JEE Journal of Ecological Engineering

Journal of Ecological Engineering 2022, 23(3), 236–250
https://doi.org/10.12911/22998993/146177
ISSN 2299–8993, License CC-BY 4.0

Received: 2021.12.29
Accepted: 2022.01.20
Published: 2022.02.01

Modeling Pollution Index Using Artificial Neural Network and Multiple Linear Regression Coupled with Genetic Algorithm

Iman Ali Abdulkareem[1], Abdulhussain A. Abbas[1*], Ammar Salman Dawood[1]

[1] Civil Engineering Department, College of Engineering, University of Basrah, Karmat Ali, Basra, Basra Governorate, 61004, Iraq

* Corresponding author's e-mail: abdulhussain.abbas@uobasrah.edu.iq

ABSTRACT

Shatt Al-Arab River in Basrah province, Iraq, was assessed by applying comprehensive pollution index (CPI) at fifteen sampling locations from 2011 to 2020, taking into consideration twelve physicochemical parameters which included pH, Tur., TDS, EC, TH, Na^+, K^+, Ca^{+2}, Mg^{+2}, Alk., SO_4^{-2}, and Cl^-. The effectiveness of multiple linear regression (MLR) and artificial neural network (ANN) for predicting comprehensive pollution index was examined in this research. In order to determine the ideal values of the predictor parameters that lead to the lowest CPI value, the genetic algorithm coupled with multiple linear regression (GA-MLR) was used. A multi-layer feed-forward neural network with backpropagation algorithm was used in this study. The optimal ANN structure utilized in this research consisted of three layers: the input layer, one hidden layer, and one output layer. The predicted equation of the comprehensive pollution index was created using the regression technique and used as an objective function of the genetic algorithm. The minimum predicted comprehensive pollution index value recommended by the GA-MLR approach was 0.3777.

Keywords: Shatt Al-Arab River, comprehensive pollution index, multiple linear regression, artificial neural network, genetic algorithm.

INTRODUCTION

Water is important for human and ecological survival and health in all aspects [Abyaneh, 2014]. According to the World Health Organization, water pollution is defined as any alteration in the physical, chemical, as well as biological characteristics of water which has a harmful impact on living beings [Salihu et al., 2017]. Water pollution is the primary cause of the water crisis. It must not be polluted to the point where it can no longer be utilized for irrigation and drinking [Singh et al., 2020]. The study of water quality provides a clear vision of the river's suitability for various uses [Al-Asadi et al., 2020].

The Shatt Al-Arab River (SAR) is the principal source of surface water in the Basrah governorate. The water supplier of the Shatt Al-Arab River comes from the Tigris and Euphrates rivers in Iraq, as well as the Karkheh and Karon rivers in Iran. Due to water scarcity, the Euphrates river was blocked as a supplier for the Shatt Al-Arab River, while Iran blocked off the waters of the Karon and Karkheh rivers from reaching Shatt Al-Arab. As a result, the Tigris river became the only supply of fresh water for Shatt Al-Arab [Al-Asadi and Alhello, 2019]. Due to the reason that the river and its branches have already become receptacles for pollutants from many sources, the river freshwater has been significantly degraded. As a result, monitoring the river pollution levels is critical for the human health in the area [Al-Asadi et al., 2020].

The neural networks technique has recently been used to a wide range of scientific fields. From the beginning in the 1990s, ANNs have been used in the fields of water engineering, and environmental sciences. When compared with conventional modeling methods, the artificial neural network is a suitable method having a flexible mathematical structure able of finding complicated nonlinear correlations among both the input and output data [Najah et al., 2013]. They are effectively utilized to predict water quality in a variety of water bodies [Kulisz et al., 2020]. In 1975, the basic concept of genetic algorithm has been first invented by John Holland when he was delivering a lecture called adapting systems theory at Michigan University [Azad et al., 2016]. The genetic algorithm is a method of searching that is dependent on Darwin's concept of evolution [Mijwel, 2016].

In this research, the comprehensive pollution index (CPI) was used to classify the Shatt Al-Arab River water pollution. Several researchers examined the water quality of the SAR [Dawood, 2017; Dawood et al., 2018; Hamdan et al., 2018; Al-Adhab et al., 2019; Dawood et al., 2020]. Researchers have implemented the Comprehensive Pollution Index to determine the water pollution [Yan et al., 2015; Mishra et al., 2016; Matta et al., 2018; Ezzat and Elkorashey, 2020; Son et al., 2020]. A number of studies on water quality prediction were performed using the ANN technique [Singh et al., 2009; Gazzaz et al., 2012; Abyaneh, 2014; Dawood et al., 2016; Hamdan and Dawood, 2016; Chen et al., 2019; Khudhur et al., 2020; Kulisz et al., 2021]. In a range of fields, many researchers have employed the genetic algorithm combined with multiple linear regression (GA-MLR) to solve optimization problems [Fisz, 2006; Zain et al., 2010; Goudarzi et al., 2012; Ghose and Samantaray, 2018; Guidea and Sarbu, 2019; Manroo and Ganiny, 2020].

The goals of the research are as follows: define the extent of water pollution in the SAR at many water treatment plants (WTPs) using the CPI, determine the optimum structure of the ANN, and determine the ideal values of the predictor parameters that lead to the lowest CPI value by using the GA-MLR method.

METHODOLOGY

Study area

The Shatt Al-Arab River rises at the confluence of the Euphrates and Tigris rivers in Qurna City and flows southwest for 101 kilometers before forming the border between Iraq and Iran for the final 91 kilometers of its main course, before flowing into the Arabian Gulf [Allafta and Opp, 2020]. The SAR lies between the latitude of (29° 45' 0" – 31° 15' 0" N) and the longitude of (47° 10' 20" – 48° 45' 0" E) [Abdulla, 2013]. The main water source in the Basrah province is the SAR, a natural river that flows through the Basrah governorate at a rate of 25–75 m^3/s [Almuktar et al., 2020]. The water quality of the Shatt Al-Arab has deteriorated dramatically during the last three decades caused by anthropogenic activities. The river is receiving growing volumes of untreated wastewater as well as runoff from the surrounding oil fields. As a result, the important functions the Shatt Al-Arab plays in maintaining healthy populations and sustaining a balanced ecology are considerably imperiled [Allafta and Opp, 2020].

