EEET ECOLOGICAL ENGINEERING & ENVIRONMENTAL TECHNOLOGY

Ecological Engineering & Environmental Technology 2024, 25(9), 13–25
https://doi.org/10.12912/27197050/189572
ISSN 2299–8993, License CC-BY 4.0

Received: 2024.05.27
Accepted: 2024.07.10
Published: 2024.08.01

Prediction of Water Quality Parameters of Tigris River in Baghdad City by Using Artificial Intelligence Methods

Noora Sadeq Jaafer[1*], Mustafa Al-Mukhtar[1]

[1] Civil Engineering Department. University of Technology, Baghdad, Iraq

* Corresponding author's e-mail: noorasadeq5@gmail.com

ABSTRACT

The purpose of this research is to assess the efficacy of five distinct artificial intelligence model techniques: AdaBoost, Gradient Boosting, Tree, Random Forest, and KNN, to estimate the water quality parameters of dissolved oxygen (DO) and biochemical oxygen demand (BOD). The performance of each model was assessed using two datasets: Al-Muthanna Bridge and Al-Aammah Bridge on the Tigris River in Baghdad City. The data was randomly divided into two categories: 70% for training and 30% for testing. Principal component analysis (PCA) was used to identify the most effective input parameters for modeling DO and BOD. The four performance criteria – coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and mean square error (MSE) – were applied in order to evaluate the models' effectiveness. It was demonstrated that the AdaBoost and Gradient Boosting models were superior for predicting DO and BOD. For DO prediction, the coefficient of determination R^2 of Gradient Boosting (AdaBoost) at Al-Muthanna Bridge and Al-Aammah Bridge were 0.994 (0.992) and 0.994 (0.991), respectively. For BOD prediction, the correlation coefficients R^2 of Gradient Boosting (AdaBoost) were 0.992 (0.982) and 0.989 (0.990), respectively. This study has shown that sophisticated machine learning techniques, such as gradient boosting and AdaBoost, are suitable for predicting water quality indices. They could also be helpful for predicting and managing the water quality parameters of different water supply systems in the future in water-related communities where artificial intelligence technology is still being thoroughly investigated.

Keywords: artificial intelligence, biochemical oxygen demand, dissolved oxygen, machine learning models, water quality parameters.

INTRODUCTION

Pollutants in water caused by human and environmental activities represent serious dangers to both the environment and human health (Biesbroek et al., 2022; Qiu et al., 2023; Zhou and Yang, 2023). As a result, the gradual rise in pollution concentrations in water produces environmental difficulties, ultimately destroying aquatic animal habitats. In 2019, water pollution led to 1.4 million deaths and about 1 billion illnesses worldwide, with low – and middle-income nations accounting for about 90% of deaths caused by pollution (Fuller et al., 2022). The dynamics of rivers, wave transport, hydrological processes, and transformation processes all have a significant impact on water pollutants that are discarded from various sources. Therefore, evaluating the features of the water quality in monitoring stations serves as the foundation for safeguarding aquatic systems because it aids in the creation of policies aimed at preventing contamination from waste discharges (Wang et al., 2019). Because human existence depends on the availability of water, surface and groundwater sources are subject to varying levels of pollution caused by various contaminants (Al-Janabi et al., 2012; Asadollah et al., 2012). Because of this, predicting water quality (WQ) has become more challenging recently, and because WQ is so important to human life, many scholars have put a lot of effort into evaluating WQ (Elbeltagi et al., 2023; Tung and Yaseen, 2020; Ding et al., 2014). Resources of water in the Iraqi region have been under a significant amount of stress for the past 20 years for a variety of reasons, including the construction of dams on the Tigris and Euphrates rivers, changes in the worldwide climate, and a decline in the local yearly rainfall and rates of precipitation (Wang et al., 2023). The recognition of surface water pollution as a problem and the growing interest in WQ assessment have led to a recent surge in the demand for reliable, accurate, flexible, and effective prediction models (Kılıç and Çetin, 2023). These models are thought to be able to adequately capture the mechanics of the WQ decrease) Montazeri et al., 2023). Because ML models are precise and dependable, researchers used them to determine the concept of surface and subsurface WQ modelling (Geshnigani et al., 2023). Researchers and scientists are interested in research that involves modelling WQ utilizing new, advanced models, and the idea of exploring new machine learning (ML) models that can solve environmental engineering challenges is always continuing (Kang et al. 2022; Liu et al. 2022) Recently developed ensemble AI algorithms, like random tree (RT), random committee (RC), and reduced error pruning tree (REPTree), which have been introduced to improve the capabilities of AI systems (Khosravi et al., 2021; Shahdad and Saber, 2022; Saha et al., 2022). Three AI models were compared by Khosravi et al. (2018): M5P, REPTree, and instance-based learning (IBK), as well as their hybridized variants, bagging-M5P, random committee-REPT, and random subspace-REPT (RS-REPT), for predicting SSL. The prediction of hourly suspended sediment was enhanced by the hybrid REPTree and RC models, according to their findings. In a different study, Chen et al. (2020) discovered that deep cascade forest (DCF), random forest, and random tree forest performed noticeably better in WQ predictions than the conventional approaches. According to Asadollahfardi et al. (2021), the patented Extra Tree Regression (ETR) model provided more precise WQI predictions all throughout the training and testing stages.

A number of recent review research publications on the advancement of machine learning for river WQ (Jamei et al., 2022; Mahdavi-Meymand et al., 2024). The literature review places a lot of emphasis on looking into new machine learning models iterations for river WQ modelling in light of the limitations of the current ML models. For example, the disadvantages of fine-tuning the internal parameters of traditional models like support vector machines (SVM), fuzzy logic (FL), and artificial neural networks (ANN) (Alavi et al., 2022).

Water pollution control requires an accurate estimate of biochemical oxygen demand (BOD) because it is a key indicator of high-quality water) Manzar et al., 2022). On the other hand, high BOD loads are bad for river water quality because they lead to low dissolved oxygen (DO) concentrations, which are unsuitable for aquatic life. Consequently, several models have been developed for forecasting of changes in water quality brought on by BOD releases (Boano et al., 2006). Analysing this parameter, especially BOD analysis, is difficult and time-consuming. BOD is a crucial indication of water pollution and gives an estimate of the quantity of organic matter that degrades naturally in the water. BOD is also recognized as the primary indicator for the health of the aquatic system, and accurate measurement of it can help establish safe and successful strategies for protecting water resources. However, BOD is noted for a minimum of five days. Since that precise WQ parameter prediction into a study field can save resources like time, money, and energy, modelling techniques are heavily considered when making these important parameter predictions (Benaafi et al., 2022). In poor countries, where financing for quality of the environment evaluation and monitoring is less than in richer countries, modelling techniques are more essential. This study is based on predicting monthly-scale DO and BOD for the Tigris River in the Iraq region. To do this, five separate ensemble machine learning models were created. The models were selected due to their widespread use, which attested to their applicability in climatological, hydrological, and environmental studies (Ramal et al., 2022, Adedeji et al., 2022).

