JEE Journal of Ecological Engineering

Journal of Ecological Engineering 2023, 24(2), 152–164
https://doi.org/10.12911/22998993/156663
ISSN 2299–8993, License CC-BY 4.0

Received: 2022.10.24
Accepted: 2022.12.10
Published: 2023.01.01

Stacking Artificial Intelligence Models for Predicting Water Quality Parameters in Rivers

Mohammad Almadani[1*], Marwan Kheimi[1],

[1] Department of Civil Engineering, Faculty of Engineering – Rabigh Branch, King Abdulaziz University, Jeddah 21589, Saudi Arabia

[*] Corresponding author's e-mail: malmadani@kau.edu.sa

ABSTRACT

Scrutinizing the changes in the quality of river water is one of the main factors of monitoring the quality of natural flows, which plays a crucial role in the sustainable management of these ecosystems. The concentration of dissolved oxygen (DO) in river water is one of the most important indicators of quality management in such water bodies. From an environmental point of view, exceeding the permissible and natural decay capacity of pollutants in natural streams leads to a decrease in DO and, consequently, causes serious risks for the survival of aquatic life in related ecosystems. Hence, in the present study, 10 daily variables with the amount of dissolved oxygen on the same day were collected and evaluated from Allen County. Moreover, half of these variables were chosen as effective inputs to the model based on statistical analysis, so as to calculate the dissolved oxygen concentration parameter. Modeling with artificial intelligence approaches was implemented in the form of four individual methods: ANFIS-PSO, OS-ELM, Bagging-RF and Boosting CART, and two ensemble-stacking methods: SMA and Meta-learner MLP. The outcomes of estimating the DO with RMSE, MAE, GRI, r, and MBE criteria and marginal-scatter and subject profile diagrams were discussed. Moreover, the efficiency of the models in estimating the outlier of the observational data was scrutinized by subject profile diagram. Finally, it was found that the Meta-learner MLP model with RMSE of 0.965 mg/L had improvement in performance by 8.8%, 8.9%, 22.3%, 24.9% and 27.6%, respectively, compared to SMA, Boosting CART, Bagging-RF, ANFIS-PSO and OS-ELM methods. This remarkable improvement led to recommendations for using stacking techniques in water quality modeling and simulation.

Keywords: dissolved oxygen; water quality; ensemble-stacking model; meta-learner.

INTRODUCTION

Natural flows are considered as one of the main sources of fresh water for diverse purposes (including drinking, agriculture, and industry). Therefore, rivers are one of the basic foundations of sustainable and environmentally friendly development in human societies, while industrial and welfare developments have led to increasing stresses on river water quality, so that these vital and valuable resources are exposed to danger. Hence, given the major impact of human activities on the changes in river water quality, evaluating river flows with a qualitative modeling approach is of a great importance in studying water resources [Abazi et al., 2022; Lusiana et al., 2022; Rahutami et al., 2022]. In the present investigation, despite diverse qualitative variables (which are included in the model), the qualitative study of river water was taken into account. At the same time, the amount of dissolved oxygen acts as the output of the model. Emphasis on the variable of dissolved oxygen and its acceptance in the role of the target parameter in modeling cover the set of reactions involved in the amount of oxygen. To put it another way, various factors affect the alterations in the concentration of dissolved oxygen in the river, in such a way that we can refer to its increase through direct absorption from the atmosphere and photosynthesis of algae (algae photosynthesis) and its decrease due to consumption in chemical and biological reactions during the processes of decay of pollution load in the river, oxidation of sediments and algae respiration [Benedini and Tsakiris, 2013]. Hence, the concentration of dissolved oxygen as a quality item is a good indicator of the condition of the water body of the river, demonstrating the resultant effect of physical, chemical and biological properties. All of the aforementioned items led to predicting the amount of dissolved oxygen in a natural flow in this study. Besides, physical variables such as water temperature and chemical variables such as phosphorus concentration were used in modeling. It is interesting to note that physical, chemical and biological classifications for qualitative variables are not an easy task at all since a number of qualitative parameters are the result of a set of physico-chemical-biological reactions and we can put them into various groups at the same time. This problem is designed while considering such an approach, so as to be in line with the comprehensiveness required in water resources management.

It is noteworthy that the use of physical, mathematical, and numerical modeling for simulating river pollution problems have been used widely during the last decades [Schaffner et al., 2009; Kisi and Parmar, 2016; Drozdov et al., 2021; Zounemat-Kermani et al., 2021a]. Mathematical water quality modeling has proved as a reliable and cost-effective approach to simulating pollutant distribution in surface waters and rivers that can be successfully employed in water resources planning and management. It should be noted that modeling is not a substitute to the field observations but it can be considered and used as a proper alternative in simulating or understanding observations under certain circumstances.

On the other hand, the demand for increasing accuracy in modeling water quality issues has led to a focus on the implementation of artificial intelligence methods in this field. During the last decades, soft computing methods and machine learning models have been successfully used and developed for modeling different areas of hydro-environment systems [Kim et al., 2014; Ahmed et al., 2019; Zounemat-Kermani et al., 2019; Bui et al., 2020; Fadaee et al., 2020; Shiri et al., 2021].

By using artificial neural network (ANN) technique, Najah et al. [2009] investigated and predicted the water qualitative variables in Johor River (Malaysia). They developed 6 architectures for neural networks, in such a way that the ANN model was used in the simulation and prediction of the parameters of total dissolved solids, electrical conductivity and turbidity, in two main stream and tributary positions. Due to the low prediction error, the outcomes of the aforementioned research proved the reliability of the model which were used in estimating the aforementioned parameters.

Sighn et al. [2009] demonstrated the capability and power of SNNs in modeling dissolved oxygen (DO) and biochemical oxygen demand (BOD) by using data gathered monthly over a 10-year period in Gomti River (India). This study indicated that optimal networks would be able to control and capture the observed long-term trends for the DO and BOD qualitative variables in time and space. Najah et al. [2014] compared the ability of the ANFIS model to predict the amount of dissolved oxygen in the Johor River basin with the MLP network and compared the capability of the ANFIS model to predict the amount of dissolved oxygen in the Johor River basin with the MLP network. For this purpose, four parameters of temperature, pH, nitrate concentration and ammonia nitrogen concentration were adopted in order to create the input compounds to modeling.