Data description

The directorate of Basrah water provided monthly data on 12 water quality parameters collected at each of the fifteen water treatment plants throughout the period of 2011–2020. There are

Table 1. Descriptive statistics of physiochemical properties

Parameters	Unit	Minimum	Maximum	Mean	Std. deviation
pH	-	7.03	8.47	7.64	± 0.23
Tur.	NTU	0.60	79	15.62	± 8.42
TDS	mg/l	200	22954	3113.79	± 2655.54
EC	µs/cm	871	34030	4897.38	± 3952.51
TH	mg/l	296	4860	980.59	± 516.38
Na^+	mg/l	62	6780	713.15	± 794
K^+	mg/l	2.50	123	12.64	± 8.11
Ca^{+2}	mg/l	59	976	199.06	± 103.90
Mg^{+2}	mg/l	36	590	117.83	± 62.73
Alk.	mg/l	90	296	157.45	± 17.97
SO_4^{-2}	mg/l	134	4449	804.76	± 495.88
Cl^-	mg/l	104	10300	1118.04	± 1209.82

twelve parameters of water quality, which include pH, Tur., TDS, EC, TH, Na^+, K^+, Ca^{+2}, Mg^{+2}, Alk., SO_4^{-2}, and Cl^-. Table 1 illustrates the statistical analysis of twelve physical and chemical parameters for raw water in this study.

Sampling sites

The physical and chemical properties were obtained at fifteen water treatment plants. Table 2 presents the coordinates of various WTPs. Figure 1 shows the locations of the WTPs considered for this study.

Comprehensive Pollution Index (CPI)

CPI was used in several studies for the categorization of water quality. The steps for calculating CPI are [Ezzat and Elkorashey, 2020]:

- The following equation should be used to compute the pollution index (PI) for every water quality parameter [Ezzat and Elkorashey, 2020]:

Table 2. Coordinates of WTPs in this study

WTP No.	WTP Name	Latitude	Longitude
1	Al-Dear	30° 48’ 6.26» N	47° 34’ 52.46» E
2	Al-Houta	30° 39’ 0.54» N	47° 45’ 42.01» E
3	Al Basrah Unified	30° 38’ 52.56» N	47° 44’ 52.71» E
4	Al-Garmma 1	30° 34’ 18.27” N	47° 44’ 45.41” E
5	Al-Garmma 2	30° 34’ 32.43” N	47° 44’ 54.18” E
6	Al-Faiha	30° 34’ 10.45” N	47° 47’ 25.37” E
7	Al-Jubailah 1	30° 33’ 0.26” N	47° 48’ 44.87” E
8	Al-Ribat	30° 32’ 8.67” N	47° 49’ 49.28” E
9	Al-Bradhiah 1	30° 30’ 8.67” N	47° 51’ 20.43” E
10	Al-Bradhiah 2	30° 30’ 9.34” N	47° 51’ 22.41” E
11	Owaisyan	30° 29’ 34.77” N	47° 52’ 1.81” E
12	Mhejran	30° 28’ 22.69” N	47° 52’ 58.25” E
13	Hamdan Bridge	30° 27’ 52.20” N	47° 54’ 10.17” E
14	Maheilah	30° 27’ 50.52” N	47° 55’ 25.19” E
15	Al-Labanie	30° 27’ 48.50” N	47° 59’ 4.78” E

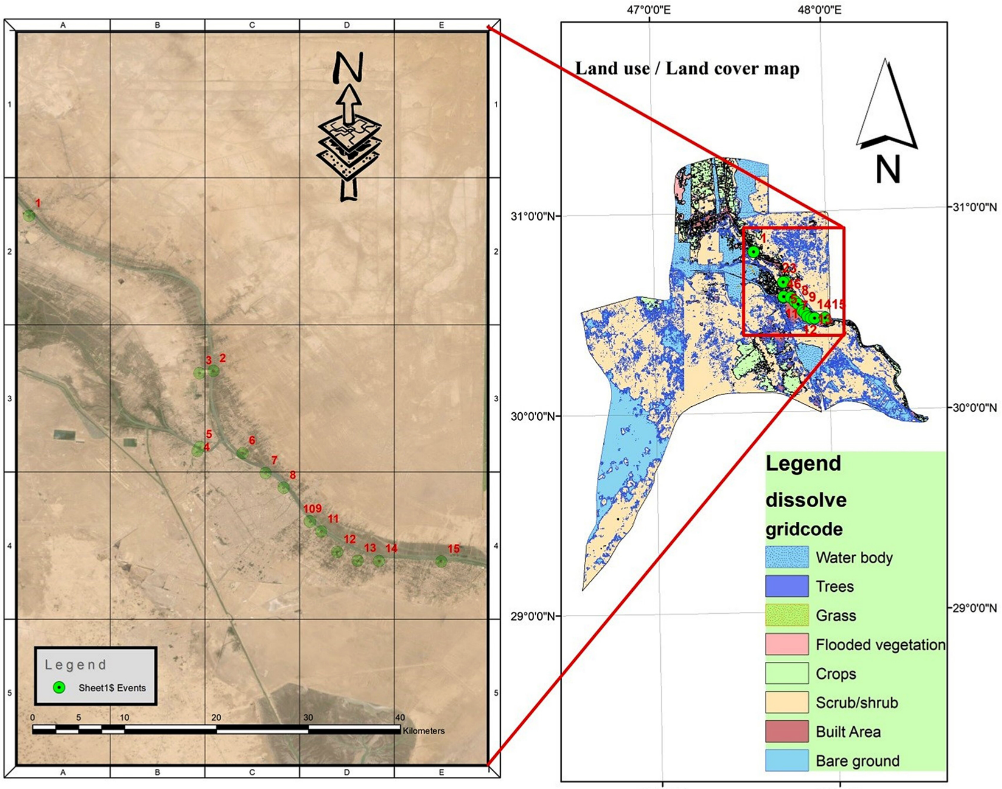

Figure 1. Locations of WTPs along the SAR in this study

$$PI = \frac{\begin{matrix}\mathit{Measured\ concentration\ of} \\ \mathit{specified\ water\ quality\ parametr}\end{matrix}}{\mathit{Standard\ permitted\ concentration}} \quad (1)$$

- The standard permitted concentrations for every parameter selected for this study were acquired from the World Health Organization (WHO 2011), as shown in Table 3 [Abbas et al., 2017; Ghalib, 2017; Mahmood et al., 2019; Ewaid and Abed, 2017].
- CPI was computed by taking the overall number of parameters into account [Ezzat and Elkorashey, 2020]:

$$CPI = \frac{1}{n}\sum_{i=1}^{n} PI \quad (2)$$

where n is the number of parameters that have been chosen.

- The CPI values could be utilized to categorize the water quality level, as shown in Table 4 [Matta et al., 2018].

