Environ Monit Assess (2019) 191: 380
https://doi.org/10.1007/s10661-019-7446-8

Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods

Mohammad Najafzadeh · Alireza Ghaemi

Received: 11 January 2019 / Accepted: 4 April 2019 / Published online: 19 May 2019

Abstract Rivers, as the most prominent component of water resources, have a key role to play in increasing the life expectancy of living creatures. The essential characteristics of water pollutants can be described by water quality indices (WQIs). Hence, a ferocious demand for obtaining an accurate prediction of WQIs is of high importance for perception of pollutant patterns in natural streams. Field studies conducted on different rivers indicated that there is no general relationship to yield water quality parameters with a permissible level of accuracy. Over the past decades, several artificial intelligence (AI) models have been employed to predict more precise estimation of WQIs rather than conventional models. In this way, through the current study, multivariate adaptive regression spline (MARS) and least square-support vector machine (LS-SVM), as machine learning methods, were used to predict indices of the five-day biochemical oxygen demand (BOD5) and chemical oxygen demand (COD). To improve the proposed approaches, 200 series of field data, collected from Karoun River southwest of Iran, pertain to the nine independent input parameters, namely electrical conductivity (EC), sodium (Na^{+}), calcium (Ca^{2+}), magnesium (Mg^{2+}), orthophosphate (PO_4^{3-}), nitrite (NO_2^{-}), nitrate nitrogen (NO_3^{-}), turbidity, and pH. The performances of the LS-SVM and MARS techniques were quantified in both training and testing stages by means of several statistical parameters. Furthermore, the results of the proposed AI models were compared with those obtained using artificial neural network (ANN), adaptive neuro-fuzzy inference system (ANFIS), and multiple regression equations. Results of the present research work indicated that the proposed artificial intelligence techniques, as machine learning classifiers, were found to be efficient in order to predict water quality parameters.

Keywords River pollution · Water quality indices · Least square-support vector machine · Multivariate adaptive regression spline · Multiple regression-based equations · Statistical criteria

Introduction

Water bodies introuduced as rivers, lakes, and oceans, are one of the most important elements of creatures' life cycle. The vast majority of human activities such as industrial and commercial affairs are a major factor of water pollution. This means that water quality plummets

M. Najafzadeh (✉) · A. Ghaemi
Department of Water Engineering, Faculty of Civil and Surveying Engineering, Graduate University of Advanced Technology, Kerman, Iran
e-mail: m.najafzadeh@kgut.ac.ir

e-mail: moha.najafzadeh@gmail.com

A. Ghaemi
e-mail: alirezaa.ghaemi@gmail.com

due to increasing the environment-devastating activities, wasteful exploitation of water bodies, and chemical, biological, and physical factors. In this way, finding the trustworthy way for controlling the pollution of water bodies has engrossed the attention of environmentalists. Among the environmental factors in assessment of water quality, the 5-day biochemical oxygen demand (BOD5), chemical oxygen demand (COD), and dissolved oxygen (DO) were introduced as the most prominent indices in environmental projects (e.g., Deininger et al. 2011; Noori et al. 2012; Feng et al. 2012; Emamgholizadeh et al. 2014; Mohammadpour et al. 2014).

The amount of DO concentration, as one of the essential indices, has inextricably bound with durability of creatures living in aquatic environments. Temperature (TEMP) is one of the effective variables on the amount of DO concentration in which the oxygen can efficiently be preserved in the cold waters in comparison with the warm ones. Moreover, BOD5 is the amount of DO being consumed in the water bodies in order to oxidize the organic matters by aerobic biological organisms. The amount of BOD5 concentration is an indicative of pollution level. Apparently, an upward trend in the BOD5 level leads to reduce the concentration of DO. However, having the preknowledge of DO is an indispensable issue in order to determine the concentration of BOD5.

Furthermore, COD is the required amount of oxygen to complete chemical decomposition process of the organic materials in water bodies. It is inevitable that the concentration level of the COD index is higher than that of the BOD5 index. Due to the complexity of pollution processing in rivers and establishing a conceptual relationship among effective variables, conventional methods maybe not efficiently applied to assess water quality indices (e.g., Singh et al. 2009; Verma and Singh 2013).

In the recent years, by the advent of artificial intelligence (AI) techniques such as artificial neural network (ANN), adaptive neuro-fuzzy inference system (ANFIS), genetic programming (GP), evolutionary polynomial regression (EPR), and model tree (MT), the vast majority of problems in the fields of water sciences were solved (e.g., Noori et al. 2009; Chau and Wu 2010; Singh et al. 2010; Laucelli and Giustolisi 2011; Azamathulla and Ghani 2011; Sreekanth and Datta 2011; Laucelli et al. 2012; Chen et al. 2013; Fallah-Mehdipour et al. 2013; Xu et al. 2013; Ebtehaj and Bonakdari 2014; Gholami et al. 2015; Taormina et al. 2015; Najafzadeh et al. 2016; Rahimikhoob 2016; Kumar et al. 2016; Wang et al. 2017; Yaseen et al. 2018).

Furthermore, researchers have developed AI techniques in order to assess the water quality in various conditions of surface waters such local geology, ecosystem, and human uses. For instance, Dogan et al. (2009) have made use of ANN in order to model BOD5 in Melen River. They concluded that ANN could estimate BOD5 index with a permissible degree of precision. Asadollahfardi et al. (2012) evaluated the two types of ANN approaches, namely multilayer perceptron (MLP) and recurrent neural network (RNN), for prediction of total dissolved solids (TDS). Their concrete results showed that the RNN had better performance than the MLP.

Orouji et al. (2013) employed ANFIS and GP to prognosticate monthly concentration of time-dependent water quality parameters of electrical conductivity (EC), sodium (Na^+), potassium (K^+), magnesium (Mg^{2+}), sulfate (SO_4^{2-}), chloride (Cl^-), nitrate (NO_3^-), TDS, and pH. Through their study, even though the equations given by GP were found to be comprehensively complex, GP was an efficient tool to predict water quality parameters in comparison with the ANFIS model.

Emamgholizadeh et al. (2014) have predicted BOD5 ad COD parameters by means of ANN and ANFIS models. They applied nine input variables, namely EC, Na^+, calcium (Ca^{2+}), Mg^{2+}, orthophosphate (PO_4^{3-}), nitrite (NO_2^-), NO_3^-, turbidity, and pH. Ultimately, the performance of the proposed models indicated that the ANFIS model has provided more accurate prediction than those obtained using ANN. Alizadeh and Kavianpour (2015) implemented the ANN and wavelet-neural network (WNN) to predict the DO concentration. Throughout the performance of the proposed models, WNN was defined as an efficient technique for the DO level prediction in comparison with the ANN model. Mohammadpour et al. (2016) have employed gene-expression programming (GEP) and ANN to predict water quality indicators (WQIs). Their findings indicated that ANN and GEP could prognosticate WQIs with permissible level of precision. Moreover, Li et al. (2016) used particle swarm optimization (PSO) to optimize the structure of ANN and SVM techniques for the prediction of DO index. The results of their study indicated that the performance of PSO-SVM was better than that of PSO-ANN. However, their research work was limited to the rivers of China with

restricted knowledge extraction of in-depth information about other water quality indices.

Yaseen et al. (2018) have applied various hybrid ANFIS models for estimation of water quality index. To model WQI, they have considered different combinations of parameters including turbidity, BOD5, COD, Ca^{2+},DO, pH, TEMP, and total solids (TS), as input variables for their proposed models. Finally, they concluded that WQI values given by the proposed ANFIS approaches were in good agreement with observed WQI values.

Furthermore, Heddam and Kisi (2018) have used three AI models of MT, MARS, and least square-support vector machine (LS-SVM) to predict monthly DO concentration in natural streams. From their study, they have applied TEMP, EC, discharge flow, and specific conductance (SC) as input parameters. They found that the proposed models had good performance in the evaluation of the DO index.

From previous investigations, one of the relatively significant disadvantages of ANN and ANFIS is that these AI techniques do not have the capability of establishing a relationship between input and output variables of complicated systems. Furthermore, the results of previous researches indicated that the main disadvantage of GEP (or GP) is to produce a nonlinear mathematical equation with a comparatively high degree of complication among the input and output variables. Moreover, it was found that the SVM outweighs ANN in the prediction of water quality indices. This is due to the nature of the ANN structure because of the lack of multiple local minima, whereas the solution to LS-SVM is unique. As a merit, the MARS approach, as a classifier model, can plummet the degree of complexity for mathematical expressions by means of linearizing all the equations and, consequently, leading to more accurate estimation rather than other artificial intelligence models.

Through application of AI models for prediction of water quality indices, the MARS model has rarely been employed to predict water quality indices. Even though rigorous endeavors have been made to predict various WQIs in different natural streams, it seems that some previous research works suffer from two main shortcomings. The first one is related to the number of input parameters. In fact, the vast majority of studies, carried out in the recent decades, used a limited number of input variables due to general difficulty of dataset measurement in rivers such as high costs of using gauged sites for recording daily concentration of various water quality parameters. The second shortcoming is that lots of AI models employed in the previous investigations were classified into black-box approaches. On the other hand, these techniques had no sufficient generalization, and additionally, these models produced an equation with complicated mathematical expressions.

In the present investigation, MARS and LS-SVM are applied to evaluate water quality indices. The qualitative and quantitative results of the proposed models are investigated in both training and testing phases. Moreover, the results of the MARS and LS-SVM techniques are implemented to compare with ANN, ANFIS, and multiple linear and nonlinear regression models. On the other hand, statistical analyses in terms of criteria of external validation and Fisher test are conducted to obtain the best proposed AI model in the prediction of WQIs.

Data description

Among the various rivers in Iran, Karoun River with roughly 950 km length and 65,230 km^2 area is considered the longest river located in Khuzestan. There is no denying the fact that a large number of investigators have made a lot of research works on Karoun River. In addition, this river plays a substantial role in supplying power and the water demands for agricultural and industrial activities in Iran. Furthermore, an overview of Karoun River is depicted in Fig. 1.