Sarkar and Pandey [2015] implemented artificial neural network (ANN) to estimate the dissolved oxygen (DO) concentrations for Mathura city, located in India. Datasets in monthly intervals including flow discharge, pH, biochemical oxygen demand (BOD), water temperature, and DO were gathered for doing the analysis. The predicted values obtained from the ANN for DO concertation, showed high level of accuracy (Pearson's correlation coefficient > 0.9) between the measured and predicted parameters. Raheli et al. [2017] predicted dissolved oxygen and BOD parameters in Langat River (Malaysia) through various models including perceptron lattice (MLP) and MLP model integrated with the glow worm metaheuristic algorithm. The results demonstrated that hybrid model was more efficient and accurate in estimating the qualitative variables of the river water by involving an optimizer.

Haghibi et al. [2018] investigated the performance of some soft computing techniques including neural networks, group method of data handling (GMDH), and SVR for the prediction of water quality indices in rivers. They claimed that the results ANN and SVR were suitable for predicting the water quality indices. Li et al. [2019] implemented a hybrid machine learning methodology embedding the metaheuristic firefly algorithm (FA) with the support vector regression (SVR) with for modeling water quality indicator prediction. The outcomes of the study showed that the SVR–FA model acted appropriately and provided promising results for the prediction of water quality index (WQI). Lu and Ma [2020] applied two hybrid tree-based soft computing models (extreme gradient boosting (XGBoost) and random forest (RF)) to predict the water quality in the Tualatin River, China. It was reported that the RF performed better than the other applied models in terms of the predicted values of DO, water temperature, and specific conductance. Moreover, stability analysis showed that the prediction stability of RF and XGBoost is higher than other benchmark models.

Varol [2020] scrutinized and assessed the effect of several stressors (such as agricultural runoff and untreated domestic sewage) on the water quality of Sürgü Stream (Turkey) with multivariate statistical techniques (MSTs) and water quality index. The majority of the studied qualitative parameters indicated significant spatial changes owing to the anthropogenic activities.

Pham et al. [2021] predicted WQI for the quality of water in wetlands using three artificial intelligence models (adaptive neuro-fuzzy system (ANFIS), ANNs, and GMDH). The results indicated that the ANFIS with (NSE = 0.9634 & MAE = 0.0219) had better performance to predict the WQI. Leong et al. [2021] applied the SVM machine learning model for predicting BOD and COD, as two WQI indices. They found that the SVM acted better than the traditional mathematical models.

By using the calculation of reflectance in remote sensing and the synchronous measurement of dissolved oxygen levels and water temperature in water bodies from 22 degrees north latitude to 45 degrees north latitude, Guo et al. [2021] developed and validated support vector regression (SVR) models and examined the effects of five climatic factors on the long-term behavior of dissolved oxygen. The results indicated the capability and generalizability of the SVR models developed in this study as well as better performance of these models in estimating dissolved oxygen by random forest methods and multiple linear regression.

Yu et al. [2022] presented a new method including decomposition of water quality data into a number of subseries by wavelet transform method, recombined by fuzzy C-means clustering and prediction (prediction) with the bidirectional gated recurrent unit method. The proposed model was assessed by qualitative data (including dissolved oxygen variable) from Poyang Lake (located in China) which indicated high accuracy in forecasting data.

Using the variables of temperature and flow rate, Dehghani et al. [2022] predicted the amount of dissolved oxygen (DO) in Cumberland River (located in the United States). In the present study, time series were monthly. Support vector regression (SVR) was responsible for modelling by itself and in combination with CSO, SSD, BWO and AIG algorithms. The four hybrid models performed better than the single model since they increased accuracy of estimation from 1.75% to 6.52%.

These studies highlight the desirable capability of data-based methods in estimating and predicting the quality variables of surface water. It can be said that by relying on the capacity of these methods, direct measurement of quality indicators can be reduced and the level of planning and quality management in natural flows can be improved. In other words, artificial intelligence models, owing to understanding the relationships governing the processes in water bodies (without the need for basic equations), have great accuracy and power in assessing and estimating water quality conditions and are considered as an effectual tool in determining the parameters of river water quality. Conversely, successful and frequent implementation of ensemble techniques including resampling methods (such as bagging and boosting), averaging and stacking has been reported for simulating and predicting the defined goals in diverse fields of hydrology [Zounemat-Kermani et al., 2021a].

All of the aforementioned issues encouraged the authors of the present study to use new modeling tools in the field of water quality in order to use artificial intelligence models in a more innovative way. The comprehensive explanation is that in the present inquiry, the amount of dissolved oxygen in the river was predicted by two groups and, subsequently, their aggregation was conducted. The first group was network-based including ANFIS-PSO and OS-ELM, and the second group was a regression tree, consisting of two models including Bagging-RF and Boosting CART. So far, no such comparison has been conducted in the field of water quality. By stacking four models, as well as ensuring innovation in the methodology used, a more accurate assessment of the combinability of the models was provided. Stacking the models was implemented using two algorithms of averaging and MLP neural network, so that the analysis of this process and its effect on the predictive power of dissolved oxygen were completed. Such a comprehensive structure was adopted for the first time in applying methods and comparing the power of models individually, in groups and collectively.

MATERIALS AND METHODS

Study area and data

The data of this study were gathered on a daily basis from January 1, 2016 to December 31, 2018 (including 1,096 recorded values of each variable) from the United States Geological Survey [USGS, 2022]. Figure 1 represents the geographical location of measuring the data of the present study in Allen, Indiana (U.S), with the following features: (Hydrologic Unit Code 04100005, Latitude 41°10'59.1", Longitude 84°52'10.9" NAD83, Drainage area 12.36 square miles, Gage datum 723.46 feet above NAVD88).

In Table 1, a summary of the statistical status of the studied parameters is available. From among the introduced parameters, DO (dissolved oxygen) is a target variable that along with other parameters, makes it possible to qualitatively model the river water; it can be said that predicting and estimating DO concentration are the responses to the interaction between the qualitative variables in the river flow field. In fact, analyzing the qualitative parameter of dissolved oxygen along with other qualitative parameters (Cl, OP, $NO_3 + NO_2$, SSC, P, T, SC, pH and $NH_3 + NH_4$), as well as the flow rate (Q) form the problem structure of this study.