Artificial Neural Network (ANN)

ANN is a mathematical programming model that mimics the functioning process of the human brain. An ANN method can perform brain processes, decide, arrive at a solution in the absence of sufficient data using current knowledge, absorb continuous data input, learn, and remember. The capability of a neural network to model complicated nonlinear relation sans making prior assumptions about the nature of the relation is its greatest advantage [Banejad and Olyaie, 2011]. An ANN is comprised of multiple nodes that represent neurons. The independent variables are represented by the input nodes, while the dependent variables are represented by the output nodes [Nwobi and Ochieze, 2018]. The main purpose of the learning procedure is to identify the best set of weights that can give the best output for the given inputs. The network output is compared to the target answer to calculate the error [Najah et al., 2013]. Different structures can be found in neural networks. Feed forward and recurrent networks can be distinguished in principle. Only forward-directed information flows from the input nodes through hidden nodes to the output nodes in feed forward networks. There are links in recurrent networks where information can travel forwards and backwards through network node connections. Feedback networks are another name for the recurrent networks [Mijwel and Alsaadi, 2019].

Table 3. Maximum permitted values for the parameters presented by WHO (2011)

Parameters	Units	WHO 2011
Tur.	NTU	5
pH	-	6.5-8.5
EC	µs/cm	1500
TDS	mg/l	1000
TH	mg/l	500
K^+	mg/l	12
Na^+	mg/l	200
Mg^{+2}	mg/l	100
Ca^{+2}	mg/l	75
Alk.	mg/l	200
Cl^-	mg/l	250
SO_4^{-2}	mg/l	250

Table 4. Classification of CPI

CPI value	Category
≤ 0.2	Clean water
0.21 – 0.40	Sub-clean water
0.41 – 1	Slightly polluted water
1.01 – 2.00	Moderately polluted water
≥ 2.01	Severely polluted water

Back Propagation Algorithm (BP)

Back propagation (BP) is the most common and widely applied learning algorithm over all neural network models among the various learning existing algorithms. This algorithm is employed in supervised learning [Banejad and Olyaie, 2011]. The primary training concept of BP is founded on gradient descent algorithm, which modifies weights to reduce Mean Square Error (MSE) [AlTobi et al., 2016]. The BP algorithm is divided in two phases: forward and backward phase. In the forward phase, the network input data is propagated to the following level and so forth. The network error is calculated after that. In the backward phase, the network error is propagated backwards, and the weights are adjusted accordingly [Gallo,2015]. As illustrated in Figure 2, the network structure is consists of three layers, each of which has n neurons.

The number of input variables determines the number of neurons in the first layer (input layer). This layer takes the input from external world and transfers them without any alteration to the hidden layer. Since they are only indirectly related to the outside environment, intermediate layers are usually known as hidden

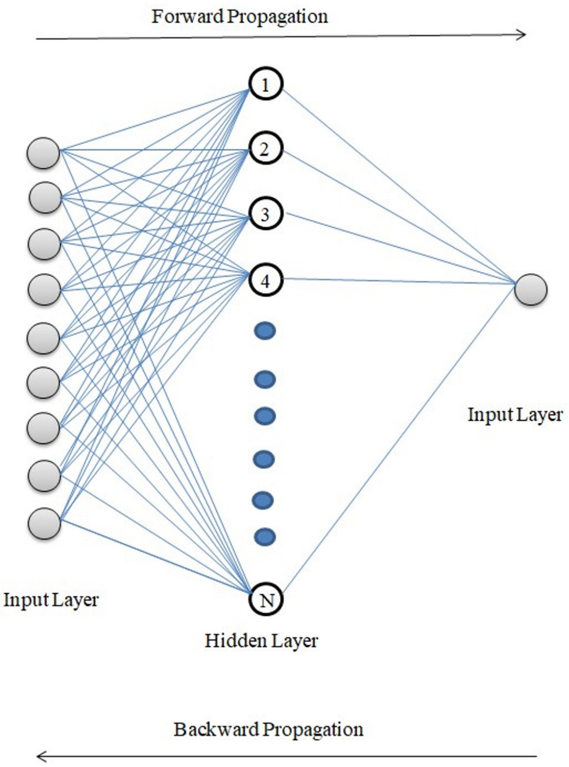

Figure 2. Multi-layer feed-forward neural network with BP algorithm

layers. The individual values are summed together and transmitted to the output layer by the output layer activation function. If the output is acceptable up to a particular level of error, it is permitted; otherwise, it is returned to the input layer for more updating of the weights and biases. It is worth noting that there is no link among nodes within the same layer. This cycle will continue until all of the limitations have been met [Chopra et al., 2019].

Performance criteria

The models were evaluated using Mean squared error (MSE) and Correlation Coefficient (R), as follows [Kulisz et al., 2021]:

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(T_j - O_j)^2 \quad (3)$$

$$R = \frac{\sum_{j=1}^{N}(T_j - \overline{T})(O_j - \overline{O})}{\left(\sqrt{\sum_{j=1}^{N}(T_j - \overline{T})^2 \sum_{j=1}^{N}(O_j - \overline{O})^2}\right)} \quad (4)$$

where: N is number of data, T is the target value, O is the output value of the network, $\overline{T}$ is the mean value of target data, and $\overline{O}$ is the mean value of network output.

Genetic Algorithm (GA)

John Holland invented Genetic Algorithm and presented his idea in his book in the year 1975 “Adaptation in Natural and Artificial Systems”. GA was suggested by Holland as a computational method dependent on the Survival of the Fittest principle [Sivanandam and Deepa, 2008]. Genetic algorithm is population-based stochastic algorithm. Selection, crossover, and mutation are the three main GA operators. Because the GA algorithm is random, one can wonder how trustworthy it is. The technique of keeping the best solutions for each generation and applying them to improve subsequent solutions is what makes this algorithm dependable and capable of estimating the global optimum for a particular problem. As a result, the entire population improves with each passing generation [Mirjalili, 2019]. The GA works with a group of chromosomes (also called individuals). Each chromosome indicates a workable solution to the problem researched. A collection of biologically based genetic operators, such as selection, crossover, and mutation, are used to generate the offspring chromosomes. The offspring are expected to inherit perfect genes from their parents, resulting in a higher average quality of solutions than previous generations. GA is iterative in their approach. A generation is the name given to each iteration. The fitness function evaluates and determines the fitness of each chromosome in each generation. A chromosome becomes fitter when its fitness function value goes up, indicating that it has a better chance of surviving in the next generation. This process of evolution is repeated until certain stopping requirements are met [Guo and Wong, 2013].