On the basis of the most recent investigations, so many water quality parameters such as EC, PO_4^{3-}, Na^+, Ca^{2+}, Mg^{2+}, NO_2^-, NO_3^-, turbidity, pH, and total suspended solids (TSS) have been applied to evaluate WQIs (e.g., Emamgholizadeh et al. 2014; Yaseen et al. 2018). In the current research, the water quality parameters mentioned in the investigation of Emamgholizadeh et al. (2014) were considered to develop AI techniques. In this way, BOD5 and COD were selected as WQIs and the remaining parameters were considered as input variables,

$$\mathrm{BOD5} = f\left(\mathrm{Ca}^{2+},\ \mathrm{Na}^{+},\ \mathrm{Mg}^{2+},\ \mathrm{NO}_2^{-},\ \mathrm{NO}_3^{-},\ \mathrm{PO}_4^{3-},\ \mathrm{EC},\ \mathrm{pH},\ \mathrm{Turbidity}\right) \quad (1)$$

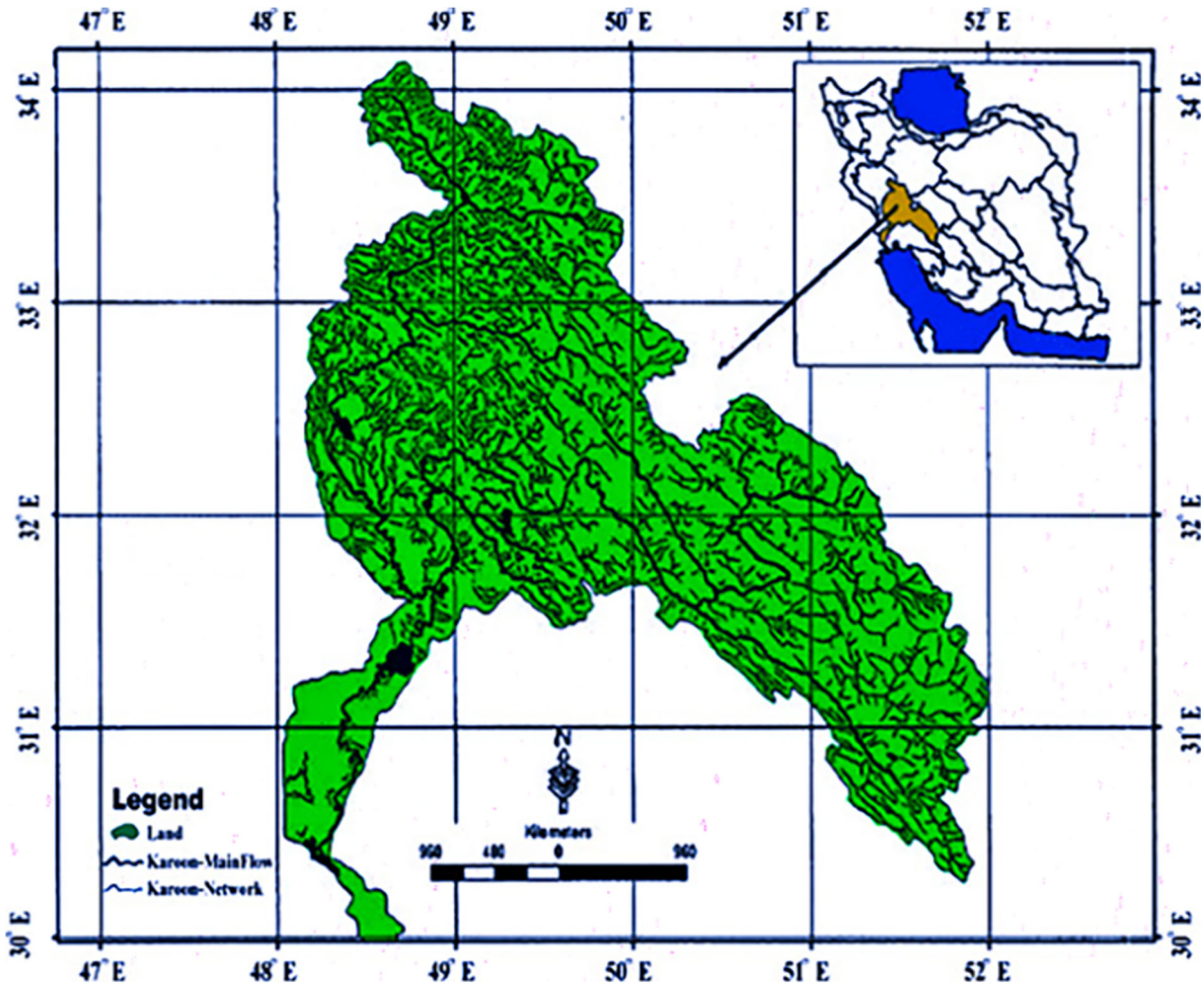

Fig. 1 Overview of the case study district

$$\mathrm{COD} = g\,(\mathrm{Ca}^{2+},\ \mathrm{Na}^{+},\ \mathrm{Mg}^{2+},\ \mathrm{NO}_2^{-},\ \mathrm{NO}_3^{-},\ \mathrm{PO}_4^{3-},\ \mathrm{EC},\ \mathrm{pH},\ \mathrm{Turbidity}) \quad (2)$$

Descriptive statistics of field dataset are seen in Table 1. From Table 1, among the parameters, NO_2^- with 0.08 mg/L stood at the minimum value of concentration. Also, standard deviation (SD) value for NO_2^- illustrated that this value was closer to its average when compared with other water quality parameters. Table 1 indicates that Mg^{2+} has the maximum values of concentration (60 mg/L) and SD (11.74).

From precious previous studies in these areas, it can be said that these input parameters mentioned in Eqs. (1) and (2) have a key role to play in the estimation of the WQIs. For instance, the levels of orthophosphates and nitrate concentration that are intolerable to local organisms are known to deplete DO levels by algae blooms. Moreover, the concentration levels of the COD index are in close connection with the orthophosphate concentration. What is more, turbidity is a significant element in water quality evaluation due to the fact that the degree of stream water turbidity is occasionally considered as a rough measure of intensity of water pollution. In fact, an increase in light absorbency level by turbid water not only ascends water temperatures but also declines oxygen concentration levels. The contents of the particles causing turbidity can also lead to a decline in dissolved oxygen levels (e.g., Emamgholizadeh et al. 2014).

Additionally, to visualize the distribution related to all the water quality parameters, frequency histograms are illustrated in Fig. 2. In the case of datasets, the sole assumption is that all the variables have no dependency on time.

Table 1 Statistical properties of water quality parameters in Karoun River

Variable	Unit	Max	Min	Ave	Std. dev
Ca^{2+}	mg/L	58.4	1	12.47	9.20
Na^{+}	mg/L	40	1.42	18.48	7.50
Mg^{2+}	mg/L	60	2.1	13.21	11.74
NO_2^-	mg/L	2.1	0.08	0.41	0.31
NO_3^-	mg/L	2.7	0.34	1.01	0.38
PO_4^{3-}	mg/L	3.21	0.13	1.21	0.78
EC	dS/m	9.26	1.7	4.72	1.62
pH	–	8.71	5.1	7.15	0.77
Turbidity	NTU	25	1	7.10	5.30
BOD5	mg/L	40.6	3.7	19.21	10.28
COD	mg/L	34.2	1.06	15.86	9.30

In this study, 200 datasets collected from Karoun River were monthly reported at eight different stations over a 16-year period. To implement the proposed models, datasets were divided into two sections in a way that three quarters of the datasets were used for the training stage, whereas the rest of the datasets are devoted to the testing models.

Proposed approaches

In this section, descriptions of two data-mining techniques, namely LS-SVM and MARS, are presented briefly. Furthermore, implementation of the proposed models to predict two indices of BOD5 and COD are given.

Implementation of the least square-support vector machine

LS-SVM is one of the supervised learning techniques for making a nonlinear relationship among input and output variables with a high level of precision. Generally, the principles of structural risk minimization (SRM) are employed in the LS-SVM in order to minimize the predicted error values, leading to reduction of overfitting occurrence. In the LS-SVM, in order to separate the dataset patterns, input variables (or vectors) are mapped into a higher dimensional feature space (Cortes and Vapnik, 1995; Suykens and Vandewalle, 1999; Kisi 2015; Mahmoudi et al. 2016; Adnan et al. 2017a, b).

For the training stage of the LS-SVM model, a series of datasets including (x_1, y_1), (x_2, y_2),….., (x_i, y_i),….., (x_N, y_N) is considered in which x_i is the input variable containing m features, y_i is the output variable that corresponded to x_i, and N is the number (or size) of data sample. The regression function in the SVM technique is introduced as (Li et al. 2016),

$$f(x) = \langle w, \mathrm{x} \rangle + h \tag{3}$$

where w and h are the vector of weights in the feature space with the dimension of x and bias term, respectively, and additionally $\langle *, * \rangle$ indicates the inner product. To minimize the regularized risk function with epsilon-insensitive loss function, Eq. (3) is defined as (Mahmoudi et al. 2016),

$$\frac{1}{2} w^T w + C\sum_{i=1}^{n} \xi_i + C\sum_{i=1}^{n} \xi_i' \tag{4}$$

$$\text{Subject to} \begin{cases} \langle w, x \rangle + h - y_i \le \varepsilon + \xi_i' \\ y_i - \langle w, x \rangle - h \le \varepsilon + \xi_i \\ \xi_i, \xi_i' \ge 0 \qquad i = 1, 2, \ldots, n \end{cases} \tag{5}$$

in which C is a non-negative constant determining the penalty for prediction error, and ξ_i and ξ_i' are slack variables to specify the distance from observed values to the corresponding boundary values of ε. It is expected that most of the points fall within the ε tube. Once a group of datasets are placed out-of-error bound, the errors ξ_i and ξ_i' are obtained. Through the LS-SVM, quadratic programming (QP) is one of the useful procedures which can solve a particular type of nonlinear optimization problem. In this way, by using Lagrangian multipliers, Eqs. (4) and (5) are transformed into the following form (Noori et al. 2011; Li et al. 2016):

$$\sum_{i=1}^{n} y_i \left(\alpha_i - \alpha_i'\right) - \varepsilon \sum_{i=1}^{n} \left(\alpha_i + \alpha_i'\right) - 0.5 \times \sum_{i.j=1}^{n} \left(\alpha_i - \alpha_i'\right)\left(\alpha_j - \alpha_j'\right) K(x, x_i) \tag{6}$$

$$\text{Subject to} \begin{cases} \sum_{i=1}^{n} \left(\alpha_i - \alpha_i'\right) = 0 \\ 0 \le \alpha_i, \alpha_i' \le C \\ i = 1, 2, \ldots, N \end{cases} \tag{7}$$

Where α_i and α_i' are Lagrange multipliers and $k(.)$, as the kernel function, is defined as,