As expected, despite the quantitative changes of water over time (Figure 2) and its effect on the

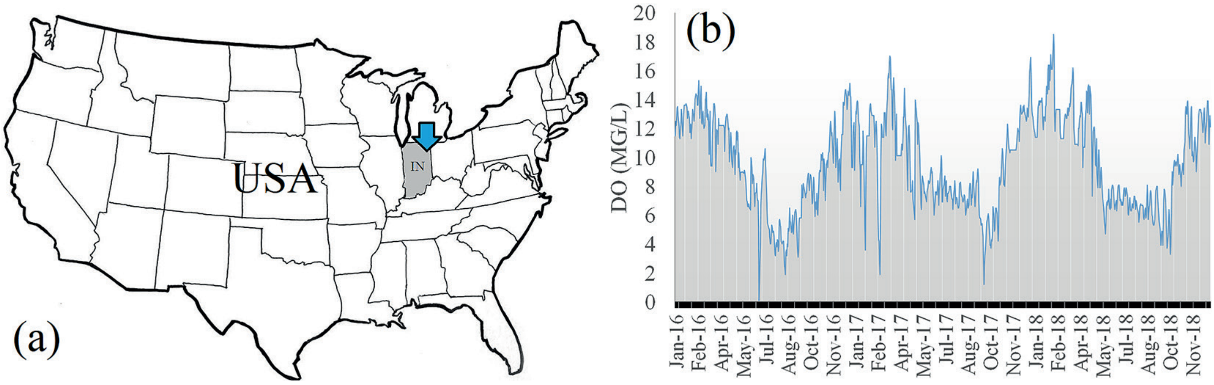

Figure 1. (a) Location of the study area in Allen County, Indiana; (b) variations of the dissolved oxygen at the study site

Table 1. Summary of the descriptive statistics of the gathered data in this study

Variable	Symbol	Mean	StDev	CoefVar	Minimum	Maximum	Skewness	Kurtosis
Discharge (m^3/s)	Q	0.339	0.895	263.81	0.001	9.798	5.66	40.52
Chloride* (ton/d)	Cl	0.742	1.335	179.99	0.000	14.900	4.08	23.08
Orthophosphate* (kg/d)	OP	7.412	27.768	374.63	0.005	344.730	6.19	46.90
Nitrate plus nitrite*, (kg/d)	NO_3+NO_2	148.30	357.40	241.05	0.000	3374.700	4.70	27.92
Suspended sediment concentration (mg/L)	SSC	39.250	112.82	287.44	2.000	1930.000	8.08	93.96
Phosphorus (mg/L)	P	0.218	0.228	104.47	0.035	2.100	3.89	21.93
Temperature, water, (Celsius)	T	12.211	8.229	67.39	0.000	26.500	0.02	-1.46
Specific conductance (µS/cm)	SC	721.70	133.04	18.43	292.0	1120.000	-0.29	-0.23
pH	pH	7.996	0.218	2.72	7.300	8.500	-0.75	0.53
Ammonia* (kg/d)	NH_3+NH_4	6.882	24.94	362.41	0.000	264.444	5.89	41.91
Dissolved oxygen (mg/L)**	DO	9.565	3.226	33.73	0.100	18.500	-0.02	-0.78

Note: * measured in dissolved water; ** the target (dependent) parameter.

self-purification potential of the river, remarkable changes in water quality are observed in low water and high-water months, in such a way that in July to September, the rate of DO was lower than other months. This is in line with the changes of flow rate in those months (Figure 3) and refers to the relationship between qualitative variables and flow in a series of time. So, in order to avoid the effect of the corresponding temporal effects of the data used (Table 1) in the models and to prevent a time trend from entering the process of predicting dissolved oxygen concentration, the chronological order of all data is disordered randomly. 75% of the beginning of the new series obtained from the data is intended for the model learning course and the final 25% is intended for testing the models.

Methodology

In this section, the methods used in order to predict the concentration of dissolved oxygen of the river in the present study are introduced from four individual models (i.e., ANFIS-PSO, OS-ELM, Bagging-RF and Boosting CART) and two stacking ones (i.e., SMA and MLP meta-learner).

Network-based models

ANFIS-PSO

Using adaptive neural network and fuzzy logic algorithms to design a nonlinear mapping between input and output spaces, an adaptive neural-fuzzy inference system (ANFIS) is developed. In the