Implementation of Genetic Algorithm

The steps below, explain what the genetic algorithm will be doing [Abuiziah and Nidal, 2013]:

- GA begins with an initial population that is generated at random.
- Calculate the population’s fitness. Fitness function is implemented to each individual chromosome to produce a fitness score.
- The solution utilized to create the next solution is chosen depending on its fitness value. The solutions with a larger fitness value have a better probability of being chosen for reproduction, whereas those with a lesser fitness

value have a reduced possibility of being chosen for reproduction.
- Identify the crossover point, which can be random.
- Identify if mutation occurs.
- The present population is replaced by the new population.
- This evolution process is replicated until a predetermined termination criterion is met. For example, satisfaction with the enhancement of the best solutions might be used as criterion. Figure 3 shows how the GA performs [Tabassum and Mathew, 2014].

Normalization data

The term normalization refers to the process of converting data values to a range between 0 and 1. The actual data is first normalized using the formula [Chopra et al., 2019]:

$$x_n = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{5}$$

where: x_i is the i^{th} data to have been normalized, x_n is the normalized value, x_{min} is the minimum value of data, and x_{max} is the maximum value of data.

RESULTS AND DISCUSSION

Comprehensive pollution index

On the basis of to the CPI classification of all WTPs in this study for ten years from 2011 to 2020, the water of the Shatt Al-Arab River is classified as moderately polluted water and seriously polluted water, as demonstrated in the Figures from 4 to 8. The year 2018 was found to be the most polluted for all WTPs compared to other years, with the highest value of TDS reaching 22 954 mg/l at Al-Labanie (WTP No. 15). This was attributable to the salt tide in this year, in addition to the pollutants resulting from domestic, industrial and agricultural activities, which led to an increase in the salinity of the river.

Estimation of CPI by multiple linear regression

The multiple linear regression model enables to investigate the impact of numerous independent variables on the dependent variable. Dependent variable: CPI, independent variables: pH, Tur., TDS, EC, TH, Na^+, K^+, Ca^{+2}, Mg^{+2}, Alk., $SO_4^{-2,}$ and Cl^-. The SPSS program

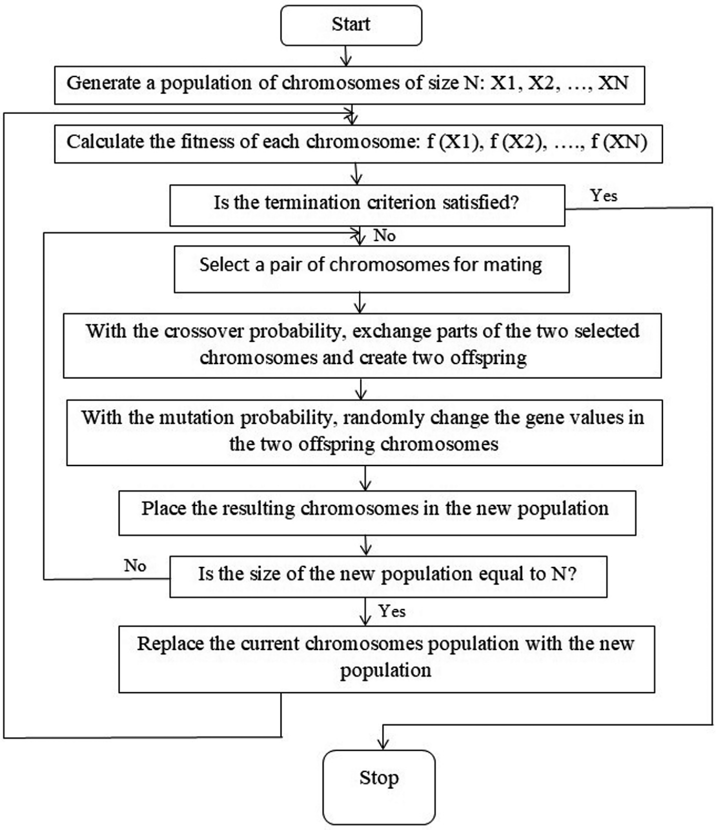
Start
Generate a population of chromosomes of size N: X1, X2, …, XN
Calculate the fitness of each chromosome: f (X1), f (X2), …., f (XN)
Is the termination criterion satisfied?
Yes
No
Select a pair of chromosomes for mating
With the crossover probability, exchange parts of the two selected chromosomes and create two offspring
With the mutation probability, randomly change the gene values in the two offspring chromosomes
Place the resulting chromosomes in the new population
No
Is the size of the new population equal to N?
Yes
Replace the current chromosomes population with the new population
Stop

Figure 3. Flow chart of GA

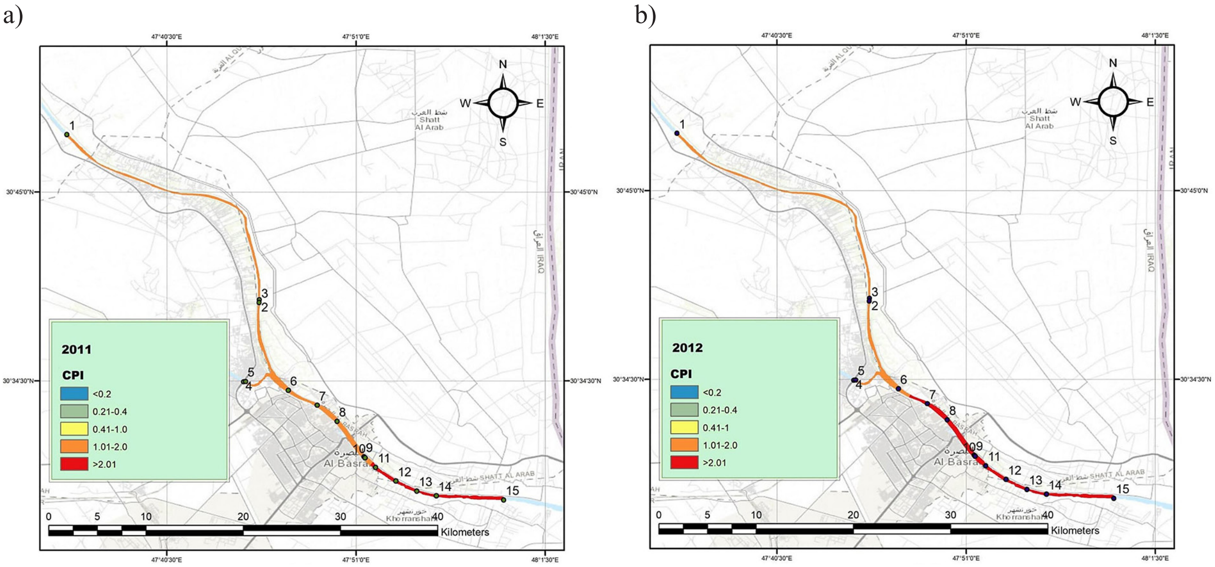

Figure 4. CPI values (average annual values) of WTPs along the SAR for (a) 2011, (b) 2012