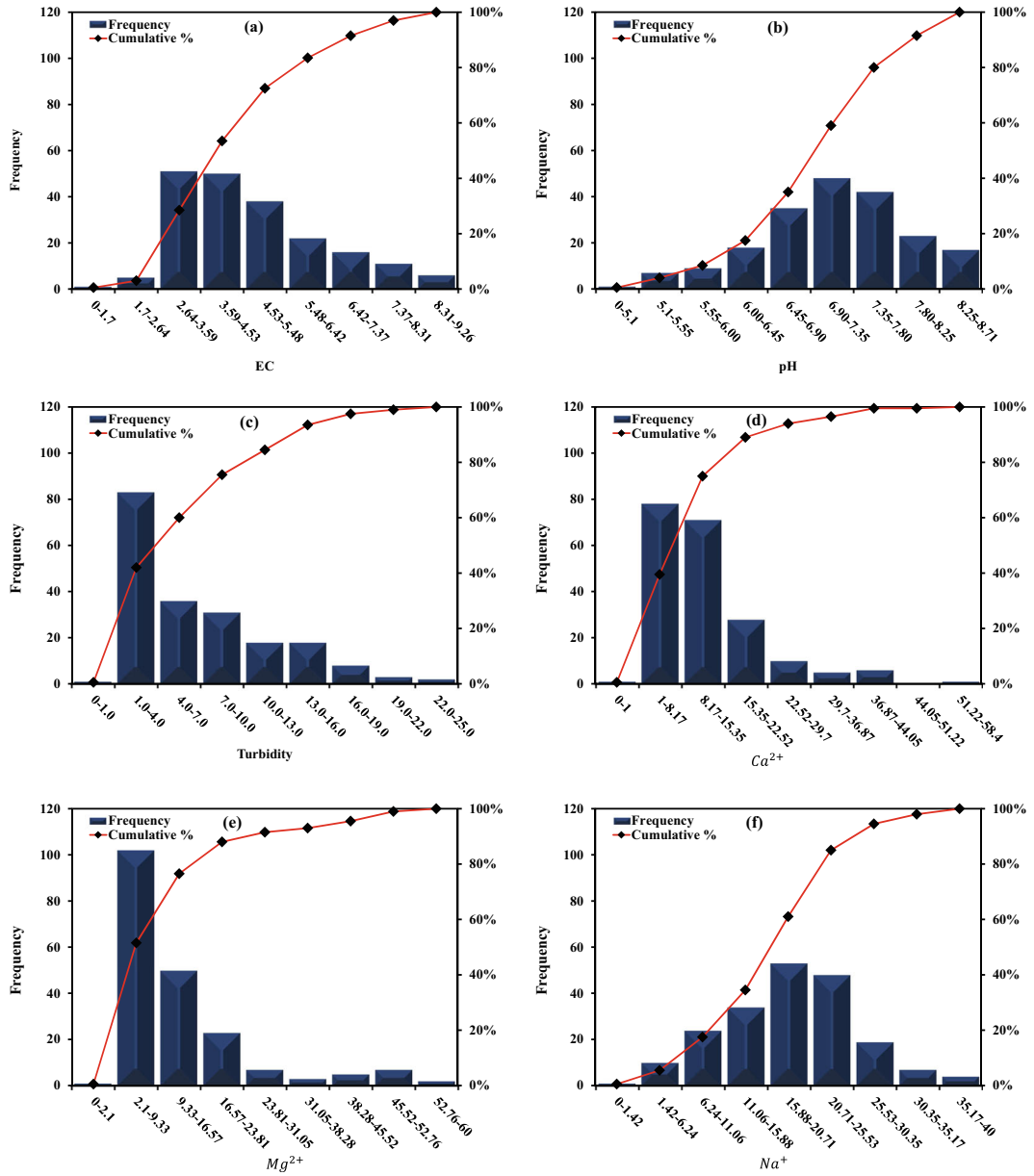

Fig. 2 Histograms of the variables in the models implementation: **a** EC, **b** pH, **c** turbidity, **d** Ca^{2+}, **e** Mg^{2+}, **e** Na^{+}, **g** PO_4^{3-}, **h** NO_3^-, **i** NO_2^-, **j** BOD5, and **k** COD

$$k(x_i, x_j) = \varnothing(x_i).\varnothing(x_j) \tag{8}$$

$$f(x) = \sum_{i.j=1}^{n} (\alpha_i - \alpha_i^{\cdot}) K(x, x_i) + h \tag{9}$$

The solution to objective function [Eq. (6)] is unique, and consequently, the following equation is an approximation of solution to Eqs. (6) and (7) (Noori et al. 2011):

Basically, various kernel functions including linear function, polynomial function, and radial basis function (RBF) are applied in the general structure of the LS-SVM technique to determine the most efficient model to

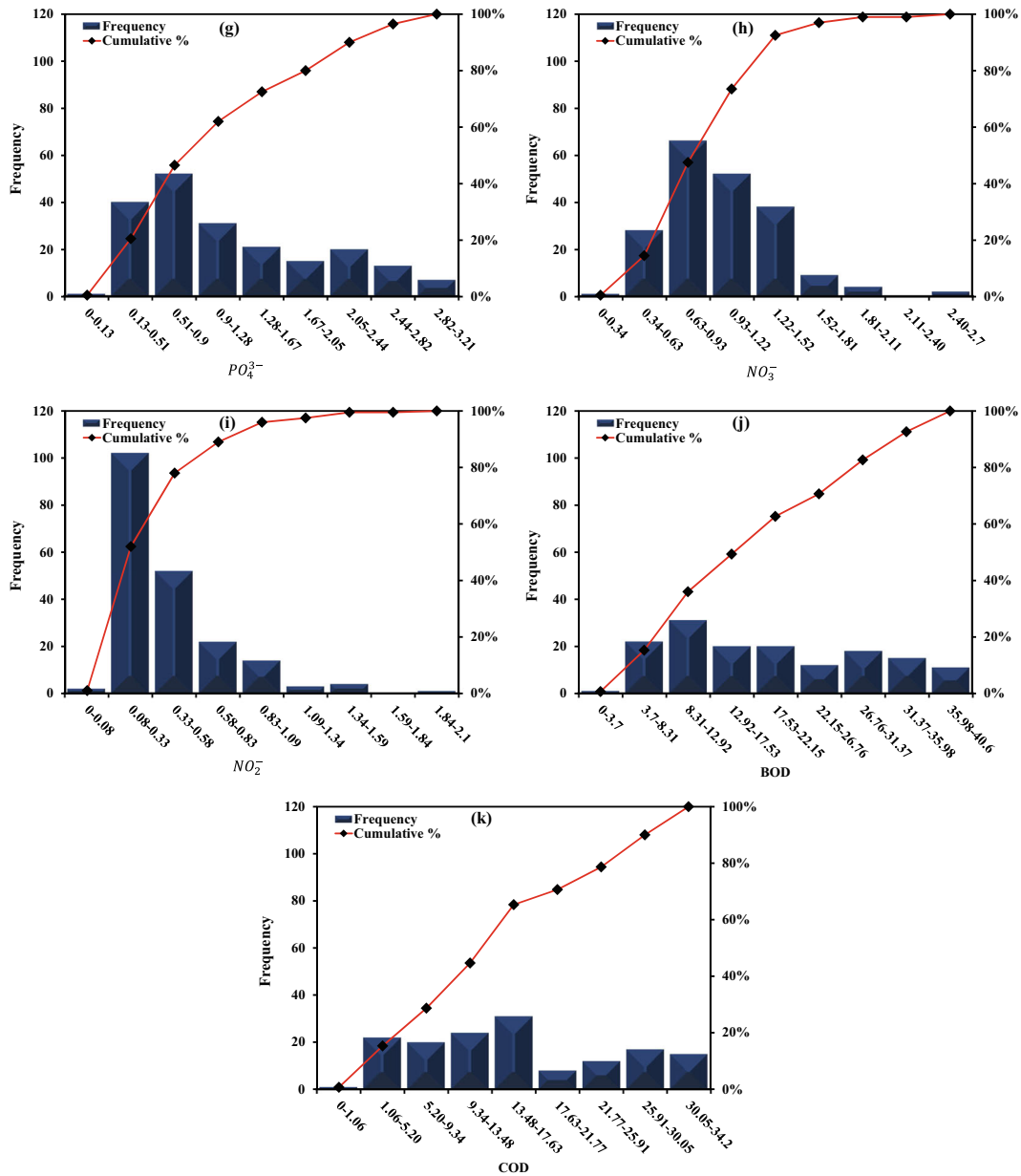

Fig. 2 continued.

predict BOD5 and COD indices. All sorts of kernels are expressed as,

Linear kernel: $k(x_i, x_j) = x_i^T x_j$
Polynomial kernel: $k(x_i, x_j) = (x_i^T x_j + \gamma)^d$, $\gamma > 0$
RBF kernel: $k(x_i. x_j) = \exp(-\gamma \parallel x_i - x_j \parallel)^2$, $\gamma > 0$
Sigmoid kernel: $k\left(x_i.x_j\right) = \tanh\left(\gamma x_i^T.x_j + r\right)$ $\gamma > 0$

where r, d, and γ (greater than 0) are kernel parameters being used to limit the complexity of the model. The performance of the LS-SVM approach in terms of precision level is highly dependent on a set of parameters including C, type of kernel function, and corresponding kernel parameters (e.g., Noori et al. 2009, 2011, 2015a, b; Li et al. 2016).

To implement the LS-SVM, indices of BOD5 and COD were considered as the dependent variables, and $Ca^{2+}, Na^{+}, Mg^{2+}, NO_2^-$, NO_3^-, PO_4^{3-}, EC, pH, and turbidity were selected as the independent variables. In the present study, to select the most accurate LS-SVM models for prediction of BOD5 and COD indices, all types of kernel functions were employed and mean squared error (MSE) criteria was used as a criteria to assess the training stages of LS-SVM models. The optimal parameters of kernel functions and the penalty parameter are given in Table 2. From Table 2, RBF (MSE = 11.27) and polynomial (MSE = 22.18) kernels produced relatively lower error for prediction of BOD5 compared with linear (MSE = 53.58) and sigmoid (MSE = 48.132) kernel functions. In the case of COD prediction, the LS-SVM approach with RBF (MSE = 13.46) and polynomial kernel (MSE = 22.74) functions had higher accuracy level rather than LS-SVM proposed by linear (MSE = 36.65) and sigmoid (MSE = 33.52) kernel functions. In addition, qualitative and quantitative comparisons of LS-SVM performance with other AI techniques and multiple regression-based equations are carried out on the basis of improved LS-SVM models by RBF and polynomial kernel functions.

Implementation of multivariate adaptive regression splines

MARS is nonparametric regression approach which can reduce the complexity degree of nonlinear systems by means of establishing a set of piecewise linear splines (segments) among system variables with various slopes (gradients). There is no particular assumption about the linear relationship being produced between the input and the output variables. The end points of the segments are introduced as knots. In fact, a knot is indicative of the end of one region related to the dataset and the beginning of another one. The locations of the knots are selected by means of an adaptive regression technique. The MARS model provides basis functions (BFs) by searching in a stepwise pattern. Basically, the MARS technique has the capability to reduce the complexity of nonlinear systems by using a two-phase approach. The first phase is related to forward manner adding functions and finding potential knots in order to strengthen the accuracy level of MARS. Also, the main goal of this step is to plummet the possibility of overfitting. The second strategy is introduced as a backward phase in which linear terms with lower effects on the performance of MARS are removed (Kisi 2015; Adnan et al. 2017a, b). On the other hand, for augmenting the accuracy level of MARS performance, the backward technique is applied to cut down the unessential datasets from the former selection using generalized cross validation (GCV). The GCV relationship is expressed as (Zhang and Goh 2013),

$$\mathrm{GCV} = \frac{\mathrm{MSE}}{\left(1-\frac{\mathrm{CM}}{N}\right)^2} \tag{10}$$

In which CM is the penalty factor being expressed as,

$$\mathrm{CM} = M + 0.5 \times d \times (M-1) \tag{11}$$

Where d and M are the determination parameter and the number of BFs (basis functions), respectively.

To construct BFs within a stepwise manner, $X(x_1, x_2, x_3, \ldots, x_N)$ is assigned as an input vector. Fundamentally, the general relationship of $Y = f(X) + \xi$ is considered to connect X with Y (output vector), in which ξ is the distribution of the model error and N is the number of training datasets. Thus, $f(X)$ function is approximated using BFs. In fact, BFs are splines whose polynomials have a smooth manner. Once BFs are generated, the piecewise linear function is considered due to the lowering complexity degree of BFs. The general form of piecewise linear functions is $\max(0, x-t)$, being introduced as the occurrence of a knot at t value. The formulation max(0) is indicative of the non-negative part of

Table 2 Setting parameters of various kernel functions used in the LS-SVM models

Kernel function	Setting parameters	BOD5	COD
Polynomial	γ	0.00127	0.000346
	d	3	3
	C	3.79	9
	MSE	22.18	22.74
RBF	γ	2.97	5.378
	C	5.242	10.88
	MSE	11.27	13.46
Sigmoid	γ	0.0031	0.0031
	r	0.555	0.00
	C	4.562	4.227
	MSE	48.132	33.52
Linear	C	0.0356	0.0686
	MSE	53.58	36.65

(0), otherwise a 0 value is inevitably acquired. Basically, the mathematical relationship of $\max(0, x-t)$ is expressed as,

$$\max(0, x-t) = \begin{cases} x-t, & \text{if } x \geq t \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

The MARS model is a linear combination of BFs whose general mathematical relationship is expressed as,

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m . \lambda_m(X) \quad (13)$$

In which $\lambda_m(X)$ is introduced as a basis function which might comprise one spline function or the product of two or more spline functions, and β parameters are constant coefficients and can be estimated by means of the least square technique. Mathematically, it should be noted that high-order basis functions can be employed once datasets justify it. To obtain a simplified $f(x)$, the second-order BFs are used at most, and generally, it can be said that this is one of the most significant assumptions in this research.

