental sensing in intelligent greenhouse. *Agriculture*, 11(8), 802. DOI 10.3390/agriculture11080802.
53. Kong, J., Yang, C., Wang, J., Wang, X., Zuo, M. et al. (2021). Deep-stacking network approach by multisource data mining for hazardous risk identification in IoT-based intelligent food management systems. *Computational Intelligence and Neuroscience*, 2021(9), 1194565. DOI 10.1155/2021/1194565.