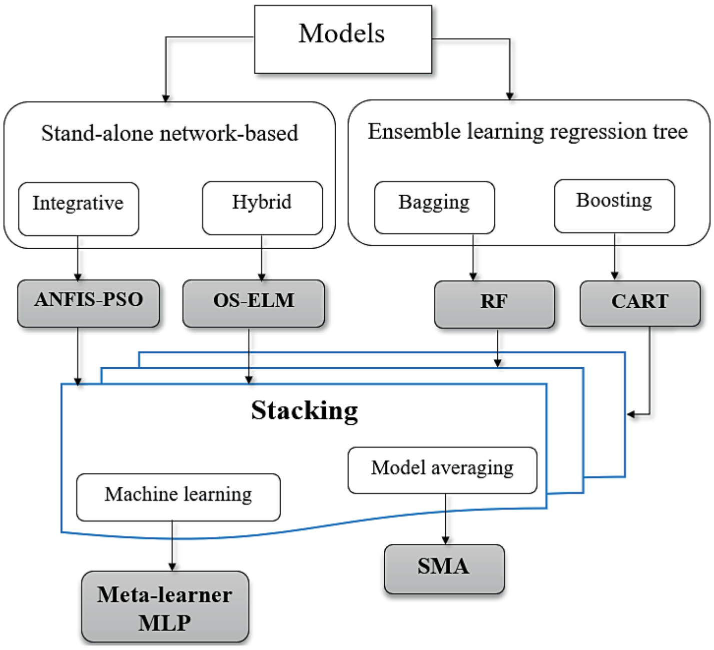

Figure 2. Variations of the discharge at the study site

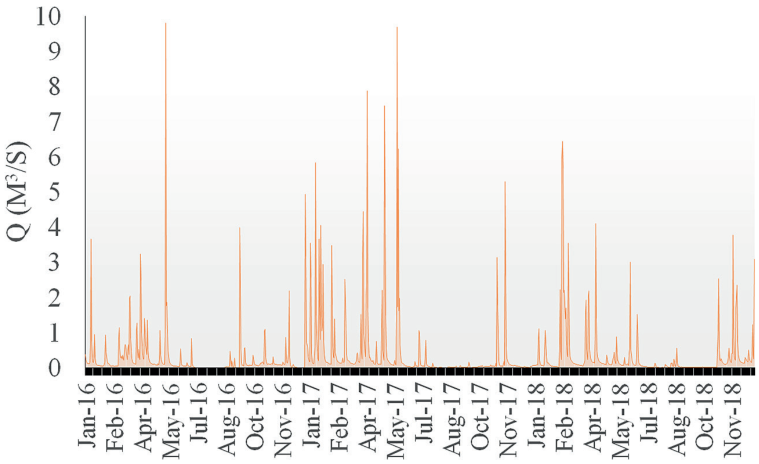

Figure 3. Schematic diagram for the methodology used in this study

learning phase, the input values are more similar to the actual values by modifying the parameters of membership degree according to the acceptable error rate. The major learning method in this system is the back propagation method under the least squares error algorithm, which corrects the parameters by returning the error value to the inputs. In order to achieve the desired framework of the neural-fuzzy system, it is essential to fit this system to the rules and functions of membership [Jang, 1993]. Hence, in the present article, the particle swarm optimization (PSO) algorithm is adopted to achieve the membership function and fuzzy rule extraction method in an optimal way, so that the algorithm can search for the optimal state by randomly creating solutions. In this study, initial population of particles is equal to 100, c1 and c2 acceleration parameters for the search space of -10 to +10 in each repetition are equal to 1 and 2, the best membership function is of the Gaussian type and the subtractive clustering method is obtained as the major partitioning technique.

OS-ELM

The extreme learning machine is a single-hidden layer feed forward neural network which determines input weights randomly and output weights analytically, except that it does not use bias for the output neuron. The ELM model decreases the network learning time remarkably by using different algorithms in calculating weights and biases. Moreover, by applying a set of weighted input signals to the network, activity functions allow for achieving a response [Huang et al., 2006]. The online sequential extreme learning machine can be trained with individual data or blocks of them in a significantly variable or fixed size. This model adopts additive hidden nodes and radial basis function (RBF) in a unified framework [Liang et al., 2006; Zounemat-Kermani et al., 2021b]. The present study has a sigmoid activating function for the additive node so as to allow for the output matrix calculation of the hidden layer in the sequential learning algorithm.

Tree-based (regression) models

Bagging-RF

In order to create a regression tree, reversal partitioning and multiple regressions are used. The decision process is repeated in each internal node from the root node, according to the tree rule, until the termination condition is satisfied. Each final node is connected to a simple regression model. At the end of the tree calling process, pruning is used to improve the generalization capacity of the trees by reducing the complexity of the structure. In order to avoid the accordance of various regression trees, the Bagging-RF model reduces the diversity of trees by creating diverse subsets of training data, which is referred to as bagging. Bagging is performed through random sampling of the main data set with replacement. Hence, some data may be used more than once in learning branches, while ineffective data may be excluded from modeling. This makes the model more stable and reliable in the face of minor changes in input data and enhances its prediction accuracy [Breiman, 2001]. In the present study, the sample size, maximum number of nodes, maximum tree depth and minimum child node size are calculated as 1, 10000, 10 and 5, respectively.

Boosting CART

The regression and classification tree model (CART) is in the form of a binary order tree that divides the problem space into segment parts [Fürnkranz et al., 2012]. This method creates its branches in a binary way and based on only one independent variable, in such a way that the information in the node is divided into two parts, based on the condition defined in each node. In the Boosting CART model, several new learners are generated from CART regression tree, which creates a more powerful algorithm by learning with previous learners. In this inquiry, maximum tree depth, number of component models for boosting and maximum surrogate in the pruning method are calculated as 5, 10 and 5, respectively. The Gini index is the impurity measure of decomposition and averaging is considered as the combining rule.

Ensemble models

SMA model (stacking)

The simple moving average (SMA) model predicts target values by averaging the available data. In stacking mode, this model considers the average of the values which are figured by the individual models at a given time as the target value at that time [Zounemat-Kermani et al., 2021a].

Meta-learner MLP model (stacking)

The multilayer perceptron neural network (MLP) is created based on a computational unit called perceptron. A perceptron takes a vector of inputs with actual values and calculates a linear combination of these inputs. In this method, calculations are performed from the input of the network to its output and, afterwards, the obtained error values are released into the previous layers in order to make the completion of the learning process possible [Barzegar and Asghari Moghaddam, 2016]. In the stacking mode, by connecting the output of individual models and defining them as input to the MLP neural network, the structure of a powerful meta-learner model is established. In the present study, the sigmoid activation function is used in the middle layer and the linear function in adopted in the output layer and the Levenberg-Marquardt optimization algorithm.

EVALUATION CRITERIA

Comparing the efficiency of the models and interpreting their abilities needs the use of error measurement criteria. Concerning this issue, as well as allowing visual comparisons with subject profile and marginal-scatter diagrams (Figures 5 and 6), quantitative metrics in Table 3 help increase accuracy in analyzing the modelling process.

In this research, the root mean square error (RMSE), mean absolute error (MAE), geometric reliability index (GRI), Pearson's correlation coefficient (r) and mean bias error (MBE) were used in order to scrutinize the results. RMSE, MAE and MBE were obtained based on the deviation of the predicted values from the observed values. Therefore, the lower the value, the more powerful the model would be, while GRI and r creates such a condition by approaching to 1. RMSE and MBE are two statistical measurements that have been widely used in environmental estimation models [Jacovides and Kontoyiannis, 1995]. Also, relative error measurements have a good level of reliability for analyzing positive data such as the values which are reported from the concentration of a variable [Jachner et al., 2007]. RMSE does not differentiate between over-estimation and under-estimation, while positive and negative MBE denote the model's tendency to over-predicted and under-predicted, respectively [Jacovides and Kontoyiannis, 1995]. GRI can also be considered as an exact simulation as a multiplicative factor in observational values, by virtue of which the corresponding predicted values are available [Jachner et al., 2007].

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(DO_m)_i-(DO_c)_i]^2} \tag{1}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|(DO_m)_i-(DO_c)_i| \tag{2}$$

$$\mathrm{GRI} = \frac{1+\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\frac{(DO_c)_i-(DO_m)_i}{(DO_c)_i+(DO_m)_i}\right]^2}}{1-\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\frac{(DO_c)_i-(DO_m)_i}{(DO_c)_i+(DO_m)_i}\right]^2}} \tag{3}$$

$$\mathrm{r} = \frac{\sum_{i=1}^{N}\begin{bmatrix}(DO_c)_i\\-(\overline{DO_c})\end{bmatrix}\begin{bmatrix}(DO_m)_i\\-(\overline{DO_m})\end{bmatrix}}{\sqrt{\sum_{i=1}^{N}\begin{bmatrix}(DO_c)_i\\-(\overline{DO_c})\end{bmatrix}^2\sum_{i=1}^{N}\begin{bmatrix}(DO_m)_i\\-(\overline{DO_m})\end{bmatrix}^2}} \tag{4}$$

$$\mathrm{MBE} = \frac{1}{N}\sum_{i=1}^{N}[(DO_c)_i-(DO_m)_i] \tag{5}$$

In Equations 1 to 5, DO_m and $\overline{DO}_m$ the concentration of the measured dissolved oxygen and its mean respectively, DO_c and $\overline{DO}_c$ the dissolved oxygen concentration calculated by the model and its mean respectively, and N are the number of actual and predicted data pairs. Based on the aforementioned equations, the difference criteria (RMSE, MAE and MBE) are expressed based on the data unit used and the relative criteria (GRI and r) are expressed without units.