The performance of the training stage for the MARS model leads to generate Eq. (13) through the forward stepwise approach. To obtain this aim, a spline is firstly generated with only the intercept, β_0, and a pair of basis which provides the lowest level of computational error in the training stages is merged. With respect to the present model with M BFs, the next pair is constructed and merged to the present model in the following form (Zhang and Goh 2013),

$$\hat{\beta}_{M+1} . \lambda_m(X) . \max\left(0, X_j - t\right) + \hat{\beta}_{M+2} . \lambda_m(X) . \max\left(0, t - X_j\right) \quad (14)$$

where the least square method is applied to estimate each β. As a new BF is put in the model space, interactions among BFs which are already in the model have high importance. BFs are continuously joined to the model space until the model obtains quite a few maximum particularized number of terms causing a decrease in overfitting of the model. After the MARS model is obtained, the overall relationship is acquired by means of combining all the BFs. Furthermore, analysis of variance (ANOVA) decomposition is employed to appraise the degree of relative importance of the input vectors and the basis functions (Zhang and Goh 2013).

In this study, the maximal number of BFs was fixed to 16 for both BOD5 and COD prediction, and additionally, only pairwise products of BFs are met in Table 3. As seen in Table 3, 10 piecewise linear BFs including the intercept terms were used to acquire optimum models for both COD and BOD5 prediction. Moreover, the results of ANOVA decomposition of the proposed MARS models for prediction of WQIs are given in Table 4. The best models given by the MARS model for prediction of BOD5 and COD indices were expressed as,

$$\begin{aligned} \text{BOD5} = {} & 47.83 - 8.804 \times \text{BF1} - 1.057 \\ & \times \text{BF2} - 4.215 \times \text{BF3} - 5.744 \times \text{BF4} \\ & + 6.027 \times \text{BF5} - 13.92 \times \text{BF6} + 2.387 \\ & \times \text{BF7} + 95.34 \times \text{BF8} - 6.979 \times \text{BF9} \\ & + 3.362 \times \text{BF10} - 20.84 \times \text{BF11} - 6.989 \\ & \times \text{BF12} + 45.15 \times \text{BF13} + 7.853 \\ & \times \text{BF14} - 0.08463 \times \text{BF15} \end{aligned} \quad (15)$$

$$\begin{aligned} \text{COD} = {} & 36.7 - 0.4042 \times \text{BF1} - 0.704 \\ & \times \text{BF2} - 1.053 \times \text{BF3} - 4.616 \times \text{BF4} \\ & + 0.4645 \times \text{BF5} + 0.5425 \times \text{BF6} - 62.3 \\ & \times \text{BF7} + 5.723 \times \text{BF8} + 54.51 \\ & \times \text{BF9} - 78.24 \times \text{BF10} + 13.96 \times \text{BF11} \\ & + 0.3141 \times \text{BF12} - 10.5 \times \text{BF13} - 19.12 \\ & \times \text{BF14} + 14.75 \times \text{BF15} \end{aligned} \quad (16)$$

With respect to Eq. (15), it can be found that pH, Mg^{2+}, and EC have no contribution in the prediction of the BOD5 index, while in Eq. (16), the turbidity parameter has no importance in the estimation of the COD index. In fact, MARS has a high level of decision-making whether all the input parameters have an important role in the prediction of WQPs or not. On the other hand, MARS can optimize the number of input variables and consequently diminishing linear equations [Eqs. (15) and (16)], compared with the equations given by multiple linear and nonlinear regression equations.

Table 3 Basis functions and corresponding linear equations of the MARS approach

BF	Equation
BOD5	
BF1	max (0, 1.39 − PO_4^{3-})
BF2	max (0, 7.6 − Turbidity)
BF3	max (0, 0.8 − NO_2^-) × max (0, Ca^{2+} − 8.9)
BF4	max (0, 14.2 − Na^+)
BF5	max (0, 0.8 − NO_2^-) × max (0, 16.9 − Na^+)
BF6	max (0, NO_3^- − 0.76)
BF7	max (0, Ca^{2+} − 18.6)
BF8	max (0, 0.8 − NO_2^-) × max (0, 0.34 − PO_4^{3-})
BF9	max (0, 18.6 − Ca^{2+}) × max (0, 0.49 − NO_2^-)
BF10	max (0, 12 − Na^+)
BF11	max (0, 1.29 − NO_3^-)
BF12	max (0, NO_2^- − 0.36)
BF13	max (0, 0.36 − NO_2^-)
BF14	max (0, 1.29 − NO_3^-) × max (0, 5.947 − Ca^{2+})
BF15	max (0, 18.6 − Ca^{2+}) × max (0, 18.4 − Na^+)
COD	
BF1	max (0, 26.6 − Ca^{2+})
BF2	max (0, Na^+ − 27.1)
BF3	max (0, 27.1 − Na^+)
BF4	max (0, 1.23 − NO_3^-) × max (0, 11.2 − Mg^{2+})
BF5	max (0, 27.1 − Na^+) × max (0, PO_4^{3-} − 1.85)
BF6	max (0, 27.1 − Na^+) × max (0, 1.85 − PO_4^{3-})
BF7	max (0, 2.51 − PO_4^{3-}) × max (0, 0.51 − NO_2^-)
BF8	max (0, 1.23 − NO_3^-) × max (0, 8.67 − Mg^{2+})
BF9	max (0, 0.39 − NO_2^-)
BF10	max (0, 1.53 − PO_4^{3-}) × max (0, NO_2^- − 0.18)
BF11	max (0, EC − 3) × max (0, 0.82 − NO_3^-)
BF12	max (0, 27.1 − Na^+) × max (0, 7.8 − pH)
BF13	max (0, 1.23 − NO_3^-) × max (0, 7.8 − pH)
BF14	max (0, 1.53 − PO_4^{3-}) × max (0, NO_3^- − 1.1)
BF15	max (0, 1.73 − PO_4^{3-})

Table 4 Results of ANOVA decomposition for the MARS approach

Function	GCV	Basis	Variable (s)
BOD5			
1	401.652	1	Ca^{2+}
2	192.133	2	Na^+
3	57.812	1	Turbidity
4	85.099	1	PO_4^{3-}
5	257.192	2	NO_3^-
6	87.913	2	NO_2^-
7	72.059	1	Ca^{2+} and Na^+
8	61.201	1	Ca^{2+} and NO_3^-
9	1191.542	2	Ca^{2+} and NO_2^-
10	338.459	1	Na^+ and NO_2^-
11	41.124	1	PO_4^{3-} and NO_2^-
COD			
1	111.503	1	Ca^{2+}
2	261.335	2	Na^+
3	316.990	1	PO_4^{3-}
4	131.690	1	NO_2^-
5	41.292	1	EC and NO_3^-
6	54.088	1	pH and Na^+
7	62.158	1	pH and NO_3^-
8	43.765	2	Mg^{2+} and NO_3^-
9	110.375	2	Na^+ and PO_4^{3-}
10	49.309	1	PO_4^{3-} and NO_3^-
11	1400.009	2	PO_4^{3-} and NO_2^-

Development of ANN

The artificial neural network is one of the most common AI models which has the capability to perceive the complicated nonlinear mathematical expressions particularly where there is no explicit formulation among the variables (Smith 1994). The general structure of an ANN has basically three various layers. The first layer is fed by the input variables where computations of their weighted sum are obtained. The second type of layer is known as the hidden layer where dataset processing is performed. The third one is the output layer, where the predicted outputs of ANN are provided and then the performance of the ANN can be evaluated. To obtain the best model of ANN, the number of neurons in the hidden layer, type of learning process, approach, and type of transfer functions are considered through a trial-and-error process. The optimal structure of ANN and its setting parameters have been defined by means of minimum values of MSE for the training stage. In fact, the lowest value of MSE related to the training stage was assigned as the criteria for the selection of the best ANN model (Basant et al. 2010).

In this study, for the BOD5 prediction, the proposed ANN includes an input layer with nine neurons and a single hidden layer with five neurons. This ANN model has been trained using the Levenberg–Marquardt (LM) algorithm, and additionally, the lowest value of MSE (46.14) was obtained. In the case of COD estimation, the proposed ANN consists of nine neurons in the input (or first) layer, one hidden layer with five neurons, and one neuron in the output layer. The lowest value of MSE was 36.48. Moreover, for both proposed ANN models, linear transfer function (purelin) has been applied in the hidden layer and a nonlinear transfer function (tansig) was used in the output layer.

Development of ANFIS

Jang (1993) has introduced the ANFIS which is a combination of ANN and fuzzy logic. In fact, the ANFIS model has the capability to successfully integrate the training capability of ANNs into a fuzzy inference system (FIS). It can be effectively applied to estimate roughly each real continuous function on a compact set to every arbitrary level of precision (e.g., Jang et al. 1997; Emamgholizadeh et al. 2014). Through the ANFIS model, FIS operations are composed of if–then rules which can be categorized into three groups as Tsukamoto, Mamdani, and Sugeno systems. In the current study, the first-order Sugeno fuzzy model is applied due to the fact that this is the most frequent model used in engineering problems. ANFIS is generated using three methods, namely the combination of the least squares and backpropagation gradient descent approaches, subtractive clustering (SC), and fuzzy c-means (FCM) clustering approach to train the ANFIS model. In this study, the FCM approach has been employed to generate the Sugeno-type FIS structure. To use this model for training in the ANFIS model, the rule extraction can be used, and additionally, the number of rules and antecedent MFs are obtained. In the training process of the ANFIS model, an adjustable value of radii is required to define a general mathematic shape of subclustering function. As mentioned in the ANN section, the training process of the ANFIS model is evaluated using the lowest value of MSE. The subtractive clustering method assigned the Gaussian membership function within the training process for the initialization of fuzzy rules. For the prediction of BOD5 index, the best ANFIS model in terms of accuracy was structured using 150 fuzzy rules, radii value of 0.2, and MSE of 46.1. Furthermore, in the case of the COD parameter, the most efficient ANFIS model had a radii value of 0.2, 28 fuzzy rules, and MSE of 114.55.