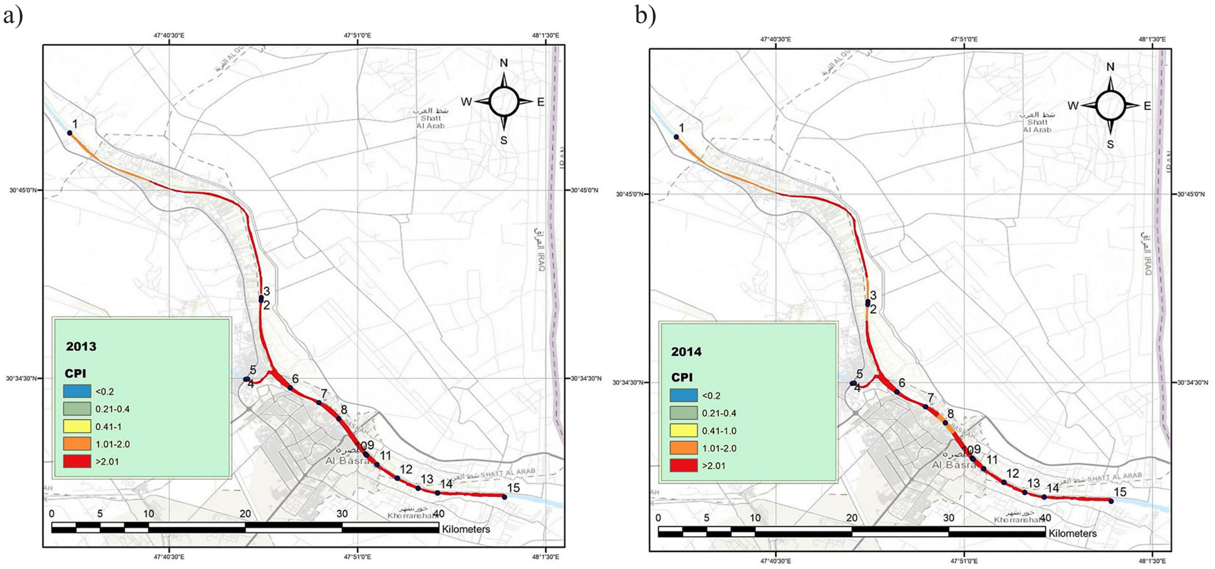

Figure 5. CPI values (average annual values) of WTPs along the SAR for (a) 2013, (b) 2014

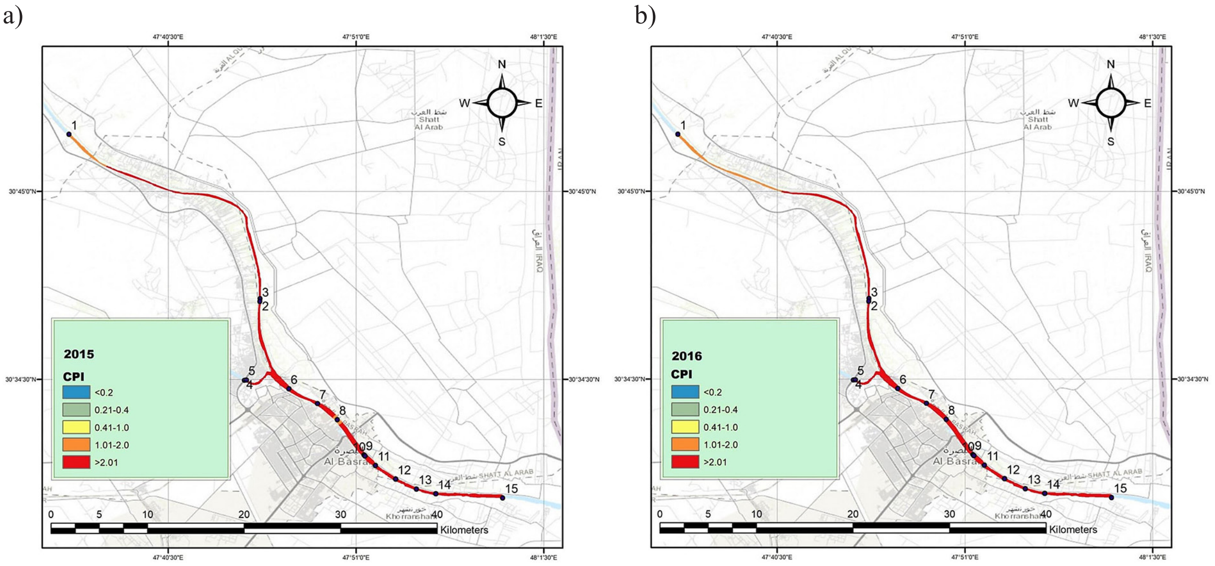

Figure 6. CPI values (average annual values) of WTPs along the SAR for (a) 2015, (b) 2016

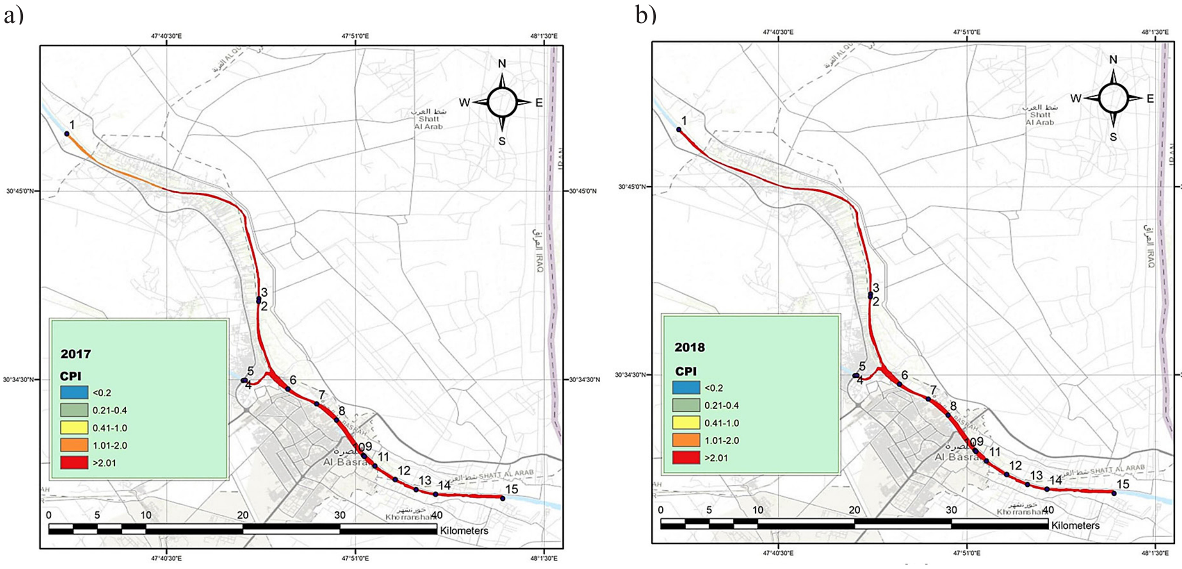

Figure 7. CPI values (average annual values) of WTPs along the SAR for (a) 2017, (b) 2018