FEATURE SELECTION PROCEDURE

In this research two methods of Pareto optimization and best subset selection methods have been applied for constructing the best input combination.

Unsupervised feature selection using Pareto optimization

Variables Q, P, T, SC and pH are the selected parameters; the degree of their effectiveness at the significance level of 15% is represented in Figure 4 by Pareto method. In this figure, the reference line with the standardized effect of 1.44 shows the minimum value for a significant relationship between the input parameter and the output variable

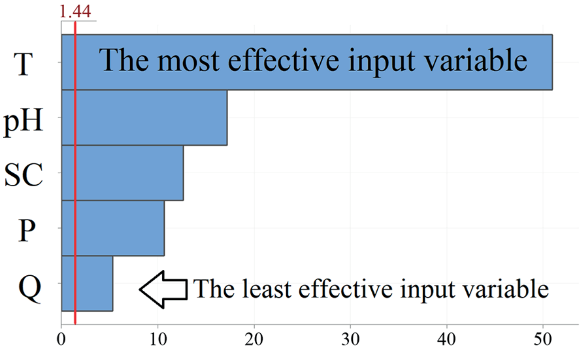

Figure 4 Pareto chart of the standardized effects for the DO as the response parameter based on the regression analysis (α=0.15)

of the model, in a way that from among the chosen variables, temperature (T) and flow rate (Q) have the most and least effects on DO, respectively.

Best subset selection method

By taking into account a series of daily input parameters including 10 variables (Q, Cl, OP, NO_3 + NO_2, SSC, P, T, SC, pH and NH_3 + NH_4), in this paper, we tried to evaluate the model of the output calculation (DO) on the same day. This requires the definition of statistical and analytical frameworks. In Table 2, the input variables with the maximum effect on dissolved oxygen values are selected with respect to the minimum Mallows' Cp statistic and in the highest correlation with DO (R-squared maximum); other parameters are not included in the modeling. In this section, the results are presented in the form of tables and graphs. Also, some explanations and clarifications are provided in order to provide the reader with a deeper understanding of numbers and figures.

RESULTS AND DISCUSSION

In subject profile diagrams (Figure 5), the outliers in the observational data are really obvious. The data which are used to draw these graphs were associated with the test phase. Scrutinizing them reveals that these DO values are on either side of the graph; i.e., they have the lowest and highest values. The lowest outlier is related to a dissolved oxygen concentration data, the correspondent parameters of which, i.e., Cl, OP, NO_3 + NO_2, SC and NH_3 + NH_4, are equal to 0.43 ton/d, 0.7 kg/d, 73 kg/d, 963 µS/cm and 1.18 kg/d. The maximum output is five data, the average value of which for the mentioned variables is equal to 0.19 ton/d, 0.3 kg/d, 24 kg/d, 814 µS/cm and 0.25 kg/d.

More chloride in outlier-Min compared to outlier-Max (0.43 vs. 0.19) increases the possibility of entering the agricultural runoff and municipal and industrial effluents to the river on the day of gathering the data of outlier-Min. The persistence of chloride in water can indicate such an event because it leads to the absence of chloride in chemical and biological reactions in the river and the presence of this element can demonstrate the presence of water pollution to some extent. Significant increase in nitrate (NO_3), nitrite (NO_2), ammonia (NH_3) and ammonium (NH_4) in the outlier-Min is in accordance with the hypothesis of entering the wastewater to the river and it refers to the nitrogen cycle, its effect on oxidation and reduction processes as well as the amount of water-soluble oxygen. It should be mentioned that it seems logical to reduce the concentration of dissolved oxygen to 1.9 mg/L and consume it by the nitrogen compounds in the effluent. Specific conductance (SC) is the rate of electrical conduction

Table 2. The effective input variables on the target parameters (DO) based on the maximum R-squared and minimum Mallows CP parameters

Variables	R-Squared	Mallows Cp	Q	Cl	OP	NO_3+NO_2	SSC	P	T	SC	pH	NH_3+NH_4
1	69.0	664.2							X			
2	75.2	311.9							X		X	
3	78.4	136.1							X	X	X	
4	80.3	31.1						X	X	X	X	
5	80.8	4.8	X					X	X	X	X	
6	80.8	5.1	X					X	X	X	X	X
7	80.8	6.4	X	X				X	X	X	X	X
8	80.8	8.0	X	X			X	X	X	X	X	X
9	80.8	9.4	X	X		X	X	X	X	X	X	X
10	80.8	11.0	X	X	X	X	X	X	X	X	X	X

Note: The five selected parameters are Q, P, T, SC, and pH.

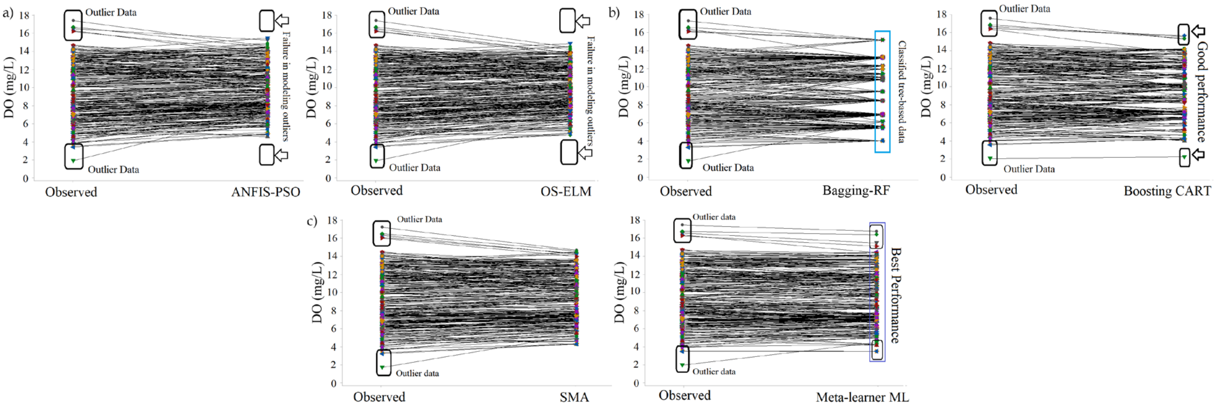

Figure 5 Subject profile plots for the observed data and the predicted results of (a): the network-based models (ANFIS-PSO & OS-ELM); (b): the regression-tree models (Bagging-RF and Boosting CART); (C): the ensemble stacking models (SMA and Meta-learner MLP)

through water-soluble salts. The SC is higher in the outlier-Min than in the outlier-Max, which is in accordance with the higher amount of chlorine ions in this case. Moreover, specific conductivity is directly related to total dissolved solids. These total dissolved solids contain organic matter and nutrients, as well as metals.