The main aim of the present work is to evaluate the AdaBoost, Gradient Boosting, Tree, Random Forest, and KNN models that were developed to predict DO and BOD on the Tigris River in the middle of Iraq. The physical and chemical water parameters, which include PH, BOD_5, DO, PO_4, Ca, Mg, NO_3, TH, Na, CL, K, E.C, Alkalinity, SO_4, TSS, TDS, and Turbidity were used as predictors. Additionally, this study used principle component analysis to evaluate various input scenarios and determine which inputs had the greatest impact on the models' accuracy of prediction. Consequently, the purpose of the current study was to support water quality monitoring by offering useful data on the performance of these models.

CASE STUDY

The case study in this research is the reach of the Tigris River, situated within Baghdad, Iraq, as seen in Figure 1. The Tigris River is the only supply of drinkable water in the city of Baghdad (Adnan et al., 2021). Baghdad, the capital city of Iraq, is located at latitude 33° 18′ 0″ from the north and longitude 44° 24′ 0″ from the east. Tigris flows from Al-Tajee, in the north, to Al-Zafaraniah, in the south, before meeting with the Diyala River. The river separates the city into two parts: Karkh (right) and Risafa (left), flowing north to south. The climate of the region is arid to semi-arid, with hot, dry summers and cool winters; the average annual rainfall is approximately 151.8 mm (Al Obaidy et al., 2016). The Tigris River is Western Asia's second-longest river, it flows through densely inhabited areas, particularly Baghdad, which has nearly 8 million people. Demand for water is at an all-time high, but Tigris discharge has significantly decreased in recent decades. Wastewater treatment plants are facing a shortage due to the rising volumes of wastewater; in Baghdad, for example, 20 percent of the sewage is thrown into the river untreated (Oleiwi and Al-Dabbas, 2022).

METHODS

Collection data and sampling locations

The dataset for this study had been gathered and monitored monthly at two sites along the Tigris River by Iraq's Central Region's Ministry of the Environment, Department of Protection, and Improvement Environment (Tab. 1). Monthly water quality dataset collected from the 2008–2022 period consists of the physical and chemical water parameters, which include PH, DO, BOD_5, NO_3, PO_4, Ca, Mg, TH, K, Na, SO_4, CL, TDS, EC, Alkalinity, TSS, and Turbidity. These variables are used to develop the AdaBoost, Gradient Boosting, Tree, Random Forest, and KNN models to estimate the dissolved oxygen and biochemical oxygen demand characteristics of water quality. Since DO and BOD have been the two most widely used WQ parameters for many years, this research focused on their prediction since precise prediction of these parameters is critical to the effectiveness of preventive measures initiatives. Table 2 presents statistical characteristics for the WQ parameters. In this paper, the total quality of water dataset for Al-Muthanna Bridge (147 samples) and Al-Aammah Bridge (139 samples) was randomly split into two groups: training and testing. The training and testing datasets comprised 70% and 30% of the samples, respectively.

Applied ensemble machine learning models

Random forest (RF) model

RF is an ensemble learning method that can be applied to regression, classification, and other tasks. Tin Kam Ho introduced it initially, and Leo

Figure 1. The study area map for the station along the Tigris River in Baghdad city

Table 1. Location and coordinates of the sampling sites

Site number	Location	Longitude (E)	Latitude (N)
1	Al-Muthanna Bridge	44°20’45.5”	33°25’43.7”
2	Al-Aammah Bridge	44°21’21.9”	33°22’29.5”

Table 2. Illustrate the statistical measurements of water quality in Baghdad City on the Tigris River

Bridge	No.	Parameters	Unit	Mean	Mode	Median	Dispersion	Min.	Max.
Al-Muthanna Bridge	1	EC	μs/cm	934.6949	863.00	923.00	0.1700	567.00	1346.00
	2	TSS	mg/L	92.10569	228.750	60.100	1.27578	1.000	1140.000
	3	TDS	mg/L	577.74	640	569	0.18	386	875
	4	Alkalinity	mg/L	149.80842	136.000	145.000	0.40554	90.000	834.000
	5	TH	mg/L	319.1324	280.00	311.00	0.2162	156.00	567.00
	6	SO_4	mg/L	205.4383	200.00	200.00	0.3085	78.00	385.00
	7	Turbidity	NTU	43.294046	48.3197	28.1967	0.878902	1.3000	190.0000
	8	Cl	mg/L	83.27478	106.000	82.000	0.28493	38.000	184.000
	9	Ca	mg/L	74.36931	64.000	73.000	0.19931	34.000	135.000
	10	Na	mg/L	57.738406	42.3150	55.0000	0.312747	3.0000	107.0000
	11	Mg	mg/L	32.74704	27.000	32.000	0.32352	9.000	80.000
	12	NO_3	mg/L	4.188352	3.1000	3.9000	0.499451	0.2600	14.2000
	13	DO	mg/L	8.572790	8.0000	8.6000	0.175743	2.1000	12.9000
	14	PH	pH Units	7.7184	8.00	7.70	0.0543	6.70	8.89
	15	K	mg/L	2.959271	2.8000	2.9000	0.314702	1.0000	8.8000
	16	BOD_5	mg/L	2.4340935	1.00000	2.00885	0.6008532	0.20000	8.30000
	17	PO_4	mg/L	0.3058318	0.30000	0.24000	1.0443461	0.00600	3.10000
Al-Aammah Bridge	1	EC	μs/cm	960.32754	840.000	915.000	0.28478	12.200	3020.000
	2	TDS	mg/L	591.7803	570.00	568.00	0.2897	58.20	1963.00
	3	TH	mg/L	342.95312	420.000	346.000	0.26582	1.400	638.000
	4	Alkalinity	mg/L	139.20497	136.000	136.000	0.31611	67.000	570.000
	5	SO_4	mg/L	214.143666	200.0000	200.0000	0.363707	55.0000	480.0000
	6	Turbidity	NTU	51.291923	11.0211	33.2000	1.062985	1.4700	446.0000
	7	Mg	mg/L	37.19986	28.000	34.000	0.89432	8.000	332.000
	8	TSS	mg/L	72.619598	283.0000	58.0000	0.837239	5.0000	329.0000
	9	Cl	mg/L	80.04873	64.000	79.000	0.27540	42.000	196.000
	10	Na	mg/L	54.01465	50.000	51.000	0.31473	26.000	156.000
	11	Ca	mg/L	85.81058	77.000	83.000	0.25346	23.000	144.000
	12	NO_3	mg/L	4.370618	3.5000	3.6700	1.257498	0.4400	65.0000
	13	K	mg/L	3.319585	2.5000	2.9800	1.060018	1.0000	43.0000
	14	DO	mg/L	8.7831537	8.00000	8.70000	0.1952830	0.70000	12.80000
	15	BOD_5	mg/L	2.225708	1.5000	2.0000	0.585744	0.3000	9.5000
	16	PH	pH Units	7.766643	7.8000	7.8000	0.049771	6.9000	8.7700
	17	PO_4	mg/L	0.3281303	0.18000	0.21000	1.6740239	0.00300	5.30000

Breiman later improved it (Breiman, 2001). It is a useful tool for solving multi-regression and prediction problems because of its simplicity and adherence to the “divide and conquer” strategy (Chen et al., 2020). The group of decision trees is produced by Random Forest. A bootstrap sample of the training data is used to generate each tree. The word “random” refers to the arbitrary set of characteristics generated during the construction of individual trees, from which the best attribute for the split is chosen (Cutler et al., 2007). RF has been successfully applied in environmental engineering (Abbas, 2013) and other areas of research (Belgiu and Drăguţ, 2016). Random forests are a technique for averaging numerous deep decision trees trained on different regions of the same training set with the purpose of reducing variation (Hastie et al., 2009). More information on how RF models are mathematically formulated can see in Goel et al. (2017).