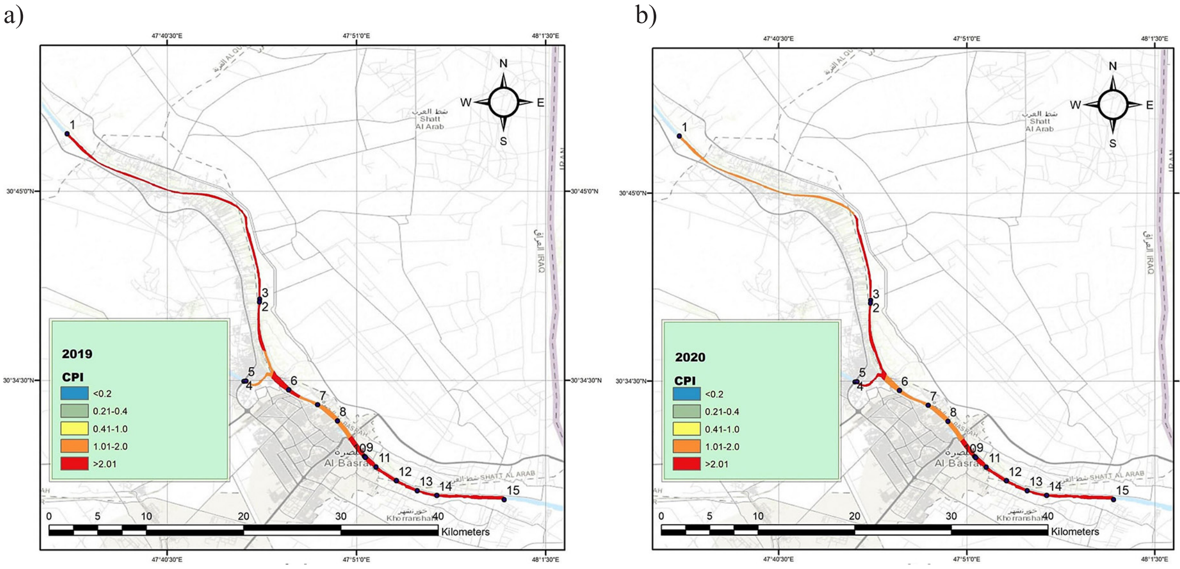

Figure 8. CPI values (average annual values) of WTPs along the SAR for (a) 2019, (b) 2020

was used to analyze the data, in this model, the multiple correlation coefficients R is 0.996, and the coefficient of determination R^2 is 0.991. The success percentage of this model is 99.1% with a 0.9% error rate.

The pH, TH, and Alk., in this model are not statistically significant, since their p-values are more than the 5% level of significance, p-value = 0.834 for pH, p-value = 0.916 for TH, and p-value = 0.848 for Alk., as illustrated in Table 5.

As a result, the following variables will be used to estimate the comprehensive pollution index: Tur., TDS, EC, Na^+, K^+, Ca^{+2}, Mg^{+2}, SO_4^{-2}, and Cl^-. Dependent variable: CPI, independent variables: Tur., TDS, EC, Na^+, K^+, Ca^{+2}, Mg^{+2}, SO_4^{-2}, and Cl^-.

The SPSS program was used to analyze the data, in this model, the multiple correlation coefficients R is 1, and the coefficient of determination R^2 is 1; this means that this model is able to predict CPI values extremely accurately. Because the p-value for all predictor variables is less than 0.001, they are statistically significant, as illustrate in Table 6. The correlation between the measured and regression variables was positive, as presented in Table 7. The equation for MLR model is as follows:

$$CPI = -0.016 + 0.106\ Tur. + 0.206\ TDS + 0.202\ EC + 0.229\ Na^+ + 0.065\ K^+ + 0.046\ Ca^{+2} + 0.013\ Mg^{+2} + 0.146\ SO_4^{-2} + 0.191\ Cl^-. \quad (6)$$

Table 5. Coefficient values for 12 independent variables

Model	Unstandardized coefficients		Standardized coefficients	t	Sig.
	B	Std. Error	Beta		
Constant	-0.015	0.001		-10.877	0
K	0.061	0.004	0.034	14.684	0
Na	0.243	0.012	0.238	20.399	0
TDS	0.190	0.051	0.184	3.702	0
SO_4	0.138	0.008	0.132	18.396	0
Cl	0.187	0.026	0.183	7.253	0
Mg	0.054	0.025	0.050	2.156	0.031
Ca	0.079	0.026	0.074	3.062	0.002
Alk.	0.001	0.003	0	0.192	0.848
TH	0	-	0	-0.106	0.916
EC	0.153	0.050	0.151	3.074	0.002
PH	0	0.002	0	-0.209	0.834
Tur.	0.105	0.002	0.093	46.676	0

Table 6. Coefficient values for 9 independent variables

Model	Unstandardized coefficients		Standardized coefficients	t	Sig.
	B	Std. Error	Beta		
Constant	-0.016	0		-217.521	0
K	0.065	0	0.037	191.673	0
Na	0.229	0.001	0.226	237.540	0
TDS	0.206	0.004	0.200	49.317	0
SO4	0.146	0.001	0.140	237.146	0
Cl	0.191	0.002	0.189	91.562	0
Mg	0.013	0.002	0.012	6.487	0
Ca	0.046	0.002	0.044	22.054	0
EC	0.202	0.004	0.201	50.202	0
Tur.	0.106	0	0.095	586.067	0

Table 7. Statistics and correlations

Variables	N	Mean	Std. deviation	Std. error mean	Correlation	Sig.
Measured	2430	0.133631	0.120147	0.00244	1	0
Regression	2430	0.133512	0.119920	0.00243		

Estimation of CPI by Artificial Neural Network

The back propagation algorithm has been used to train the created ANN models. Multiple linear regression analysis was used to determine the number of input variables. Tur., TDS, EC, Na^+, K^+, Ca^{+2}, Mg^{+2}, SO_4^{-2}, and Cl^- were utilized as input variables to predict the CPI. 70% of the data was used for training set, 20% for testing, and 10% for validation set, because this proportion produced the best performance in terms of least MSE and highest R values. In order to find the optimal number of nodes in the hidden layer, many ANN models were created and evaluated. The effectiveness of the ANN models was assessed utilizing the coefficient of correlation (R) and the mean squared error (MSE).