Development of multiple regression equations

In this section, linear and nonlinear regression equations were fitted by means of training datasets. In this way, SPSS 16 software was applied to present multiple linear and nonlinear regression equations. In this study, multiple linear regression (MLR) equation is considered as follows,

$$\begin{aligned} WQI = {} & o_1 + (EC)^{o_2} + (pH)^{o_3} + \left(Ca^{2+}\right)^{o_4} + \left(Mg^{2+}\right)^{o_5} \\ & + (Na^{+})^{o_6} + (Turbidity)^{o_7} + \left(PO_4^{3-}\right)^{o_8} \\ & + \left(NO_3^{-}\right)^{o_9} + \left(NO_2^{-}\right)^{o_{10}} \end{aligned} \quad (17)$$

With the aid of the least square (LS) method, MLR equations for prediction of BOD5 and COD were obtained, respectively, as,

$$\begin{aligned} BOD5 = {} & 0.9633\, EC + 0.525\ pH + 0.1403\, Ca^{2+} \\ & + 0.1226\, Mg^{2+} + 0.1345\, Na^{+} + 0.2589\, Turbidity \\ & + 4.5232\, PO_4^{3-} + 1.1885\, NO_3^{-} + 11.0624\, NO_2^{-} \\ & + 5.5639 \end{aligned} \quad (18)$$

$$COD = 0.69971\,EC + 1.1731\,pH + 0.392053\,Ca^{2+} + 0.04651\,Mg^{2+} + 0.1554\,Na^{+} + 0.1187\,Turbidity + 4.684\,PO_4^{3-} + 0.20539\,NO_3^{-} + 5.6991\,NO_2^{-} + 5.267 \quad (19)$$

In the case of the multiple nonlinear regression (MNLR) technique, the general form of the equation presenting a mathematical relationship among input and output variables was introduced as,

$$WQI = \alpha_1 \times (EC)^{\alpha_2} \times (pH)^{\alpha_3} \times \left(Ca^{2+}\right)^{\alpha_4} \times \left(Mg^{2+}\right)^{\alpha_5} \times (Na^{+})^{\alpha_6} \times (Turbidity)^{\alpha_7} \times \left(PO_4^{3-}\right)^{\alpha_8} \times \left(NO_3^{-}\right)^{\alpha_9} \times \left(NO_2^{-}\right)^{\alpha_{10}} \quad (20)$$

Then, coefficients of equations were obtained by means of the LS model. Multiple nonlinear regression relationships for estimation of BOD5 and COD were expressed, respectively, as,

$$BOD5 = 3.313 \times (EC)^{0.1359} \times (pH)^{0.2987} \times \left(Ca^{2+}\right)^{0.2257} \times \left(Mg^{2+}\right)^{0.1697} \times (Na^{+})^{0.1367} \times (Turbidity)^{0.1681} \times \left(PO_4^{3-}\right)^{0.2761} \times \left(NO_3^{-}\right)^{0.6798} \times \left(NO_2^{-}\right)^{0.2484} \quad (21)$$

$$COD = 7.7119 \times (EC)^{0.1437} \times (pH)^{-0.3131} \times \left(Ca^{2+}\right)^{0.2515} \times \left(Mg^{2+}\right)^{0.1606} \times (Na^{+})^{0.2274} \times (Turbidity)^{-0.1973} \times \left(PO_4^{3-}\right)^{0.436} \times \left(NO_3^{-}\right)^{0.6596} \times \left(NO_2^{-}\right)^{0.2877} \quad (22)$$

Results and discussion

In this section, in the first place, quite a few statistical measures were introduced to evaluate the quantitative performance of the proposed models. Furthermore, results of the LS-SVM and MARS techniques are compared with those obtained using ANN, ANFIS, MLR, and MNLR models.

Definition of statistical parameters

To determine the best approach with an acceptable level of precision for both training and testing stages, some statistical criteria including correlation coefficient (R), root mean square error (RMSE), mean absolute error (MAE), and overall index of model performance (OI) were applied as (e.g., Mattar and Alamoud 2015; Olyaie et al. 2015; Chen et al. 2015),

$$R = \frac{\sum_{I=1}^{N}\left(WQI_{Pre}^{i} - WQI_{Pre}^{mean}\right)\left(WQI_{Obs}^{i} - WQI_{Obs}^{mean}\right)}{\sqrt{\sum_{i=1}^{N}\left(WQI_{Obs}^{i} - WQI_{Obs}^{mean}\right)^2 \sum_{i=1}^{N}\left(WQI_{Pre}^{i} - WQI_{Pre}^{mean}\right)^2}} \quad (23)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(WQI_{Obs}^{i} - WQI_{Pre}^{i}\right)^2}{N}} \quad (24)$$

$$MAE = \frac{\sum_{i=1}^{N}\left|WQI_{Pre}^{i} - WQI_{Obs}^{i}\right|}{N} \quad (25)$$

$$OI = \frac{1}{2}\left(2 - \frac{RMSE}{WQI_{Obs}^{max} - WQI_{Obs}^{min}} - \frac{\sum_{i=1}^{N}\left(WQI_{Obs}^{i} - WQI_{Pre}^{i}\right)^2}{\sum_{i=1}^{N}\left(WQI_{Obs}^{i} - WQI_{Obs}^{mean}\right)^2}\right) \quad (26)$$

Where WQI_{Obs} and WQI_{Pre} are the observed WQIs and predicted ones by the proposed models, respectively. WQI^{mean}, WQI^{min}, and WQI^{max} are indicative of the mean, minimum, and maximum values of WQI, respectively.

Additionally, in order to determine the best technique which has more efficient performance than other AI models, the objective function (OBJ) is applied as (Gandomi et al. 2010):

$$OBJ = \left(\frac{N_{Training} - N_{Testing}}{N_{Total}}\right)\frac{MAE_{Training}}{R^2{}_{Training}} + \frac{2N_{Testing} \times MAE_{Testing}}{N_{Total} \times R^2{}_{Testing}} \quad (27)$$

in which N_{Total} is the total number of observations, and $N_{Training}$ and $N_{Testing}$ are the number of samples devoted to the training and testing stages, respectively.

Evaluation of the AI proposed techniques

In this section, the performances of SVM and MARS to predict the BOD5 and COD were investigated for both training and testing stages. Results of statistical evaluations were given in Tables 5 and 6.

In the training phase, for the prediction of BOD5, LS-SVM developed by the RBF kernel function provided more accurate performance (R = 0.95 and RMSE = 3.357) in comparison with LS-SVM-Poly (R = 0.89 and RMSE = 4.71) and MARS (R = 0.89 and RMSE = 4.567). From Table 5, the MAE value (2.44) indicated the superiority of LS-SVM-RBF to the LS-SVM-Poly (MAE = 3.494)

and MARS (MAE = 3.594). Statistical parameters given by the MARS model are relatively the same as those obtained by LS-SVM-Poly. In fact, Table 5 demonstrated that LS-SVM-Poly had slightly higher precision with MAE of 3.494 and OI of 0.829 than that of MARS (MAE = 3.594 and OI = 0.838).

In the training stage, the LS-SVM-RBF model has provided COD prediction with a higher level of accuracy in terms of R (0.92) and RMSE (3.669) rather than the LS-SVM-Poly (R = 0.86 and RMSE = 4.768) and MARS (R = 0.88 and RMSE = 4.412). Moreover, MAE (2.57) and OI (0.869) values given by LS-SVM-RBF indicated a lower level of computational error when compared with LS-SVM-Poly (MAE = 3.559 and OI = 0.8) and MARS (MAE = 3.349 and OI = 0.824). As seen in Table 5, it can be said that MARS predicted the COD index with relatively higher level of accuracy than those acquired by LS-SVM-Poly. Furthermore, qualitative performances of the proposed models for the training stage have been presented in Fig. 3a, b.

Results of the methods' performance for the testing stage are summarized in Table 6. From quantitative comparisons, LS-SVM-Ploy has provided BOD5 values with more relatively accurate prediction in terms of R (R = 0.85) and RMSE (5.463) in comparison with the LS-SVM-RBF (R = 0.83 and RMSE = 5.725). Furthermore, the OI value given by LS-SVM-Ploy indicated a slightly better performance than the LS-SVM-RBF (OI = 0.761). Equation (15) obtained by MARS has produced BOD values with relatively lower computational error (RMSE = 6.719 and MAE = 5.399) rather than LS-SVM-RBF (RMSE = 5.725 and MAE = 3.959) and LS-SVM-Poly (RMSE = 5.463 and MAE = 4.508). Also, according to R and OI values, both LS-SVM models have the capability of estimating the BOD5 index in comparison with the MARS [Eq. (15)].

In the case of COD prediction, results of performance testing related to the LS-SVM-RBF indicated a higher level of accuracy in terms of RMSE (4.461) and MAE (3.165) compared with LS-LS-Ploy (RMSE = 4.49 and MAE = 3.399) and MARS (RMSE = 5.306 and MAE = 4.045). With respect to R and OI, Table 6 demonstrates that both LS-SVM models had the same performance in the prediction of COD. Furthermore, Eq. (16) extracted from the MARS model had relatively larger computational error (R = 0.82 and OI = 0.719) in comparison with the LS-SVM-Ploy (R = 0.85 and OI = 0.788) and LS-SVM-RBF (R = 0.85 and OI = 0.79). Qualitative comparisons of the performance of the AI models for the testing stage are depicted in Fig. 4a, b.

Comparisons of the proposed models with artificial intelligence models

In this section, the performances of the LS-SVM and MARS models were compared with those obtained using the ANN and ANFIS approaches. In the case of BOD5 prediction, the LS-SVM-RBF technique had better performance in terms of RMSE (5.725) and MAE (3.959) in comparison with the ANN (RMSE = 6.946 and MAE = 5.94) and ANFIS (RMSE = 6.118 and MAE = 4.727). Furthermore, the superiority of the LS-SVM-Poly model to the ANN and ANFIS approaches had the same pattern as that of the LS-SVM-RBF. As seen in Table 6, Eq. (15) given by the MARS model gave slightly more accurate prediction of the BOD5 parameter (R = 0.79 and OI = 0.688) rather than the ANN model (R = 0.74 and OI = 0.671). The ANFIS model has produced relatively lower computational error of the BOD5 prediction with MAE of 6.118 and OI of 0.733 compared to the MARS approach (MAE = 4.727 and OI = 0.688).

Statistical parameters presented in Table 6 demonstrated that the ANFIS model has provided prediction of the COD with the lowest level of precision (R = 0.63 and RMSE = 10.703) when compared with the other AI approaches. The performance of the LS-SVM-RBF (R = 0.85 and RMSE = 4.461) and LS-SVM-Poly (R = 0.85 and RMSE = 4.491) indicated a higher level of accuracy rather than ANN (R = 0.75 and RMSE = 5.688). Moreover, the MARS approach with RMSE of 5.306 and OI of 0.719 had better performance in the prediction of COD rather than the ANN (RMSE = 5.688 and OI = 0.684) and ANFIS (RMSE = 10.703 and OI = 0.38). Qualitative performances of the ANN and ANFIS models are illustrated in Fig. 5a, b.

Comparisons of the proposed models with multiple regression-based equations

In this section, quantitative and qualitative comparisons of the AI models with the MLR and MNLR approaches were given. In the case of the BOD5 prediction, Eq. (18) given by MLR provided the largest computational error in terms of RMSE (15.775) and MAE (14.52) when compared with the artificial intelligence models. Similarly, Eq. (21) extracted from the MNLR model had the lowest level of precision in the prediction of

BOD5 with RMSE of 20.871 and MAE of 15.40 rather than the other proposed models. The R value (0.59) obtained by Eq. (21) has stood at the minimum level of performance in comparison with the LS-SVM-RBF ($R = 0.83$) and LS-SVM-Poly ($R = 0.85$). Moreover, from Table 6, it can be said that the performances of the MLR (OI = − 0.391) and NMLR (OI = − 1.333) models have no superiority to the ANN (OI = 0.671), ANFIS (OI = 0.733), and MARS (OI = 0.688).