It should be noted that the outlier mentioned in this article means that the actual DO value is further away from the normal and expected data range, which may have been owing to an external factor (such as pollution) and a change in normal conditions in the water body. This approach introduces two approaches in order to address subject profile diagrams. To put it another way, we can divide these diagrams into two main parts: the major part of them consists of normal data and the minor and most important part contains outlier data. The importance of this view is reflected in tracking pollution on the days when the DO is severely reduced and, in fact, abnormal fluctuations happen in the concentration of dissolved oxygen. Comparing minimum outlier with the maximum outlier is performed, so that the data are of the same type; i.e., the river quality conditions are not normal and the data indicate the days that show a significant increase or decrease in the intensity of the effects of external factors. In completing this view, the maximum outlier indicates the days when the least pollution entered the river and its amount was less than the self-purification capacity of the river.

Also, in such a qualitative approach, the normal range of DO concentration can be interpreted as the equilibrium condition between the amount of pollutants and the self-purification capacity of the water body. This represents the value of subject profile diagrams (not yet seen in similar studies) that are highly in line with the RMSE and MAE error criteria in the test phase (Table 3).

Table 3. Train and test results of DO

Category	Type	Model	Phase	RMSE (mg/L)	MAE (mg/L)	GRI	r	MBE (mg/L)
Stand-alone-Integrative	Network-based	ANFIS-PSO	Training	1.247	0.881	1.189	0.925	-0.054
			Testing	1.284	0.989	1.180	0.923	0.413
Stand-alone- Hybrid	Network-based	OS-ELM	Training	1.186	0.849	1.172	0.932	-0.000
			Testing	1.333	1.048	1.187	0.927	0.514
Ensemble-Bagging	Regression Tree	Bagging-RF	Training	1.369	0.993	1.202	0.909	0.136
			Testing	1.242	0.919	1.162	0.917	0.019
Ensemble-Boosting	Regression Tree	Boosting CART	Training	1.176	0.865	1.183	0.934	0.072
			Testing	1.059	0.834	1.128	0.940	-0.060
Ensemble-Stacking	Model averaging	SMA	Training	1.101	0.771	1.168	0.942	0.039
			Testing	1.058	0.807	1.144	0.945	0.221
Ensemble-Stacking	Meta-learner	ANN (MLP)	Training	1.047	0.733	1.171	0.948	0.004
			Testing	0.965	0.698	1.132	0.950	-0.063

According to Figures 5b and 5c, the Meta-learner MLP model has the best performance and Boosting CART method was more powerful in predicting DO for normal and outlier data. Also, the SMA technique, despite having RMSE = 1.058 mg/L and MAE = 0.807 mg/L, did not act satisfactorily in estimating outliers. Actually, increasing horizontal lines in subject profile diagrams were associated with enhancing the performance of the models. ANFIS-PSO and OS-ELM models, in addition to having weakness in determining the amount of outlier data, did not have a good performance with RMSE = 1.284 mg/L and RMSE = 1.333 mg/L, respectively. Figure 5b shows that the Bagging-RF technique, unlike the other tree algorithm (Boosting CART), presents a state of classification in the results that accumulated and increased the error.

According to Table 3, the highest bias in the test phase can be observed in the network-based models, where the deviation of the computational values from the 1 : 1 line and the tendency to overestimating (placing the maximum points of the graph above the 1 : 1 line) are quite clear in Figure 6a. The symmetry of the points relative to the 1 : 1 line (Bagging-RF model, Figure 6b) causes the MBE values to approach 0. However, the centralization of the points on this line (Meta-learner MLP in Figure 6c; Boosting CART in Figure 6b) as well as satisfying the insignificance of MBE (MBE = -0.06) reveals higher effectiveness of the model. This is in line with the high conformity of the box plots drawn for the Boosting CART and the Meta-learner MLP to the box plot of observational data. It is interesting to note that the Boosting CART and Meta-learner MLP methods were mostly in line with the lower branch (the connecting line between the lower whisker and Q1) and the upper branch (the connection between Q3 and the higher whisker) of box plot of the actual data, respectively. In the test phase of the superior models (Meta-learner MLP and Boosting CART), the criterion r indicated high correlation between the estimated and actual values (r = 0.950 and r = 0.940) and the GRI criterion showed the highest geometric similarity (GRI = 1.132 and GRI = 1.128).

According to Table 3, the OS-ELM and ANFIS-PSO methods (with the highest RMSE and MAE in the test phase) were the weakest models. These models had negative MBE in training phase and underestimated the desired items. However, in the test phase, only the most accurate models (Meta-learner MLP and Boosting CART) had negative mean bias error values. It is likely that in the data related to the dissolved oxygen concentration, there was a tendency to decrease due to natural refining and oxygen consumption. Therefore, even the best models were in a state of underestimation during the training