AdaBoost model

Yoav Freund and Robert Schapire designed the "adaptive boosting" widget as a machine-learning approach. It can be combined with other learning algorithms to improve effectiveness (Hastie et al., 2009). The classifier's correct separation of samples reduces their weight, while misclassification increases their weight. This allows the learning algorithm to focus on difficult training samples and learn them in future studies. Weighted voting merges weaker options into each round, resulting in a stronger final option (Bishop, 2006). AdaBoost works for both classification and regression.

Gradient boosting model

Gradient boosting is a machine learning approach for regression and classification problems that constructs a prediction model from a collection of weak prediction models, typically decision trees. It produces a prediction model in the form of an ensemble of weak prediction models, that is, models with very few data assumptions, often simple decision trees (Hastie et al., 2009).

Tree mode

A tree is a basic method of separating data into nodes based on class purity. It is a precursor to Random Forest. Trees can handle both categorical and numerical collections. It can also be applied to classification and regression tasks (Hastie et al., 2009).

k-nearest neighbours algorithm (k-NN) model

Regression and classification issues are both resolved by the k-NN approach. The input is always the collection of k closest training samples found in a dataset. Regression or classification with k-NN produces different results. The object's property value is obtained by k-NN regression. This is the average of the values of the k closest neighbours. If k equals one, the output is simply set to the value of the nearest neighbour (Hastie et al., 2009). A helpful method for classification and regression is to weight neighbor contributions so that closer neighbors contribute more to the average than those who are farther apart. For example, a popular weighting strategy applies a weight of 1/d to each neighbour, where d represents the distance between them (Blu et al., 2004).

Parameters selection

In this study, the most influential predictor's parameters on predictand was identified using Principal component analysis (PCA). PCA is a technique for reducing the dimensionality of such datasets while enhancing interpretability and avoiding information loss. It achieves this by creating new uncorrelated variables that gradually optimize variance) Jolliffe and Cadima, 2016). The PCA approach's mathematical technique works on the basis of allocating the least amount of error between observed and predicted values. Because of the variation in the principal component (Bhagat et al., 2020). Many studies employ the first two major components to plot data in two dimensions and visually show clusters of closely related data points (Jolliffe and Cadima, 2016).

Modelling performance criteria

The performance of AdaBoost, Gradient Boosting, Tree, Random Forest, and KNN configurations was evaluated using four error measures as explained below (Al-Mukhtar et al., 2024):

1. Coefficient of determination (R^2), this shows the degree of relationship between predicted and measured values Equation 1.
2. Root mean square error (RMSE), Which is preferable in many iterative prediction and optimization strategies Equation 2.
3. Mean absolute error (MAE), This is a metric generally understood in engineering applications Equation 3.
4. Mean square error (MSE), is the average of the squared errors or variances (the difference between the estimator and the value that is estimated) Equation 4.

$$R^2 = \frac{\sum_{i=1}^{n}(O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^{n}(O_i - \bar{O})^2 \sum_{i=1}^{n}(P_i - \bar{P})^2}} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{n}} \tag{2}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|O_i - P_i| \tag{3}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(O_i - P_i)^2 \quad (4)$$

when: N is the number of data points, O are the observed values, P are the predicted values, and the bar sign is the variable's mean.

RESULTE AND DISCUSSION

Feature selection

AI models perform much worse when there are redundant and irrelevant predictors included, and prediction models have overfitting issues as a result. Consequently, as it reduces the amount of time needed for data collection and calculation, it could be useful to extract a smaller group of predictors that includes the most relevant predictors (Bhagat et al., 2021). To increase the accuracy of the surface DO and BOD water quality prediction in the Tigris River, in this work, five PCs were combined with five distinct artificial intelligence models for ensemble learning (AdaBoost, Gradient Boosting, Tree, Random Forest, and KNN). It is important to note that the dataset span employed in this study had sufficient information to support the creation of machine learning models and the learning process. The monthly amount of the 15 years of observations in this study was sufficient to build the machine learning models. In this study, the scree plot is used to choose the number of components. The first five principal components (PCs) were extracted with eigenvalues > 1, as seen in Tables 3 and 4, for prediction of BOD in Al-Muthanna Bridge and Al-Aammah, explaining 69.55% and 71.19%, respectively, of the total variance in the water quality data set. Similarly, for DO in Al-Muthanna Bridge and Al-Aammah, the variance was explained by 68.05% and 70.36%, respectively, as shown in Table 4. Furthermore, Figures 2 and 3 show the creation of the scree plot, which depicts the majority of the variability in the data. The *x*-axis depicts the component, while the *y*-axis shows how important it is. The chart shows that after the second component, the incremental influence of each subsequent component decreases significantly.

Model performances

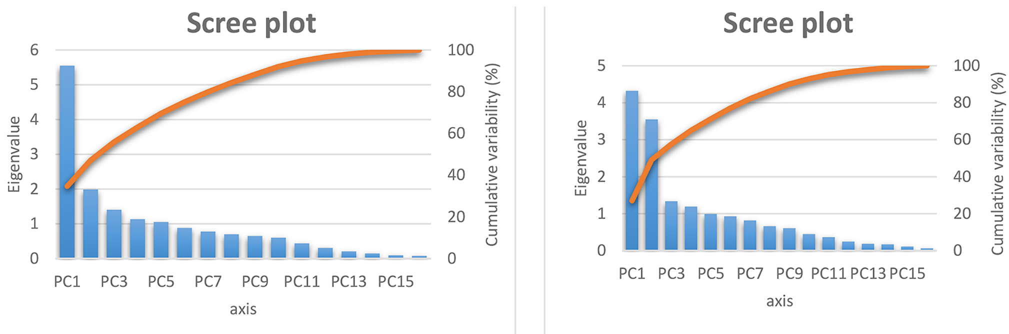
This work intends to evaluate the efficiency of AdaBoost, Gradient Boosting, Tree, Random Forest, and KNN in predicting DO and BOD concentrations in the Tigris River at Al-Muthanna Bridge and Al-Aammah Bridge. The datasets for 15 years were divided into two groups: 70% for training and 30% for testing. Table 2 summarizes the concentration of parameters for Al-Muthanna Bridge and Al-Aammah Bridge used in this study. The two evaluated parameters exhibited distinct patterns of influence with respect to the input parameters due to different sources of pollution and population variation along the river stretch. In DO prediction, during training, AdaBoost performed extremely well, followed by GB, and