In the case of COD prediction, Eq. (19) given by the multiple linear regression technique has provided the largest level of computation error with RMSE of 19.76 and MAE of 18.93 when compared with the LS-SVM and MARS models. Similarly, with respect to the values of RMSE and MAE, the MLR model [Eq. (19)] had significantly poor performance in comparison with ANN (RMSE = 5.688 and MAE = 4.821) and ANFIS (RMSE = 10.703 and MAE = 7.554). Moreover, Eq. (22) given by the MNLR approach has produced the COD predictions slightly more precise (RMSE = 8.221 and MAE = 6.067) than those obtained using the ANFIS model (RMSE = 10.703 and MAE = 7.554). Statistical parameters given in Table 6 demonstrated that the performance of LS-SVM-RBF ($R = 0.85$ and OI = 0.79), LS-SVM-Poly ($R = 0.85$ and OI = 0.788), and MARS ($R = 0.85$ and OI = 0.79) had superiority to that acquired by the MNLR technique ($R = 0.64$ and OI = 0.401). An illustrative performance of the MLR and MNLR models for both BOD5 and COD parameters is illustrated in Fig. 5a, b.

Evaluation of OBJ criterion for the proposed models

Values of OBJ are presented in Table 7. From Table 7, it can be said that LS-SVM-Poly (OBJ = 5.291) and LS-SVM-RBF (OBJ = 4.176) demonstrated an acceptable level of accuracy in the prediction of BOD5 values, compared with the other models. OBJ values given by MARS (OBJ = 6.526) and ANFIS (OBJ = 6.521) indicated that these two artificial intelligence techniques have the same performance. Furthermore, OBJ values obtained by MLR (OBJ = 26.09) and MNLR (OBJ = 38.73) resulted in poorer performance than the machine learning models. In the case of COD prediction, results of OBJ showed that LS-SVM-RBF (OBJ = 3.649) had more precise prediction than LS-SVM-Poly (OBJ = 4.695) and MARS (OBJ = 5.211). Table 7 indicated that MARS model predicted the COD index with higher level of accuracy in comparison with the ANN (OBJ = 8.158) and ANFIS (OBJ = 10.47). Additionally, OBJ values given by MLR (OBJ = 34.25) and MNLR (OBJ = 13.91) were indicative of being the largest error of COD prediction compared with the other approaches.

Table 5 Evaluation of the proposed models in the training phase

Model	R	RMSE	MAE	OI
BOD5				
LS-SVM-RBF	0.95	3.357	2.440	0.900
LS-SVM-Poly	0.89	4.710	3.494	0.829
MARS	0.89	4.567	3.594	0.838
COD				
LS-SVM-RBF	0.92	3.669	2.570	0.869
LS-SVM-Poly	0.86	4.768	3.559	0.800
MARS	0.88	4.412	3.449	0.824

Table 6 Evaluation of the proposed models in the testing phase

Model	R	RMSE	MAE	OI
BOD5				
LS-SVM-RBF	0.83	5.725	3.959	0.761
LS-SVM-Poly	0.85	5.463	4.508	0.778
MARS	0.79	6.719	5.399	0.688
ANN	0.74	6.946	5.940	0.671
ANFIS	0.81	6.118	4.727	0.733
MLR	0.78	15.775	14.52	− 0.391
MNLR	0.59	20.871	15.40	− 1.333
COD				
LS-SVM-RBF	0.85	4.461	3.165	0.790
LS-SVM-Poly	0.85	4.491	3.399	0.788
MARS	0.82	5.306	4.045	0.719
ANN	0.75	5.688	4.821	0.684
ANFIS	0.63	10.703	7.554	0.380
MLR	0.75	19.76	18.93	− 2.00
MNLR	0.64	8.221	6.067	0.401

Comparative results of the Fisher test for the proposed approaches

In this section, the ANOVA technique has been employed to evaluate the statistical reliability and significance related to the proposed models. To appraise the hypothesis asserting that the value of variation expressed by the regression model is higher than the

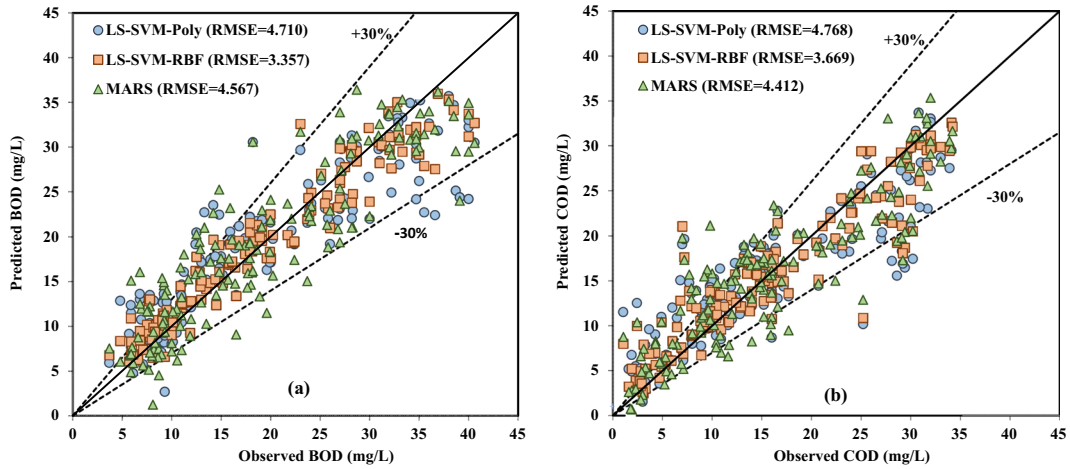

Fig. 3 Illustrative performances of proposed models for the training stage: **a** prediction of BOD5 and **b** prediction of COD

variation expressed by the averages, the F ratio has been applied. In the Fisher test, it is claimed that the null hypothesis is accepted if $F_0 > F_{\propto,\, k,\, n-p}$, where $\propto$ is the significant level, k is the number of independent variable, p is the $k+1$, and n is the number of datasets. For all the proposed artificial intelligence approaches and multiple regression-based equations, values of $\propto$, k, and $n-p$ are fixed at 0.05, 9, and 40, respectively. Therefore, the $F_{0.05,9,\,40}$ value is 2.1240 with the aid of the F distribution table. Furthermore, formulation of F_0 is expressed as (Hair et al. 1995),

$$F_0 = \frac{MS_R}{MS_E} \tag{28}$$

In which MS_R is the regression mean square and MS_E is the error mean square. MS_R and MS_E are computed as,

$$\mathrm{MS_R} = \frac{\mathrm{SSR}}{k} \tag{29}$$

$$\mathrm{MS_E} = \frac{\mathrm{SSE}}{n-p} \tag{30}$$

In which SSR is the sum of squares regression and SSE is the sum of squares of error. These parameters are calculated as,

$$\mathrm{SSR} = \sum_{i=1}^{n} \left(\mathrm{WQI}^{i}_{\mathrm{Pre}} - \mathrm{WQI}^{i}_{\mathrm{Obs}}\right)^2 \tag{31}$$

$$SSE = \sum_{i=1}^{n} \left(WQI^{i}_{Pre} - \overline{WQI}_{Obs}\right)^2 \tag{32}$$

Statistical performance of the F test is detailed in Table 8. For the BOD5 prediction, with respect to

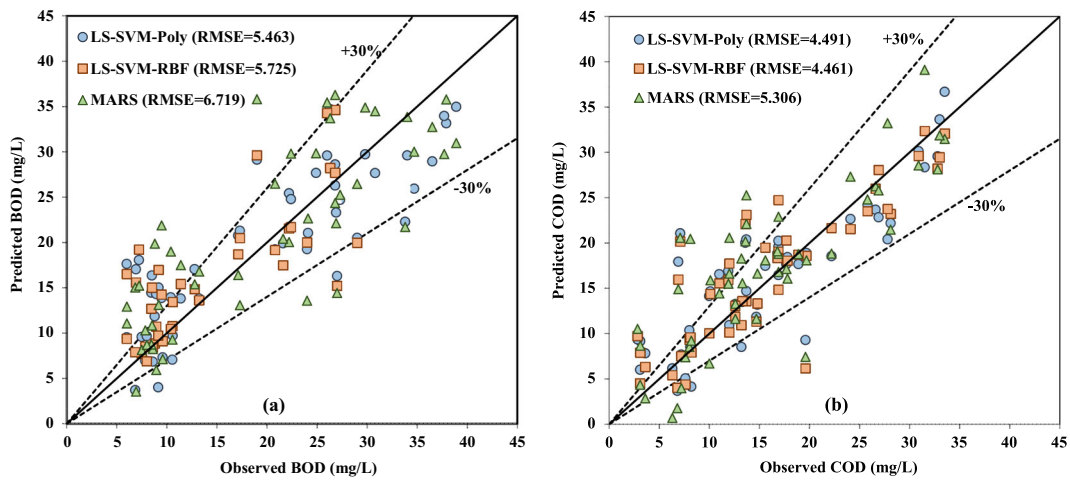

Fig. 4 Qualitative performances of the proposed models for the testing stage: **a** prediction of BOD5 and **b** prediction of COD

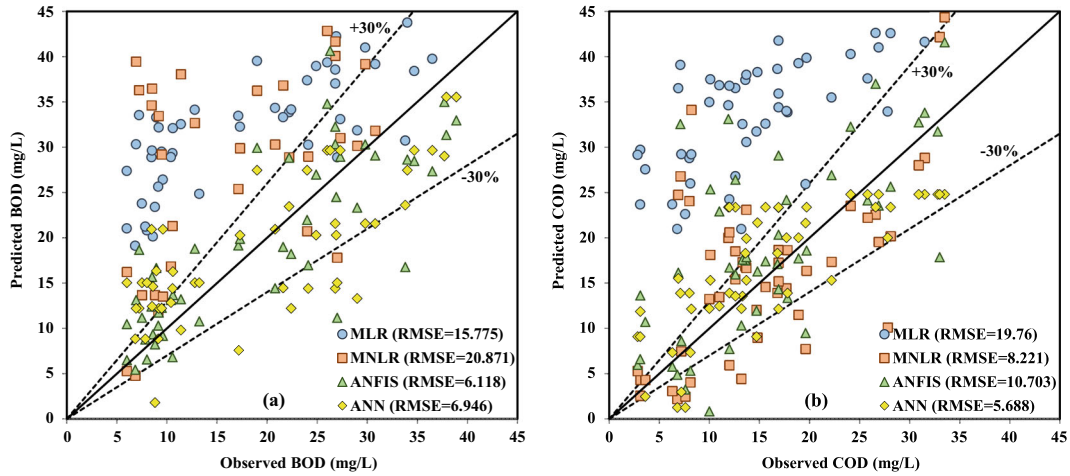

Fig. 5 Qualitative comparisons of the ANN, ANFIS, MLR, and MNLR for testing the stage: **a** prediction of BOD5 and **b** prediction of COD

$F_{0.05,9,\ 40}$, F_0 given by the LS-SVM models and the MARS approach has accepted the hypothesis and additionally showed a precise prediction of the BOD5 index compared with multiple regression-based equations. As seen in Table 8, the performance of the proposed ANN model with F_0 of 3.697 has rejected the hypothesis of the F test. The MNLR model [Eq. (21)] provided a large amount of F_0 (10.596) rather than $F_{0.05,9,\ 40}$ value, and additionally, this method is indicative of the lower capability in BOD5 prediction. Furthermore, in the case of COD evaluation, Table 8 indicates that LS-SVM-RBF ($F_0 = 1.48$), LS-SVM-Poly ($F_0 = 1.6$), and MARS ($F_0 = 1.66$) have accepted the hypothesis of F test with the significant level of 0.05 rather than ANFIS, ANN, and multiple regression equations. In fact, F_0 obtained by the proposed models rejecting F test hypothesis was stood at the slightly lower level of $F_{0.05,9,\ 40}$.