The maximum regression coefficient and minimum mean squared error for the training set, validation set, and testing set, for each training functions used for one and two hidden layers are presented in Tables 8 and 9, respectively.

From Tables 8 and 9, the optimum prediction model was found in the 9-16-1 network structure. The Levenberg Marquardt algorithm trained this network, because when compared to other training functions, it produced the best performance in terms of least MSE and highest R values. In the hidden layer, logsig was chosen as activation function and purelin activation

Table 8. Best results with one hidden layer for the training functions of the ANN model

Algorithms	No. of neurons	MSE (training set)	MSE (validation set)	MSE (testing set)	R (testing set)	Epochs
trainlm	16	9.755×10^{-7}	1.945×10^{-7}	8.388×10^{-8}	1	226
trainbfg	11	6.734×10^{-6}	1.054×10^{-5}	8.078×10^{-6}	0.9996	88
traincgb	12	1.053×10^{-6}	1.519×10^{-5}	8.843×10^{-6}	0.9995	95
traincgf	10	2.817×10^{-5}	1.249×10^{-5}	1.561×10^{-5}	0.9993	180
traincgp	14	2.708×10^{-5}	1.882×10^{-5}	5.623×10^{-5}	0.9968	75
traingdm	12	0.0039	0.0043	0.0029	0.9414	5000
traingda	11	8.964×10^{-4}	5.375×10^{-4}	8.761×10^{-4}	0.9473	160
traingdx	15	0.0020	0.0014	0.0040	0.9453	120
trainoss	9	2.019×10^{-5}	1.350×10^{-5}	1.568×10^{-5}	0.9993	98
trainrp	10	3.020×10^{-5}	2.838×10^{-5}	2.008×10^{-5}	0.9990	260
trainscg	15	4.153×10^{-5}	3.361×10^{-5}	4.898×10^{-5}	0.9972	210

Table 9. Best results with two hidden layer for the training functions of the ANN model

Algorithms	No. of neurons	MSE (training set)	MSE (validation set)	MSE (testing set)	R (testing set)	Epochs
trainlm	[10 17]	2.524×10^{-8}	4.089×10^{-7}	4.147×10^{-9}	1	298
trainbfg	[9 10]	8.929×10^{-6}	1.153×10^{-5}	6.142×10^{-6}	0.9997	145
traincgb	[13 12]	7.815×10^{-5}	6.050×10^{-5}	5.590×10^{-5}	0.9961	90
traincgf	[9 13]	2.754×10^{-5}	3.572×10^{-5}	2.648×10^{-5}	0.9979	220
traincgp	[11 13]	1.053×10^{-4}	6.814×10^{-4}	1.328×10^{-4}	0.9944	77
traingdm	[14 10]	0.0087	0.0088	0.0080	0.5935	5000
traingda	[16 9]	0.0013	5.926×10^{-4}	0.0040	0.8583	398
traingdx	[17 11]	4.356×10^{-4}	$1.998 * 10^{-4}$	5.779×10^{-4}	0.9742	187
trainoss	[12 12]	4.175×10^{-5}	3.383×10^{-5}	9.625×10^{-5}	0.9969	150
trainrp	[9 20]	2.486×10^{-4}	2.376×10^{-4}	2.689×10^{-4}	0.9879	170
trainscg	[17 14]	5.499×10^{-5}	1.803×10^{-4}	3.988×10^{-5}	0.9984	86

function was chosen in the output layer. This structure produced the lowest MSE value of 9.755×10^{-7} for the training set, 1.945×10^{-7} for the validation set, and 8.388×10^{-8} for testing set as shown in the Figure 9, and maximum R value of 0.99996 for training set, 0.99998 for validation set, and 1 for testing set, as shown in the Figure 10. Table 10 represents the properties of the selected ANN model in this study. Figure 11 indicates the optimal ANN structure performed in this study.

Table 10. The properties of the selected ANN model in this study

Character	Description
Artificial neural network type	feedforward multi-layer NN
Performance function	MSE
Training function	trainlm
Activation function in hidden layer	logsig
Activation function in output layer	purelin
ANN architecture	9-16-1

Genetic algorithm optimization solution

GA is a method to resolving the optimization problems that are both constrained and unconstrained. The goal of the optimization procedure in this study was to identify the optimum values for the independent variables that lead to the minimum WPI value. The comprehensive pollution index estimation model described in Eq. (6) is chosen as the fitness function and written as follows:

$$\text{Minimize CPI (Tur., T.D.S, EC, Na}^{+}\text{, K}^{+}\text{, Ca}^{+2}\text{, Mg}^{+2}\text{, SO}_4^{-2}\text{, Cl}^{-}\text{)} = \min(-0.016 + 0.106\ \text{Tur.} + 0.206\ \text{TDS} + 0.202\ \text{EC} + 0.229\ \text{Na}^{+} + 0.065\ \text{K}^{+} + 0.046\ \text{Ca}^{+2} + 0.013\ \text{Mg}^{+2} + 0.146\ \text{SO}_4^{-2} + 0.191\ \text{Cl}^{-}). \quad (7)$$

The reduction of the objective function value is exposed to the limits of predictor variable values. The range of values of measured predictor variables are chosen to illustrate the constraints of the optimization solution, as presented in Table 11.

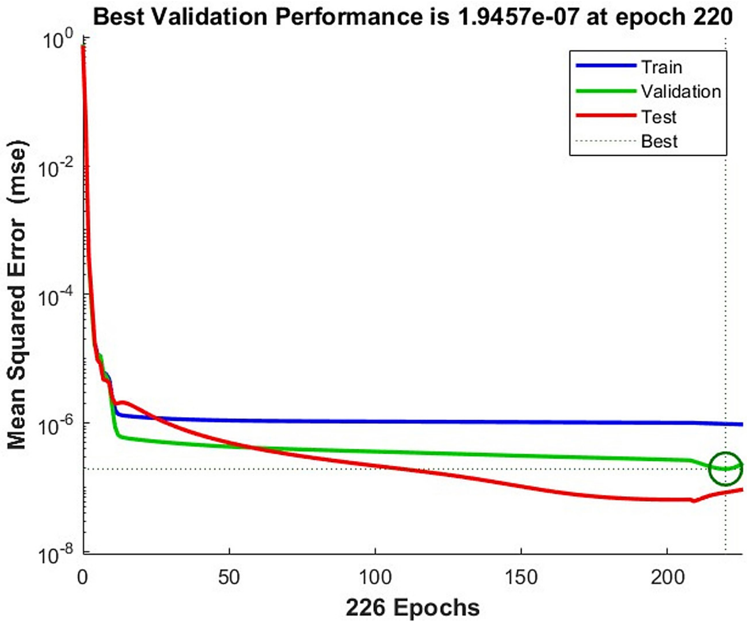

Figure 9. MSE values for training, validation, and testing sets with one hidden layer for the ANN model selected