Table 3. Eigen values of PCA for input parameters for predicting BOD

Principal component	Al-Muthanna Bridge			Al-Aammah Bridge		
	Eigenvalue	Variability (%)	Cumulative %	Eigenvalue	Variability (%)	Cumulative %
PC1	5.550	34.685	34.685	4.321	27.008	27.008
PC2	1.988	12.425	47.110	3.549	22.182	49.190
PC3	1.404	8.776	55.886	1.336	8.351	57.541
PC4	1.134	7.085	62.971	1.191	7.445	64.986
PC5	1.053	6.579	69.550	0.994	6.211	71.198
PC6	0.881	5.508	75.059	0.930	5.811	77.009
PC7	0.776	4.851	79.910	0.818	5.112	82.120
PC8	0.698	4.365	84.275	0.662	4.137	86.257
PC9	0.650	4.060	88.335	0.607	3.797	90.054
PC10	0.599	3.746	92.081	0.448	2.801	92.856
PC11	0.437	2.732	94.814	0.365	2.282	95.138
PC12	0.307	1.916	96.730	0.246	1.536	96.673
PC13	0.206	1.288	98.018	0.186	1.165	97.838
PC14	0.147	0.916	98.934	0.171	1.068	98.906
PC15	0.094	0.587	99.521	0.112	0.702	99.608
PC16	0.077	0.479	100.000	0.063	0.392	100.000

Table 4. Eigen Values of PCA for input parameters to prediction DO

Principal component	Al-Muthanna Bridge			Al-Aammah Bridge		
	Eigenvalue	Variability (%)	Cumulative %	Eigenvalue	Variability (%)	Cumulative %
PC1	5.489	34.308	34.308	4.357	27.232	27.232
PC2	2.003	12.516	46.824	3.415	21.344	48.576
PC3	1.329	8.308	55.133	1.308	8.176	56.752
PC4	1.135	7.096	62.228	1.157	7.234	63.986
PC5	0.932	5.827	68.055	1.020	6.375	70.361
PC6	0.883	5.520	73.576	0.954	5.964	76.325
PC7	0.860	5.373	78.948	0.862	5.389	81.714
PC8	0.743	4.644	83.592	0.719	4.492	86.206
PC9	0.682	4.262	87.854	0.596	3.724	89.930
PC10	0.623	3.896	91.750	0.445	2.782	92.712
PC11	0.429	2.682	94.431	0.371	2.320	95.031
PC12	0.370	2.315	96.746	0.259	1.619	96.650
PC13	0.204	1.278	98.024	0.190	1.186	97.836
PC14	0.146	0.915	98.939	0.172	1.073	98.909
PC15	0.093	0.581	99.520	0.112	0.702	99.611
PC16	0.077	0.480	100.000	0.062	0.389	100.000



(a) Al-Muthanna Bridge **(b) Al-Aammah Bridge**

Figure 2. Scree plot of Principal component analysis (PCA) for the input parameter used for prediction BOD (a) in Al-Muthanna Bridge(b) in Al-Aammah Bridge

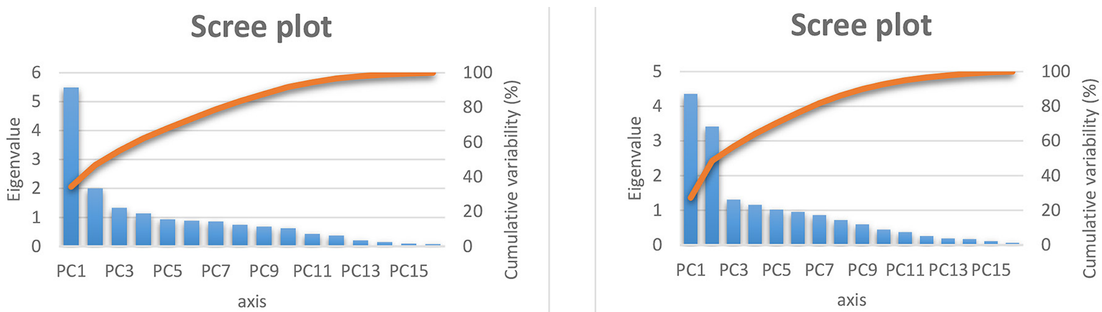


Figure 3. Scree plot of Principal component analysis (PCA) for the input parameter used for prediction DO (a) in Al-Muthanna Bridge(b) in Al-Aammah Bridge

Tree surpassed RF and the last one, KNN. However, gradient boosting was the most successful in the tests (R^2 = 0.994, MAE = 0.108, RMSE = 0.13, MSE = 0.018) in Al-Muthanna Bridge and in Al-Aammah Bridge (R^2 = 0.994, MAE = 0.092, RMSE = 0.14, MSE = 0.013). AdaBoost followed closely in performance (R^2 = 0.992, MAE = 0.047, RMSE = 0.147, MSE = 0.022) in Al-Muthanna Bridge and in Al-Aammah Bridge (R^2 = 0.991, MAE = 0.048, RMSE = 0.145, MSE = 0.021). Tree performance performed less accurately in Al-Muthanna Bridge (R^2 = 0.866, MAE = 0.432, RMSE = 0.606, MSE = 0.367) and in Al-Aammah Bridge (R^2 = 0.941, MAE = 0.280, RMSE = 0.370, MSE = 0.137). RF performance was not superior in Al-Muthanna Bridge (R^2 = 0.866, MAE = 0.432, RMSE = 0.606, MSE = 0.367) and in Al-Aammah Bridge (R^2 = 0.941, MAE = 0.280, RMSE = 0.370, MSE = 0.137). KNN performance lagged in Al-Muthanna Bridge (R^2 = 0.646, MAE = 0.808, RMSE = 0.986, MSE = 0.973) and in Al-Aammah Bridge (R^2 = 0.528, MAE = 0.861, RMSE = 1.042, MSE = 1.086). In summary, GB and AdaBoost outperformed other methods for DO predictions in both training and testing (Tab. 5), indicating that they should be used in the present study. In BOD prediction, during training, AdaBoost performed extremely well, followed by GB and Tree, while RF and KNN lagged. In testing, GB outperformed other models in Al-Muthanna Bridge (R^2 = 0.992, MAE = 0.096, RMSE = 0.119, MSE = 0.014) and in Al-Aammah Bridge (R^2 = 0.989, MAE = 0.128, RMSE = 0.152, MSE = 0.023). Surprisingly, AdaBoost became the second-best in testing in Al-Muthanna Bridge (R^2 = 0.982, MAE = 0.063, RMSE = 0.174, MSE = 0.030) and in Al-Aammah Bridge (R^2 = 0.990, MAE = 0.066, RMSE = 0.150, MSE = 0.022), followed by Tree (R^2 = 0.969, MAE = 0.177, RMSE = 0.229, MSE = 0.052 in Al-Muthanna Bridge) and (R^2 = 0.849, MAE = 0.312, RMSE = 0.572, MSE = 0.328 in Al-Aammah Bridge). While RF performance was not good with respect to the remainder models in Al-Muthanna Bridge (R^2 = 0.788, MAE = 0.474, RMSE = 0.601, MSE = 0.361) and in Al-Aammah Bridge (R^2 = 0.795, MAE = 0.391, RMSE = 0.667, MSE = 0.445), The KNN model was the least effective (R^2 = 0.511, MAE = 0.736, RMSE = 0.914, MSE = 0.835) in Al-Muthanna Bridge and in Al-Aammah Bridge (R^2 = 0.665, MAE = 0.531, RMSE = 0.852, MSE = 0.727). Table 6 shows the model's performance comparison without overlapping findings, allowing model selection to pick GB and AdaBoost as the best models, superior for the purpose of BOD prediction testing and training.