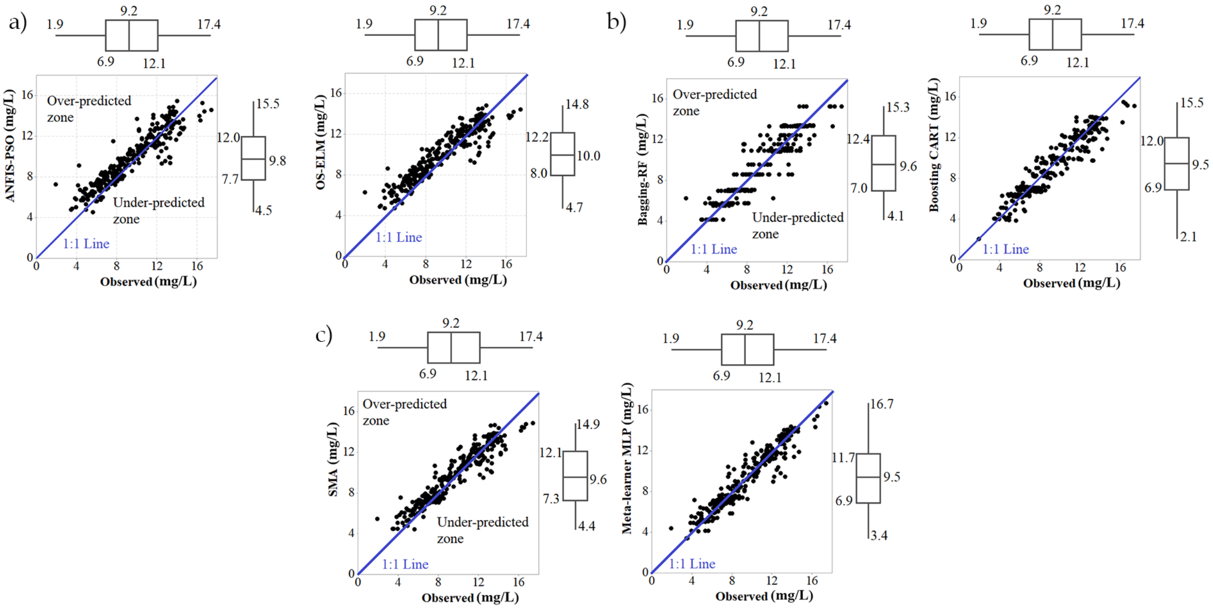

Figure 6. Marginal-scatter plots for the predicted results of (a): the network-based models (ANFIS-PSO & OS-ELM); (b): the regression-tree models (Bagging-RF & Boosting CART); (c): the ensemble stacking models (SMA & Meta-learner MLP); the numbers on the box-plots illustrate the lower whisker, Q1, median, Q3, and higher whisker values

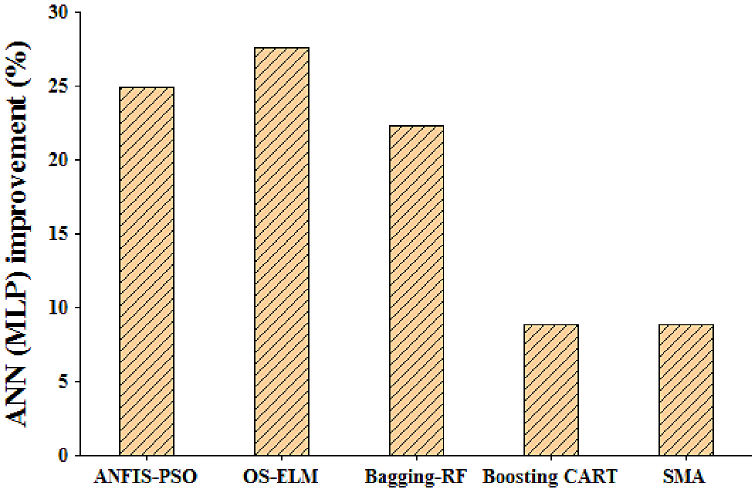

Figure 7. Meta-learner MLP improvement to other models (%) based on RMSE criterion in the testing status

Table 4. Improvement percentage of RMSE criterion in the testing results of Meta-learner MLP model in analogy to other methods

Model	RMSE (mg/L)	Ratio ANN (MLP) Error to other models (%)	ANN (MLP) improvement to other models (%)
ANFIS-PSO	1.284	75.10	24.90
OS-ELM	1.333	72.38	27.62
Bagging-RF	1.242	77.67	22.33
Boosting CART	1.059	91.12	8.88
SMA	1.058	91.18	8.82
ANN (MLP)	0.965	-	-

and testing phases. Finally, in the present study, the highest conformity was obtained between the performance of the models in DO prediction with the RMSE criterion, revealing the high validity of this measurement in modeling measurement. Hence, the relative improvement in the performance of the best model compared to other models in the test phase is represented in Figure 7 and Table 4.

CONCLUSIONS

In arranging test models in the order of desirability, the RMSE criterion can be relied upon (Meta-learner MLP is the most desirable model and OS-ELM is the least accurate method). At the same time, by analyzing the distribution diagrams and subject profile, it is possible to make a more accurate judgment and, according to the outlier examination, it was found that Boosting CART had much more predictive power than SMA. However, these models had almost the same RMSE. In the following, the superiority of the models was discussed based on the category and type. Network-based models had less accuracy than tree methods, which could be owing to the type of tree algorithm learning. It should be noted that there was little error in the training phase for network-based models, but the effect of the type of learning led to the superiority of regression tree models in the test phase. Stacking models also intensified the ensemble effect to the point that the simplest ensemble-stacking model had better performance than network-based and Bagging-RF models. Also, stacking performed under neural network experienced improved performance in modelling, which indicated not only the importance of ensemble modeling, but also the validity of combining models by stacking them.

Finally, it is suggested that in future studies, stacking models should be equipped with an optimization algorithm and the results obtained in two modes should be compared and evaluated, so as to predict the parameters of water quality. For instance, the MLP model should be integrated with particle swarm optimization (PSO) and gray wolf optimizer (GWO) algorithms.

Acknowledgments

This Project was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University Jeddah, under grant no. (G: 644-829-1443). The authors, therefore, acknowledge with thanks DSR for technical and financial support.