Table 7 Performance of the OBJ criterion for the proposed models

Model	OBJ
BOD5	
LS-SVM-Poly	5.291
LS-SVM-RBF	4.176
MARS	6.526
ANN	10.02
ANFIS	6.521
MLR	26.09
MNLR	38.73
COD	
LS-SVM-Poly	4.695
LS-SVM-RBF	3.649
MARS	5.211
ANN	8.158
ANFIS	10.47
MLR	34.25
MNLR	13.91

External validation and evaluation of the proposed approaches

External validation is the way of conducting a fair comparison between the predicted values and the actual ones. Performance of external validation is evaluated by means of the testing datasets and then several criteria are computed (Golbraikh and Tropsha 2002; Sattar 2014). All the criteria need to be validated in a specific range, and these criteria should be calculated for all the proposed artificial intelligence methods and multiple regression equations. To conduct external validation, in the first place, the gradient of the regression lines among the predicted WQIs and the observed ones can be computed as,

$$K = \sum_{i=1}^{n} \left(\mathrm{WQI}^{i}_{\mathrm{Obs}} \times \mathrm{WQI}^{i}_{\mathrm{Pre}}\right) / \overline{\mathrm{WQI}}^{2}_{\mathrm{Pre}} \tag{33}$$

Also, the gradient value given by Eq. (33) can be obtained from the observed WQIs and predicted ones as,

$$K' = \sum_{i=1}^{n} \left(\mathrm{WQI}^{i}_{\mathrm{Obs}} \times \mathrm{WQI}^{i}_{\mathrm{Pre}}\right) / \overline{\mathrm{WQI}}^{2}_{\mathrm{Obs}} \tag{34}$$

At least, one of Eq. (33) or Eq. (34) can be evaluated through the external validation process. The valid range of K (or K') is between 0.85 and 1.15 according to the Golbraikh and Tropsha (2002) investigations. Furthermore, the most widely acceptable value of K (or K') is approximately close to 1. Secondly, indices m and n are required to become lower than 0.1 as,

$$\mathrm{m} = \left(\mathrm{R}^2 - \mathrm{R}_0^2\right)/\mathrm{R}^2 \tag{35}$$

$$\mathrm{n} = \left(\mathrm{R}^2 - R_0'^2\right)/\mathrm{R}^2 \tag{36}$$

Roy and Roy (2008) proposed a confirmation index for the external predictability of the proposed models (R_m) in a way that the satisfying performance is met when $R_m > 0.5$. The R_m index is computed as,

$$\mathrm{R_m} = \mathrm{R}^2 \times \left(1 - \sqrt{\left|\mathrm{R}^2 - \mathrm{R}_0^2\right|}\right) > 0.5 \tag{37}$$

in which R_0^2 is introduced as the squared correlation coefficients between the predicted values and the observed ones, and $R_0'^2$ is the squared correlation coefficients between the observed values and the predicted ones. Moreover, one of the R_0^2 and $R_0'^2$ is required to be close to 1 for the selection of the best model. Formulations of the two mentioned squared correlation coefficients are expressed as,

$$R_0^2 = 1 - \sum_{i=1}^{n} \mathrm{WQI}_{\mathrm{Pre}}^{i\,2}(1-K)^2 \Big/ \sum_{i=1}^{n} \left(\mathrm{WQI}_{\mathrm{Pre}}^{i} - \mathrm{WQI}_{\mathrm{Pre}}^{\mathrm{mean}}\right)^2 \tag{38}$$

$$R_0'^2 = 1 - \sum_{i=1}^{n} \mathrm{WQI}_{\mathrm{Obs}}^{i\,2}\left(1-K'\right)^2 \Big/ \sum_{i=1}^{n} \left(\mathrm{WQI}_{\mathrm{Obs}}^{i} - \mathrm{WQI}_{\mathrm{Obs}}^{\mathrm{mean}}\right)^2 \tag{39}$$

All the criteria for the evaluation of external validation related to the proposed artificial intelligence models and multiple regression-based equations are given in Table 9. According to Table 9, in the case of BOD5 prediction, all the criteria given by LS-SVM-Ploy (K = 0.99 and R_m = 0.51) and LS-SVM-RBF (K = 0.97 and R_m = 0.54) were compromisingly satisfied to meet the best conditions for the selection of the most accurate approach when compared with the other techniques. Moreover, the criterion of the R_m value for MARS (R_m=0.47) was obtained slightly lower than 0.5, and consequently, this condition was met. For the ANN model, even though values of K, K', n, and m were in agreement with the required conditions, R (0.74) and R_m (0.39) values were not validated in the permissible

Table 8 Analysis of variance for WQI prediction

Model	MS_R	MS_E	F_0	State of hypothesis
BOD5				
LS-SVM-Poly	165.803	89.618	1.85	Accept
LS-SVM-RBF	182.143	94.98	1.917	Accept
MARS	250.817	136.82	1.83	Accept
ANN	268.071	72.505	3.697	Reject
ANFIS	207.99	106.42	1.954	Accept
MLR	199.129	86.944	2.29	Reject
MNLR	906.875	85.581	10.596	Reject
COD				
LS-SVM-Poly	112.08	70.051	1.6	Accept
LS-SVM-RBF	110.6	74.68	1.48	Accept
MARS	156.43	93.95	1.66	Accept
ANN	179.79	61.65	2.916	Reject
ANFIS	636.42	235.55	2.702	Reject
MLR	57.67	126.73	2.197	Reject
MNLR	375.52	134.78	2.786	Reject

ranges. With respect to R and R_m values, the ANFIS model had the capability to predict the BOD5 index with a permissible level of validated criteria in comparison with the ANN approach. Ultimately, Table 9 demonstrates that all the criteria values for both MLR and MNLR equations were not met to select them as the most accurate model. For instance, m and n values given by MLR [Eq. (17)] were obtained at 0.54 and 6.15, respectively, while for MNLR [Eq. (19)], the values were 1.1 and 1.79.

In addition, Table 9 indicates that the proposed LS-SVM models can provide more accurate values of the COD index on the basis of criteria values. For example, m and n values given by LS-SVM-RBF were fixed at −0.34 and −0.36, respectively, and for LS-SVM-Poly, both m and n were obtained at −0.36. The MARS model has stood at the third rank in terms of precision level ($R = 0.82$, $K = 0.89$, and $K' = 1.04$) just after the LS-SVM models. ANN was not capable of estimating the COD with an acceptable R (0.75) and R_m (0.41) as external validation criteria. As seen in Table 9, the criteria of R, R_m, and K values for the proposed ANFIS model were fixed at 0.63, 0.38, and 0.71, respectively. In fact, these criteria did not have permissible values so that the ANFIS model was assigned as the best model. Table 9 indicates that both MLR and MNLR equations did not have an acceptable level of external validation for the criteria of R, R_m, and K. Overall, statistical indices in Table 9 are indicative of the higher performance of the LS-SVM models and MARS in the estimation of BOD5 and COD indices.

Conclusion

In the present research work, various artificial intelligence approaches, such as machine learning classifiers, were investigated to assess water quality indices. In this way, LS-SVM and MARS techniques were implemented to estimate BOD5 and COD parameters; 200 samples taken from Karoun River, located in the southwest of Iran, have been used to conduct training and testing stages. Through this study, nine independent variables were determined as input variables for the development of the proposed approaches. The LS-SVM model was improved by four types of kernel functions including linear, polynomial, RBF, and sigmoid. Results of LS-SVM performance indicated that the LS-SVM technique with polynomial and RBF kernel functions have stood at higher level of accuracy rather than the LS-SVM approaches developed by other types of kernel functions. Results of training and testing phases

Table 9 External validation statistical measures for WQIs in prediction models

Model	R ($R > 0.8$)	K ($0.85 < K < 1.15$)	K' ($0.85 < K' < 1.15$)	m ($m < 0.1$)	n ($n < 0.1$)	R_m ($R_m > 0.5$)
BOD5						
LS-SVM-RBF	0.83	0.97	0.95	−0.42	−0.42	0.51
LS-SVM-Poly	0.85	0.99	0.94	−0.37	−0.36	0.54
MARS	0.79	0.92	0.98	−0.54	−0.58	0.47
ANN	0.74	1.02	0.87	−0.79	−0.59	0.39
ANFIS	0.81	0.97	0.94	−0.51	−0.49	0.50
MLR	0.78	0.59	1.49	0.54	6.15	0.25
MNLR	0.59	0.51	1.58	1.100	1.79	0.13
COD						
LS-SVM-RBF	0.85	0.95	0.99	−0.34	−0.36	0.55
LS-SVM-Poly	0.85	0.96	0.98	−0.36	−0.36	0.56
MARS	0.82	0.89	1.04	−0.38	−0.46	0.50
ANN	0.75	0.94	0.95	−0.72	−0.73	0.41
ANFIS	0.63	0.71	1.079	−0.59	−1.47	0.38
MLR	0.75	0.472	1.84	1.40	28.85	0.16
MNLR	0.64	0.82	1.00	−1.09	−1.42	0.37

indicated that LS-SVM-RBF and LS-SVM-Poly methods have provided relatively precise prediction for BOD5 and COD indices, compared to the MARS models. Moreover, the performances of the proposed models were compared with those obtained using ANN, ANFIS, MLR, and MNLR. For the prediction of BOD5 and COD parameters, the LS-SVM and MARS models [Eqs. (15) and (16)] had better performance in terms of external validation criteria and *F* test in comparison with ANN, ANFIS, and multiple regression-based equations. To put it another way, the proposed LS-SVM-RBF and LS-SVM-Poly approaches had comparatively successful performance in terms of statistical measures compared with the MARS. ANN and ANFIS models, as a representation of black-box systems, did not have the capability to give an accurate prediction as well as the machine learning models. From qualitative results, it should be noted that the performance of MLR and MNLR was highly dependent on a range of variables and, consequently, led to a high level of overprediction.

Generally, it can be said that LS-SVM and MARS models used in the current study demonstrated satisfactory assessments to present comparatively precise approximates in one of the most attention-grabbing subjects in the field of environmental sciences. On the other hand, explicit equations given by the MARS techniques can be practically employed by environmentalists in order to evaluate the water quality of natural streams.

As a demerit, there were two main shortcomings rooted from the datasets in the present research. The first one was related to the number of data samples. It takes a huge amount of time and effort to take more dataset samples from Karoun River within further field investigations. In this way, it causes to perform the proposed AI models using a broad range of data samples, and additionally, it can obtain not only a more accurate prediction of water quality indices but also a higher level of generalization. The second shortcoming corresponds to the number of parameters. In fact, in this study, some water quality parameters such as TDS, TS, and TSS were not considered as input or output variables due to either barriers of collecting data samples from Karoun River or the lack of experimental facilities for measurement of different WQIs.