Scatter plot analysis for model outputs

The current section aims to create scatter plots according to the result that the Gradient Boosting and AdaBoost models performed effectively in the testing phase for the prediction of DO and BOD, as depicted in Figures 4 and 5. BOD prediction models performed well in testing, with GB and AdaBoost outperforming Tree, RF, and KNN in terms of peak capture. As depicted in figures, GB confirmed superiority (R^2 = 0.992) in Al-Muthanna Bridge and (R^2 = 0.989) in Al-Aammah Bridge. AdaBoost became the second-best in testing at Al-Muthanna Bridge (R^2 = 0.982) and at Al-Aammah Bridge (R^2 = 0.990). Tree closely followed with R^2 = 0.969 and 0.849 in Al-Muthanna Bridge and Al-Aammah Bridge, respectively. The RF model outperformed with R^2 = 0.788 and 0.795 at Al-Muthanna Bridge and Al-Aammah Bridge,

Table 5. Performance of the model for DO during training and testing

	Training data					Testing data				
	Model	MSE	RMSE	MAE	R2	Model	MSE	RMSE	MAE	R2
Al-Muthanna Bridge	AdaBoost	0.084	0.290	0.161	0.959	Gradient boosting	0.018	0.133	0.108	0.994
	Gradient boosting	0.084	0.290	0.231	0.958	AdaBoost	0.022	0.147	0.047	0.992
	Tree	0.213	0.462	0.294	0.895	Tree	0.367	0.606	0.432	0.866
	Random forest	0.286	0.535	0.387	0.859	Random forest	0.567	0.753	0.562	0.794
	kNN	0.579	0.761	0.554	0.714	kNN	0.973	0.986	0.808	0.646
Al-Aammah Bridge	AdaBoost	0.070	0.264	0.123	0.968	Gradient boosting	0.013	0.114	0.092	0.994
	Gradient boosting	0.152	0.390	0.294	0.930	AdaBoost	0.021	0.145	0.048	0.991
	Tree	0.233	0.483	0.304	0.892	Tree	0.137	0.370	0.280	0.941
	Random forest	0.269	0.519	0.380	0.876	Random forest	0.341	0.584	0.473	0.852
	kNN	0.650	0.806	0.564	0.700	kNN	1.086	1.042	0.861	0.528

Table 6. Performance of the model for BOD during training and testing

	Training data					Testing data				
	Model	MSE	RMSE	MAE	R2	Model	MSE	RMSE	MAE	R2
Al-Muthanna Bridge	AdaBoost	0.100	0.317	0.145	0.957	Gradient Boosting	0.014	0.119	0.096	0.992
	Tree	0.117	0.342	0.262	0.950	AdaBoost	0.030	0.174	0.063	0.982
	Gradient Boosting	0.194	0.441	0.358	0.916	Tree	0.052	0.229	0.177	0.969
	Random Forest	0.319	0.565	0.418	0.863	Random Forest	0.361	0.601	0.474	0.788
	kNN	0.888	0.943	0.704	0.618	kNN	0.835	0.914	0.736	0.511
Al-Aammah Bridge	AdaBoost	0.007	0.086	0.027	0.995	AdaBoost	0.022	0.150	0.066	0.990
	Gradient Boosting	0.073	0.271	0.221	0.950	Gradient Boosting	0.023	0.152	0.128	0.989
	Tree	0.172	0.415	0.263	0.883	Tree	0.328	0.572	0.312	0.849
	Random Forest	0.308	0.555	0.375	0.792	Random Forest	0.445	0.667	0.391	0.795

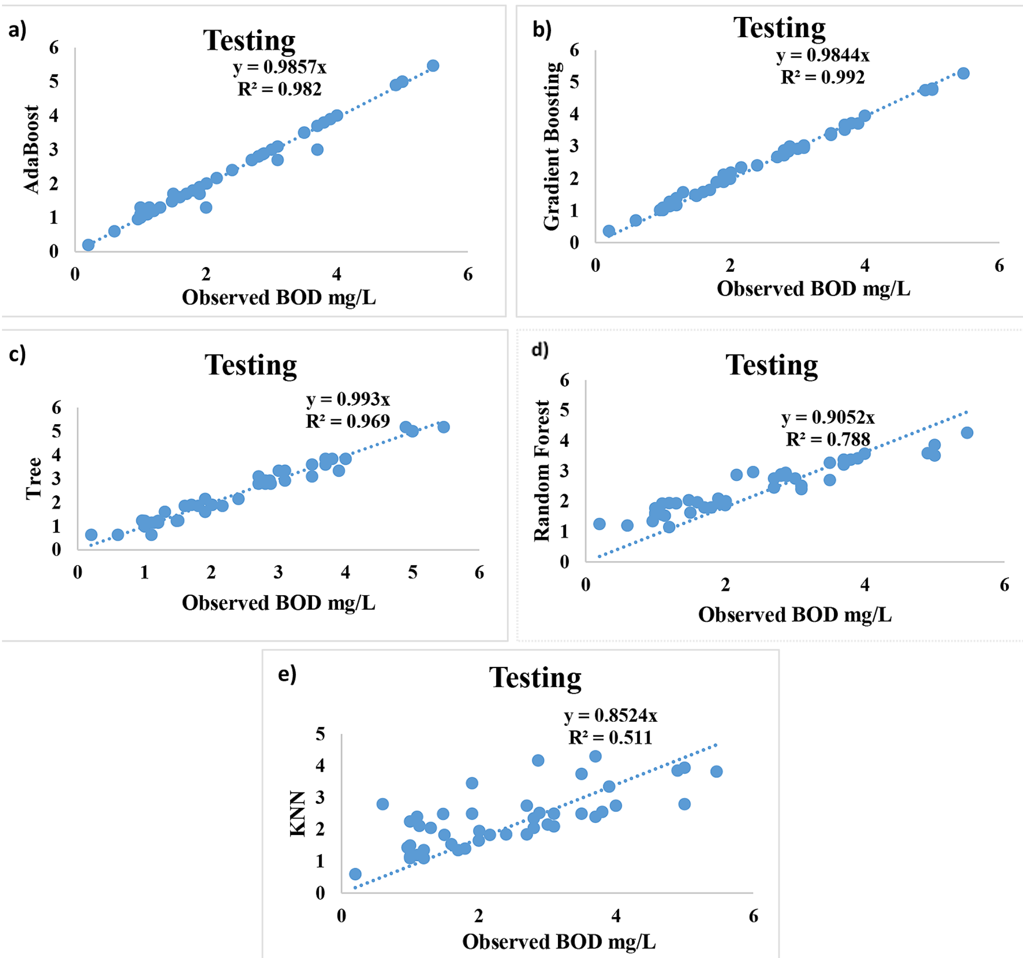


Figure 4. Scatter plot for the model predicted and observed BOD levels in Al-Muthanna Bridge(a) AdaBoost, (b) Gradient Boosting, (c)Tree, (d)Random Forest, (e) KNN