REFERENCES

1. Abazi A.S., Gashi B., Spahiu M.H., Bytyçi P., Dreshaj A. 2022. Analysis of the impact of ferronicel industrial activity on Drenica River quality. Journal of Ecological Engineering, 23(7), 312–322.
2. Ahmed A.N., Othman F.B., Afan H.A., Ibrahim R.K., Fai C.M., Hossain M.S., Ehteram M., Elshafie A. 2019. Machine learning methods for better water quality prediction. Journal of Hydrology, 578, 124084.
3. Benedini M., Tsakiris G. 2013. Water Quality Modelling for Rivers and Streams. Springer Science+Business Media Dordrecht, 290.
4. Bui D.T., Khosravi K., Tiefenbacher J., Nguyen H., Kazakis N. 2020. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Science of The Total Environment, 721, 137612.
5. Barzegar R., Asghari Moghaddam A. 2016. Combining the advantages of neural networks using the concept of committee machine in the groundwater salinity prediction. Modeling Earth Systems and Environment, 2, 26.
6. Breiman L. 2001. Random forests. Machine Learning, 45, 5–32.
7. Dehghani R., Torabi Poudeh H., Izadi Z. 2022. Dissolved oxygen concentration predictions for running waters with using hybrid machine learning techniques. Modeling Earth Systems and Environment, 8, 2599–2613.
8. Drozdov I.Y., Aleksakhin A.V., Aleksakhina Y.V., Petrusevich D.A. 2021. Mathematical models of water pollution evaluation. In IOP Conference Series: Earth and Environmental Science 684(1), 012026, IOP Publishing.
9. Fadaee M., Mahdavi-Meymand A., Zounemat-Kermani M. 2020. Seasonal short-term prediction of dissolved oxygen in rivers via nature-inspired algorithms. CLEAN–Soil, Air, Water, 48(2), 1900300.
10. Fürnkranz J., Gamberger D., Lavrač N. 2012. Foundations of Rule Learning. Springer-Verlag Berlin Heidelberg, 344.
11. Guo H., Huang J.J., Zhu X., Wang B., Tian S., Xu W., Mai Y. 2021. A generalized machine learning approach for dissolved oxygen estimation at multiple spatiotemporal scales using remote sensing. Environmental Pollution, 288, 117734.
12. Haghiabi A.H., Nasrolahi A.H., Parsaie A. 2018. Water quality prediction using machine learning methods. Water Quality Research Journal, 53(1), 3–13.
13. Huang G.-B., Zhu Q.-Y., Siew C.-K. 2006. Extreme learning machine: Theory and applications. Neurocomputing, 70(1–3), 489–501.
14. Jachner S., Gerald van den Boogaart K., Petzoldt T. 2007. Statistical methods for the qualitative assessment of dynamic models with time delay (R Package qualV). Journal of Statistical Software, 22(8), 1–30.
15. Jacovides C.P., Kontoyiannis H. 1995. Statistical procedures for the evaluation of evapotranspiration computing models. Agricultural Water Management, 27(3–4), 365–371.
16. Jang J.-S.R. 1993. ANFIS: Adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man, and Cybernetics, 23(3), 665–685.
17. Kim Y.H., Im J., Ha H.K., Choi J.K., Ha S. 2014. Machine learning approaches to coastal water quality monitoring using GOCI satellite data. GIScience & Remote Sensing, 51(2), 158–174.
18. Kisi O., Parmar K.S. 2016. Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. Journal of Hydrology, 534, 104–112.
19. 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.
20. Li J., Abdulmohsin H.A., Hasan S.S., Kaiming L., Al-Khateeb B., Ghareb M.I., Mohammed M.N. 2019. Hybrid soft computing approach for determining water quality indicator: Euphrates River. Neural Computing and Applications, 31(3), 827–837.
21. Liang N.-Y., Huang G.-B., Saratchandran P., Sundararajan N. 2006. A fast and accurate online sequential learning algorithm for feedforward networks. IEEE Transactions on Neural Networks, 17(6), 1411–1423.
22. Lu H., Ma X. 2020. Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere, 249, 126169.
23. Lusiana E.D., Mahmudi M., Hutahaean S.M., Darmawan A., Buwono N.R., Arsad S., Musa M. 2022. A multivariate technique to develop hybrid water quality index of the Bengawan Solo River, Indonesia. Journal of Ecological Engineering, 23(2), 123–131.
24. Najah A., Elshafie A., Karim O.A., Jaffar O. 2009. Prediction of Johor River water quality parameters

using artificial neural networks. European Journal of Scientific Research, 28(3), 422–435.

25. Najah A., El-Shafie A., Karim O.A., El-Shafie A.H. 2014. Performance of ANFIS versus MLP-NN dissolved oxygen prediction models in water quality monitoring. Environmental Science and Pollution Research, 21, 1658–1670.
26. Pham Q.B., Mohammadpour R., Linh N.T.T., Mohajane M., Pourjasem A., Sammen S.S., Anh D.T., Nam V.T. 2021. Application of soft computing to predict water quality in wetland. Environmental Science and Pollution Research, 28(1), 185–200.
27. 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, 503.
28. Rahutami S., Said M., Ibrahim E., Herpandi. 2022. Actual status assessment and prediction of the Musi River water quality, Palembang, South Sumatra, Indonesia. Journal of Ecological Engineering, 23(10), 68–79.
29. Sarkar A., Pandey P. 2015. River water quality modelling using artificial neural network technique. Aquatic procedia, 4, 1070–1077.
30. Schaffner M., Bader H.P., Scheidegger R. 2009. Modeling the contribution of point sources and non-point sources to Thachin River water pollution. Science of the Total Environment, 407(17), 4902–4915.
31. Shiri N., Shiri J., Yaseen Z.M., Kim S., Chung I.M., Nourani V., Zounemat-Kermani M. 2021. Development of artificial intelligence models for well groundwater quality simulation: Different modeling scenarios. Plos one, 16(5), e0251510.
32. Singh K.P., Basant A., Malik A., Jain G. 2009. Artificial neural network modeling of the river water quality—A case study. Ecological Modelling, 220(6), 888–895.
33. USGS (U.S. Geological Survey). 2022. Available at: ⟨https://nwis.waterdata.usgs.gov/in/nwis/dv/?site_no=04183038&agency_cd=USGS&amp;referred_module=sw ⟩ (accessed 26 March 2022).
34. Varol M. 2020. Use of water quality index and multivariate statistical methods for the evaluation of water quality of a stream affected by multiple stressors: A case study. Environmental Pollution, 266, 115417.
35. Yu J.-W., Kim J.-S., Li X., Jong Y.-C., Kim K.-H., Ryang G.-I. 2022. Water quality forecasting based on data decomposition, fuzzy clustering and deep learning neural network. Environmental Pollution, 303, 119136.
36. Zounemat-Kermani M., Seo Y., Kim S., Ghorbani M.A., Samadianfard S., Naghshara S., Kim N.W., Singh V.P. 2019. Can decomposition approaches always enhance soft computing models? Predicting the dissolved oxygen concentration in the St. Johns River, Florida. Applied Sciences, 9(12), 2534.
37. Zounemat-Kermani M., Batelaan O., Fadaee M., Hinkelmann R. 2021a. Ensemble machine learning paradigms in hydrology: A review. Journal of Hydrology, 598, 126266.
38. Zounemat-Kermani M., Alizamir M., Fadaee M., Sankaran Namboothiri A., Shiri J. 2021b. Online sequential extreme learning machine in river water quality (turbidity) prediction: A comparative study on different data mining approaches. Water and Environment Journal, 35(1), 335–348.