As a framework for future research works, it is possible to apply the group method of data handling (GMDH), which is one of the most powerful self-organizing approaches, to predict water quality indices. In fact, the GMDH model can easily combine with evolutionary algorithms such as genetic algorithm (GA), PSO, and harmonic search (HS). GMDH has the capability of presenting explicit equations for the prediction of WQIs.

References

Adnan, R. M., Yuan, X., Kisi, O., & Yuan, Y. (2017a). Streamflow forecasting using artificial neural network and support vector machine models. *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS), 29*(1), 286–294.

Adnan, R. M., Yuan, X., Kisi, O. and Anam, R. (2017b). Improving accuracy of river flow forecasting using LSSVR with gravitational search algorithm. *Advances in Meteorology*, 2017.

Alizadeh, M. J., & Kavianpour, M. R. (2015). Development of wavelet-ANN models to predict water quality parameters in Hilo Bay, Pacific Ocean. *Marine Pollution Bulletin, 98*(1), 171–178.

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.

Azamathulla, H. M., & Ghani, A. A. (2011). Genetic programming for predicting longitudinal dispersion coefficients in streams. *Water Resources Management, 25*(6), 1537–1544.

Basant, N., Gupta, S., Malik, A., & Singh, K. P. (2010). Linear and nonlinear modeling for simultaneous prediction of dissolved oxygen and biochemical oxygen demand of the surface water—a case study. *Chemometrics and Intelligent Laboratory Systems, 104*(2), 172–180.

Chau, K., & Wu, C. (2010). A hybrid model coupled with singular spectrum analysis for daily rainfall prediction. *Journal of Hydroinformatics, 12*(4), 458–473.

Chen, Z., Shi, R., & Zhang, S. (2013). An artificial neural network approach to estimate evapotranspiration from remote sensing and AmeriFlux data. *Frontiers of Earth Science, 7*(1), 103–111.

Chen, X., Chau, K., & Busari, A. (2015). A comparative study of population-based optimization algorithms for downstream river flow forecasting by a hybrid neural network model. *Engineering Applications of Artificial Intelligence, 46*, 258–268.

Cortes, C., & Vapnik, V. (1995). Support vector networks. *Machine Learning, 20*, 273–297.

Deininger, R. A., Lee, J., & Clark, R. M. (2011). Rapid detection of bacteria in drinking water and water contamination case studies. *Frontiers of Earth Science, 5*(4), 378–389.

Dogan, E., Sengorur, B., & Koklu, R. (2009). Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique. *Journal of Environmental Management, 90*(2), 1229–1235.

Ebtehaj, I., & Bonakdari, H. (2014). Performance evaluation of adaptive neural fuzzy inference system for sediment transport in sewers. *Water Resources Management, 28*(13), 4765–4779.

Emamgholizadeh, S., Kashi, H., Marofpoor, I., & Zalaghi, E. (2014). Prediction of water quality parameters of Karoon River (Iran) by artificial intelligence-based models.

International Journal of Environmental Science and Technology, 11(3), 645–656.

Fallah-Mehdipour, E., Bozorg Haddad, O., & Mariño, M. (2013). Prediction and simulation of monthly groundwater levels by genetic programming. *Journal of Hydro-Environment Research, 7*(4), 253–260.

Feng, C., Wu, F., Zhao, X., Li, H., & Chang, H. (2012). Water quality criteria research and progress. *Science China Earth Sciences, 55*(6), 882–891.

Gandomi, A. H., Alavi, A. H., Sahab, M. G., & Arjmandi, P. (2010). Formulation of elastic modulus of concrete using linear genetic programming. *Journal of Mechanical Science and Technology, 24*(6), 1273–1278.

Gholami, V., Chau, K., Fadaee, F., Torkaman, J., & Ghaffari, A. (2015). Modeling of groundwater level fluctuations using dendrochronology in alluvial aquifers. *Journal of Hydrology, 529*, 1060–1069.

Golbraikh, A., & Tropsha, A. (2002). Beware of q2! *Journal of Molecular Graphics & Modelling, 20*(4), 269–276.

Hair, J. F., Anderson, R. E., Tathan, R. L., & Black, W. (1995). *Multivariate data analysis with readings. 4th edn.* Englewood Cliffs: Prentice Hall.

Heddam, S., & Kisi, O. (2018). Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. *Journal of Hydrology, 559*, 499–509.

Jang, J. S. R. (1993). ANFIS: adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics, 23*, 665–685.

Jang, J. S. R., Sun, C. T., & Mizutani, E. (1997). *Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence*. Upper Saddle River: Prentice Hall.

Kisi, O. (2015). Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. *Journal of Hydrology, 528*, 312–320.

Kumar, P. S., Praveen, T., & Prasad, M. A. (2016). Artificial neural network model for rainfall-runoff—a case study. *International Journal of Hybrid Information Technology, 9*(3), 263–272.

Laucelli, D., & Giustolisi, O. (2011). Scour depth modelling by a multi-objective evolutionary paradigm. *Environmental Modelling and Software, 26*(4), 498–509.

Laucelli, D., Berardi, L., Doglioni, A. and Giustolisi, O. (2012). EPR-MOGA-XL: an excel based paradigm to enhance transfer of research achievements on data-driven modeling. Proceedings of 10th international conference on hydroinformatics HIC 2012, 14-18 July, Hamburg, Germany, R. Hinkelmann, M.H. Nasermoaddeli, S.Y. Liong, D. Savic, P. Fröhle (Eds).

Li, X., Sha, J. and Wang, Z.L (2016). A comparative study of multiple linear regression, artificial neural network and support vector machine for the prediction of dissolved oxygen. *Hydrology Research*, nh2016149.

Mahmoudi, N., Orouji, H., & Fallah-Mehdipour, E. (2016). Integration of shuffled frog leaping algorithm and support vector regression for prediction of water quality parameters. *Water Resources Management, 30*(7), 2195–2211.

Mattar, M. A., & Alamoud, A. I. (2015). Artificial neural networks for estimating the hydraulic performance of labyrinth-channel emitters. *Computers and Electronics in Agriculture, 114*, 189–201.

Mohammadpour, R., Shaharuddin, S., Chang, C. K., Zakaria, N. A., & Ab-Ghani, A. (2014). Spatial pattern analysis for water quality in free surface constructed wetland. *Water Science and Technology, 70*(7), 1161–1167.

Mohammadpour, R., Shaharuddin, S., Zakaria, N., Ghani, A., Vakili, M., & Chan, N. (2016). Prediction of water quality index in free surface constructed wetlands. *Environmental Earth Sciences, 75*(2), 1–12.

Najafzadeh, M., Laucelli, D. B. and Zahiri, A. (2016). Application of model tree and evolutionary polynomial regression for evaluation of sediment transport in pipes. *KSCE Journal of Civil Engineering*, 1–8.

Noori, R., Karbassi, A., Farokhnia, A., & Dehghani, M. (2009). Predicting the longitudinal dispersion coefficient using support vector machine and adaptive neuro-fuzzy inference system techniques. *Environmental Engineering Science, 26*(10), 1503–1510.

Noori, R., Karbassi, A., Moghaddamnia, A., Han, D., Zokaei-Ashtiani, M., Farokhnia, A., & Gousheh, M. G. (2011). Assessment of input variables determination on the SVM model performance using PCA, gamma test, and forward selection techniques for monthly stream flow prediction. *Journal of Hydrology, 401*(3), 177–189.

Noori, R., Karbassi, A., Ashrafi, K., Ardestani, M., Mehrdadi, N., & Bidhendi, G. R. N. (2012). Active and online prediction of BOD 5 in river systems using reduced-order support vector machine. *Environmental Earth Sciences, 67*(1), 141–149. https://doi.org/10.1007/s12665-011-1487-9.

Noori, R., Deng, Z., Kiaghadi, A., & Kachoosangi, F. T. (2015a). How reliable are ANN, ANFIS, and SVM techniques for predicting longitudinal dispersion coefficient in natural rivers? *Journal of Hydraulic Engineering, 142*(1), 04015039. https://doi.org/10.1061/(ASCE)HY.1943-7900.0001062.

Noori, R., Yeh, H. D., Abbasi, M., Kachoosangi, F. T., & Moazami, S. (2015b). Uncertainty analysis of support vector machine for online prediction of five-day biochemical oxygen demand. *Journal of Hydrology, 527*, 833–843. https://doi.org/10.1016/j.jhydrol.2015.05.046.

Olyaie, E., Banejad, H., Chau, K.-W., & Melesse, A. M. (2015). A comparison of various artificial intelligence approaches performance for estimating suspended sediment load of river systems: a case study in United States. *Environmental Monitoring and Assessment, 187*(4), 189.

Orouji, H., Bozorg Haddad, O., Fallah-Mehdipour, E., & Mariño, M. (2013). Modeling of water quality parameters using data-driven models. *Journal of Environmental Engineering, 139*(7), 947–957.

Rahimikhoob, A. (2016). Comparison of M5 model tree and artificial neural network's methodologies in modelling daily reference evapotranspiration from NOAA satellite images. *Water Resources Management, 30*(9), 3063–3075.

Roy, P. P., & Roy, K. (2008). On some aspects of variable selection for partial least 684 squares regression models. *QSAR and Combinatorial Science, 27*(3), 302–313.

Sattar, A. M. (2014). Gene expression models for the prediction of longitudinal dispersion coefficients in transitional and turbulent pipe flow. *Journal of Pipeline Systems Engineering and Practice, 5*(4013011), 1–10.

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.

Singh, K. K., Pal, M., & Singh, V. P. (2010). Estimation of mean annual flood in Indian catchments using backpropagation neural network and M5 model tree. *Water Resources Management, 24*(10), 2007–2019.

Smith, M. (1994). *Neural networks for statistical modelling* (p. 245). New York: Van Nostrand Reinhold.

Sreekanth, J., & Datta, B. (2011). Comparative evaluation of genetic programming and neural network as potential surrogate models for coastal aquifer management. *Water Resources Management, 25*(13), 3201–3218.

Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters, 9*(3), 293–300.

Taormina, R., Chau, K.-W., & Sivakumar, B. (2015). Neural network river forecasting through baseflow separation and binary-coded swarm optimization. *Journal of Hydrology, 529*, 1788–1797.

Verma, A., & Singh, T. (2013). Prediction of water quality from simple field parameters. *Environmental Earth Sciences, 69*(3), 821–829.

Wang, W.-C., Chau, K.-W., Xu, D.-M., Qiu, L., & Liu, C.-C. (2017). The annual maximum flood peak discharge forecasting using Hermite projection pursuit regression with SSO and LS method. *Water Resources Management, 31*(1), 461–477.

Xu, J., Chen, Y., Li, W., Peng, P. Y., Yang, Y., Wei, C., & Hong, Y. (2013). Combining BPANN and wavelet analysis to simulate hydro-climatic processes—a case study of the Kaidu River, North-west China. *Frontiers of Earth Science, 7*(2), 227–237.

Yaseen, Z. M., Ramal, M. M., Diop, L., Jaafar, O., Demir, V., & Kisi, O. (2018). Hybrid adaptive neuro-fuzzy models for water quality index estimation. *Water Resources Management., 32*(7), 2227–2245.

Zhang, W. G., & Goh, A. T. C. (2013). Multivariate adaptive regression splines for analysis of geotechnical engineering systems. *Computers and Geotechnics, 48*, 82–95.