respectively. Lastly, KNN performed less accurately in Al-Muthanna Bridge and in Al-Aammah Bridge, with R^2 = 0.511 and 0.665, respectively. On the other hand, DO predictions proved the models' robustness. GB and AdaBoost excelled, followed by tree, while RF and KNN lagged. The results of R^2 values from scatter plots (Fig. 6 and Fig. 7) affirmed GB dominance in Al-Muthanna

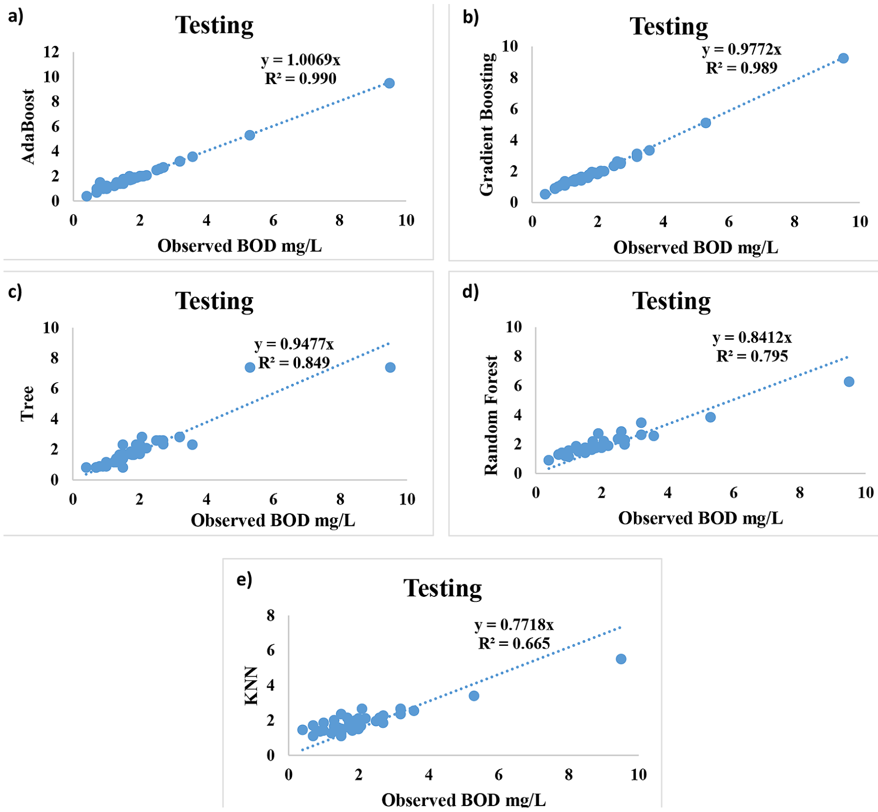


Figure 5. Scatter plot for the model predicted and observed BOD levels in Al-Aammah Bridge (a) AdaBoost, (b) Gradient Boosting, (c)Tree, (d) Random Forest, (e) KNN

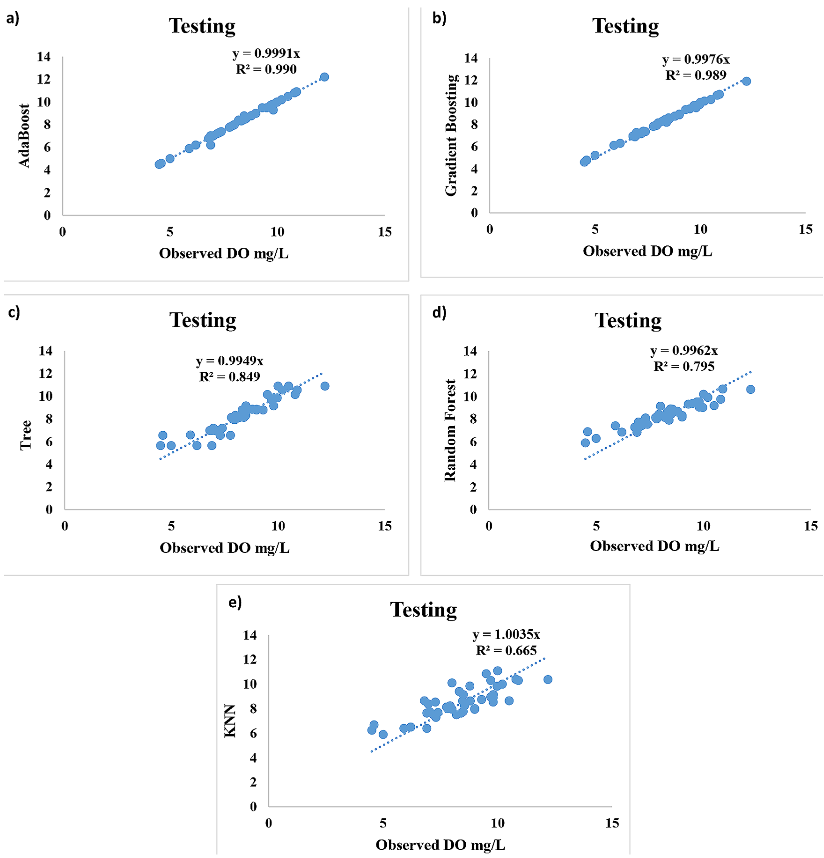


Figure 6. Scatter plot for the model predicted and observed DO levels in Al-Muthanna Bridge (a) AdaBoost, (b) Gradient boosting, (c)Tree, (d) Random forest, (e) KNN

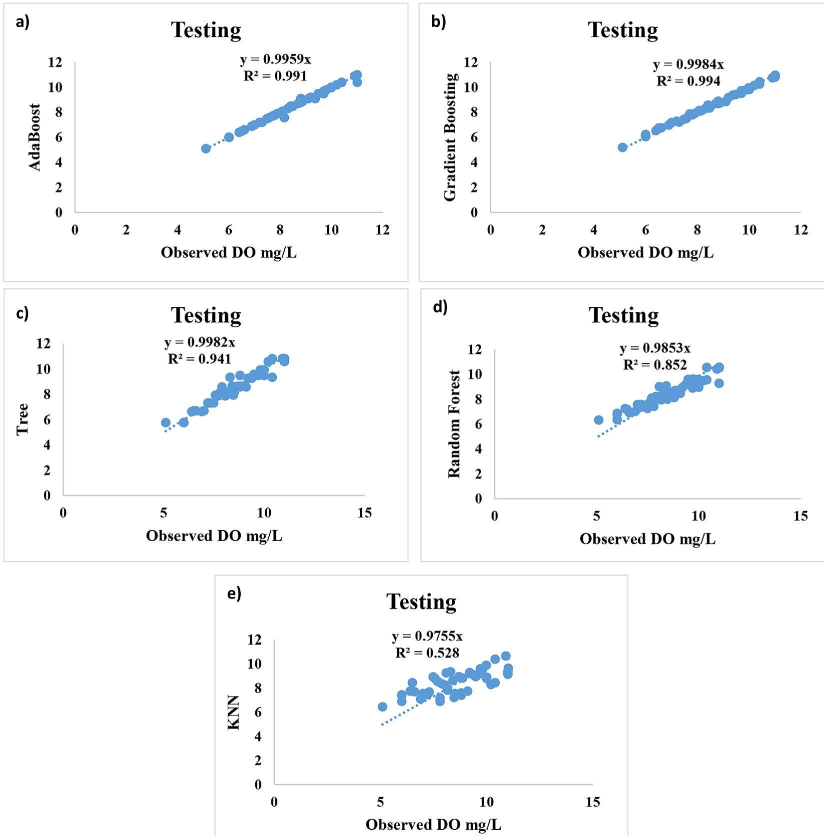


Figure 7. Scatter plot for the model predicted and observed DO levels in Al-Aammah Bridge (a) AdaBoost, (b) Gradient Boosting, (c)Tree, (d)Random Forest, (e) KNN

Bridge (R^2 = 0.994) and in Al-Aammah Bridge (R^2 = 0.994). AdaBoost followed closely in performance in Al-Muthanna Bridge (R^2 = 0.992) and in Al-Aammah Bridge (R^2 = 0.991), followed by Tree (R^2 = 0.866 in Al-Muthanna Bridge and R^2 = 0.941 in Al-Aammah Bridge). While RF performance lagged with R^2 = 0.866 in Al-Muthanna Bridge and 0.941 in AI-Aammah Bridge, followed by KNN with R^2 = 0.646 in Al-Muthanna Bridge and 0.528 in Al-Aammah Bridge. Overall, visual and statistical assessments agreed, indicating that the models performed well in predicting BOD and DO values.

CONCLUSIONS

Five different forms of artificial intelligence were evaluated in this study i.e. AdaBoost, Gradient Boosting, Tree, Random Forest, and KNN to calculate and predict DO and BOD concentrations in the Tigris River at Al-Muthanna and Al-Aammah Bridges. These models were evaluated in this paper as a more reliable technique to predicting WQ parameters than laboratory analysis. The input qualities for the suggested models have been selected from a several types of water factors, including chemical, physical, and biological. The model was constructed using laboratory data over a 15-year period, from 2008 to 2022. The evaluation employed four assessment criteria, including: MSE, RMSE, MAE, and R^2. It was found that AdaBoost and Gradient Boosting performed better than the other assessed approaches. In another words, Gradien

Acknowledgements

The Iraqi Ministry of Environment and the University of Technology in Baghdad, Iraq, are to be thanked by the authors for their cooperation in completing this work.

REFERENCES

1. Qiu, D., Zhu, G., Lin, X., Jiao, Y., Lu, S., Liu, J., Zhang, W., Ye, L., Li, R., Wang, Q., Chen, L. 2023. Dissipation and movement of soil water in artificial forest in arid oasis areas: Cognition based on stable isotopes. CATENA, 228, 107178.

2. Zhou, G., Yang, Z. 2023. Analysis for 3-D morphology structural changes for underwater topographical in Culebrita Island. International Journal of Remote Sensing, 44(7), 2458–2479.
3. Biesbroek, R., Wright, S.J., Eguren, S.K., Bonotto, A., Athanasiadis, I.N. 2022. Policy attention to climate change impacts, adaptation and vulnerability: a global assessment of National Communications (1994–2019). Climate Policy, 22(1), 97–111.
4. Fuller, R., Landrigan, P.J., Balakrishnan, K., Bathan, G., Bose-O'Reilly, S., Brauer, M., et al. 2022. Pollution and health: a progress update. The Lancet Planetary Health, 6(6), e535–e547.
5. Wang, P., Yao, J., Wang, G., Hao, F., Shrestha, S., Xue, B., Xie, G., Peng, Y. 2019. Exploring the application of artificial intelligence technology for identification of water pollution characteristics and tracing the source of water quality pollutants. Science of the Total Environment, 693, 133440.
6. Al-Janabi, Z.Z., Al-Kubaisi, A.R., Al-Obaidy, A.H.M.J. 2012. Assessment of water quality of Tigris River by using water quality index (CCME WQI). Al-Nahrain Journal of Science, 15(1), 119–126.
7. 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.
8. Elbeltagi, A., Al-Mukhtar, M., Kushwaha, N.L., Al-Ansari, N., Vishwakarma, D.K. 2023. Forecasting monthly pan evaporation using hybrid additive regression and data-driven models in a semi-arid environment. Applied Water Science, 13(2), 42.
9. Tung, T.M., Yaseen, Z.M. 2020. A survey on river water quality modelling using artificial intelligence models: 2000–2020. Journal of Hydrology, 585, 124670.
10. Ding, Y.R., Cai, Y.J., Sun, P.D., Chen, B. 2014. The use of combined neural networks and genetic algorithms for prediction of river water quality. Journal of applied research and technology, 12(3), 493–499.
11. Wang, J., Shi, Y., Zhang, R., Wu, Z., Ye, H., Li, S. 2023. CVT on-line error measurement hybrid-driven by domain knowledge and Stacking Model. Engineering Applications of Artificial Intelligence, 125, 106710.
12. Kılıç, H., Çetin, A. 2023. A novel graph-based ensemble token classification model for keyword extraction. Arabian Journal for Science and Engineering, 48(8), 10673–10680.
13. Montazeri, A.H., Emami, S.K., Zaghiyan, M.R., Eslamian, S. 2023. Stochastic learning algorithms. In Handbook of Hydroinformatics. Elsevier. 385–410.
14. Geshnigani, F.S., Golabi, M.R., Mirabbasi, R., Tahroudi, M.N. 2023. Daily solar radiation estimation in Belleville station, Illinois, using ensemble artificial intelligence approaches. Engineering Applications of Artificial Intelligence, 120, 105839.
15. Kang, Y., Song, J., Li, K., Zhai, X.A., Li, Y. 2022. Research on water quality prediction model based on echo state network. Journal of Computational Methods in Sciences and Engineering, 22(3), 901–910.
16. Liu, C., Xu, M., Liu, Y., Li, X., Pang, Z., Miao, S. 2022. Predicting groundwater indicator concentration based on long short-term memory neural network: A case study. International Journal of Environmental Research and Public Health, 19(23), 15612.
17. Khosravi, K., Miraki, S., Saco, P.M., Farmani, R. 2021. Short-term River streamflow modeling using Ensemble-based additive learner approach. Journal of Hydro-environment Research, 39, 81–91.
18. Shahdad, M., Saber, B. 2022. Drought forecasting using new advanced ensemble-based models of reduced error pruning tree. Acta Geophysica, 70(2), 697–712.
19. Saha, T.K., Pal, S., Sarda, R. 2022. Impact of river flow modification on wetland hydrological and morphological characters. Environmental Science and Pollution Research, 29(50), 75769–75789.
20. Khosravi, K., Mao, L., Kisi, O., Yaseen, Z.M., Shahid, S. 2018. Quantifying hourly suspended sediment load using data mining models: case study of a glacierized Andean catchment in Chile. Journal of Hydrology, 567, 165–179.
21. Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., Liu, F., Zuo, M., Zou, X., Wang, J., Zhang, Y., Chen, D., Chen, X., Yong feng, D., Ren, H., Ren, H. 2020. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water research, 171, 115454.
22. Asadollahfardi, G., Taklify, A., Ghanbari, A. 2012. Application of artificial neural network to predict TDS in Talkheh Rud River. Journal of Irrigation and Drainage Engineering, 138(4), 363–370.
23. Mahdavi-Meymand, A., Sulisz, W., Zounemat-Kermani, M. 2024. Hybrid and integrative evolutionary machine learning in hydrology: A systematic review and meta-analysis. Archives of Computational Methods in Engineering, 31(3), 1297–1340.
24. Jamei, M., Karbasi, M., Alawi, O.A., Kamar, H.M., Khedher, K.M., Abba, S.I., Yaseen, Z.M. 2022. Earth skin temperature long-term prediction using novel extended Kalman filter integrated with Artificial Intelligence models and information gain feature selection. Sustainable Computing: Informatics and Systems, 35, 100721.
25. Alavi, J., Ewees, A.A., Ansari, S., Shahid, S., Yaseen, Z.M. 2022. A new insight for real-time wastewater quality prediction using hybridized kernel-based extreme learning machines with advanced optimization algorithms. Environmental Science

and Pollution Research, 29(14), 20496–20516.

26. Manzar, M.S., Benaafi, M., Costache, R., Alagha, O., Mu'azu, N.D., Zubair, M., Abdullahi J., Abba, S.I. 2022. New generation neurocomputing learning coupled with a hybrid neuro-fuzzy model for quantifying water quality index variable: A case study from Saudi Arabia. Ecological Informatics, 70, 101696.
27. Boano, F., Revelli, R., Ridolfi, L. 2006. Stochastic modelling of DO and BOD components in a stream with random inputs. Advances in Water Resources, 29(9), 1341–1350.
28. Benaafi, M., Yassin, M.A., Usman, A.G., Abba, S.I. 2022. Neurocomputing Modelling of Hydrochemical and Physical Properties of Groundwater Coupled with Spatial Clustering, GIS, and Statistical Techniques. Sustainability, 14(4), 2250.
29. Ramal, M.M., Jalal, A.D., Sahab, M.F., Yaseen, Z.M. 2022. River water turbidity removal using new natural coagulant aids: case study of Euphrates River, Iraq. Water Supply, 22(3), 2721–2737.
30. Adedeji, I.C., Ahmadisharaf, E., Sun, Y. 2022. Predicting in-stream water quality constituents at the watershed scale using machine learning. Journal of Contaminant Hydrology, 251, 104078.
31. Adnan, T.A., Mohammed, E.A., Al-Madhhachi, A.S.T. 2021. Water quality index of tigris river within baghdad city: a review. Journal of Engineering and Sustainable Development, 25(3), 34–43.
32. Al Obaidy, H.M., S Awad, E., Zahraw, Z. 2016. Impact of Medical City and Al-Rasheed power plant effluents on the water quality index value of Tigris River at Baghdad city. Engineering and Technology Journal, 34(4A), 715–724.
33. Oleiwi, A.S., Al-Dabbas, M. 2022. Assessment of contamination along the Tigris River from Tharthar-Tigris canal to Azizziyah, middle of Iraq. Water, 14(8), 1194.
34. Breiman, L. 2001. Random forests. Machine Learning, 45(1), 5–32.
35. Chen, W., Li, Y., Xue, W., Shahabi, H., Li, S., Hong, H., Wang, X., Bian, H., Zhang, S., Pradhan, B., Ahmad, B.B. 2020. Modeling flood susceptibility using data-driven approaches of naïve bayes tree, alternating decision tree, and random forest methods. Science of The Total Environment, 701, 134979.
36. Cutler, D.R., Edwards Jr, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J. 2007. Random forests for classification in ecology. Ecology, 88(11), 2783–2792.
37. Abbas, J. 2013. Assessment of water quality in Tigris River-Iraq by using GIS mapping. Natural Resources, 2013.
38. Belgiu, M., Drăguţ, L. 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS journal of photogrammetry and remote sensing, 114, 24–31.
39. Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H. 2009. The elements of statistical learning: data mining, inference, and prediction. Springer, 2, 1–758.
40. Goel, E., Abhilasha, E., Goel, E., Abhilasha, E. 2017. Random forest: A review. International Journal of Advanced Research in Computer Science and Software Engineering, 7(1), 251–257.
41. Hastie, T., Rosset, S., Zhu, J., Zou, H. 2009. Multi-class adaboost. Statistics and its Interface, 2(3), 349–360.
42. Bishop, C.M. 2006. Pattern recognition and machine learning. Springer google schola, 2, 1122–1128.
43. Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R., Friedman, J. 2009. Boosting and additive trees. The elements of statistical learning: data mining, inference, and prediction, 337–387.
44. Blu, T., Thévenaz, P., Unser, M. 2004. Linear interpolation revitalized. IEEE Transactions on Image Processing, 13(5), 710–719.
45. Jolliffe, I.T., Cadima, J. 2016. Principal component analysis: a review and recent developments. Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.
46. Al-Mukhtar, M., Srivastava, A., Khadke, L., Al-Musawi, T., Elbeltagi, A. 2024. Prediction of irrigation water quality indices using random committee, discretization regression, REPTree, and additive regression. Water Resources Management, 38(1), 343–368.
47. Bhagat, S.K., Tiyasha, T., Tung, T.M., Mostafa, R.R., Yaseen, Z.M. 2020. Manganese (Mn) removal prediction using extreme gradient model. Ecotoxicology and Environmental Safety, 204, 111059.
48. Bhagat, S.K., Paramasivan, M., Al-Mukhtar, M., Tiyasha, T., Pyrgaki, K., Tung, T.M., Yaseen, Z.M. 2021. Prediction of lead (Pb) adsorption on attapulgite clay using the feasibility of data intelligence models. Environmental Science and Pollution Research, 28, 31670–31688.