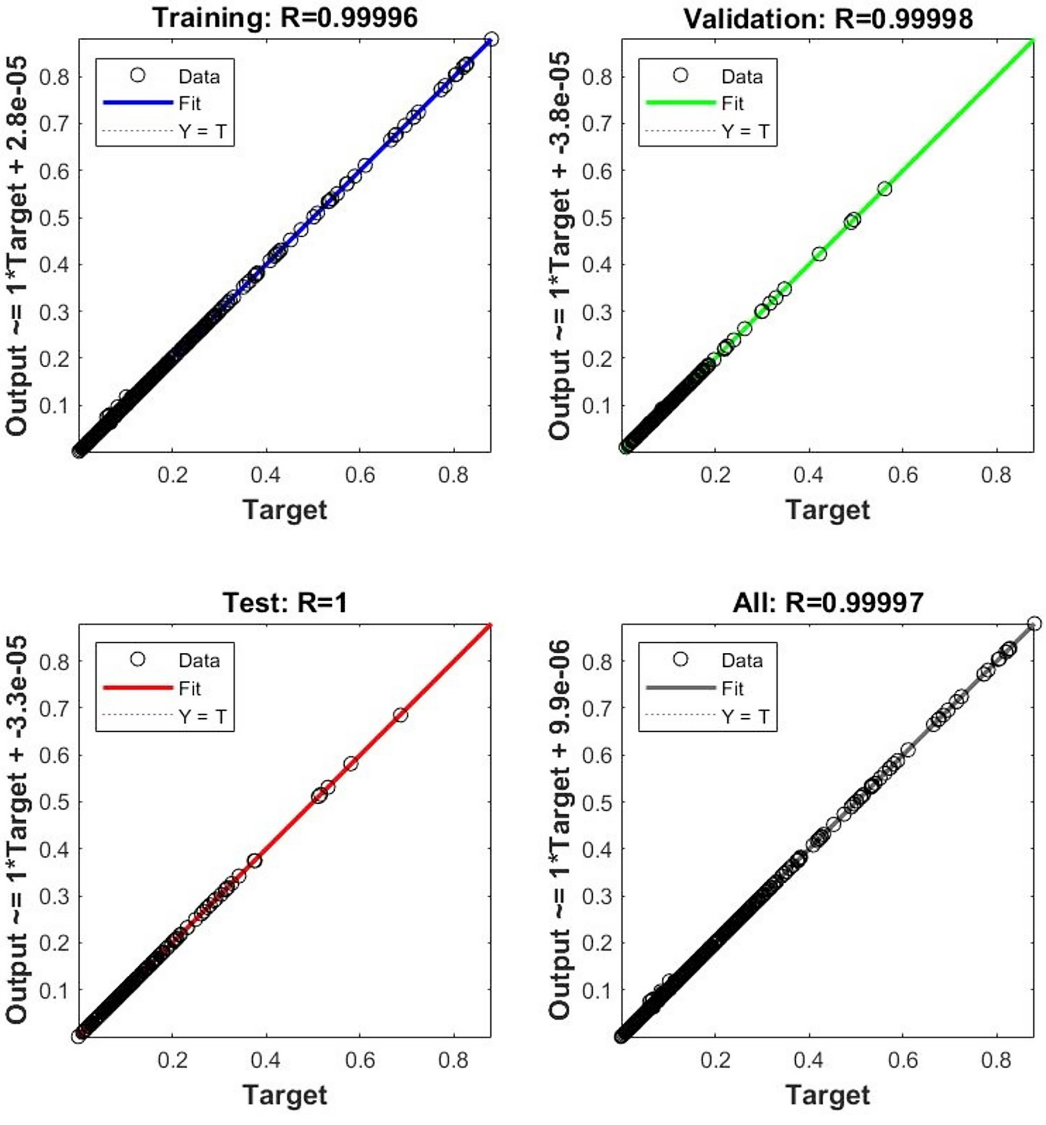

Figure 10. R values for training, validation, and testing sets with one hidden layer for the ANN model selected

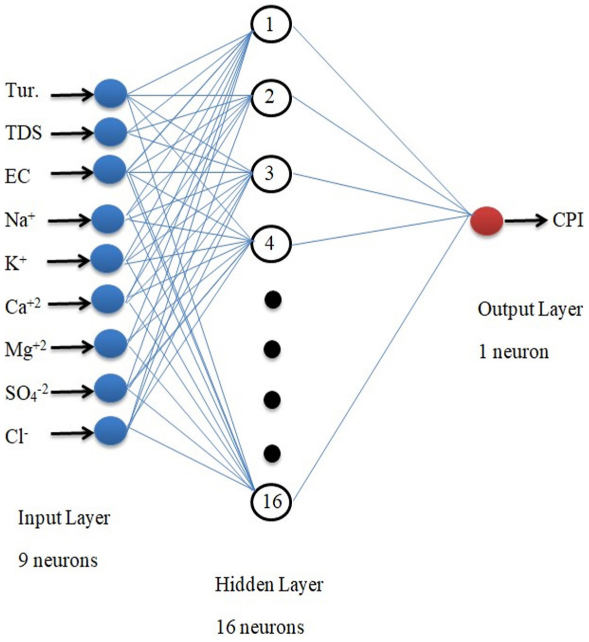

Figure 11. The ideal network architecture for prediction CPI value

Table 11. Limitations of predictor variables for GA optimization solution

Independent variable	Limitations (normalized values)
Tur.	0 ≤ Tur. ≤1
TDS	0 ≤ TDS ≤1
EC	0 ≤ EC ≤ 1
Na^+	0 ≤ Na^+ ≤ 1
K^+	0 ≤ K^+ ≤ 1
Ca^{+2}	0 ≤ Ca^{+2} ≤ 1
Mg^{+2}	0 ≤ Mg^{+2} ≤ 1
SO_4^{-2}	0 ≤ SO_4^{-2} ≤ 1
Cl^-	0 ≤ Cl^- ≤ 1

Table 12. The finest combination of GA parameters that resulted in the optimal solution

Parameter	Setting
Population size	100
Scaling function	Shift linear
Selection function	Roulette
Crossover function	Constraint
Crossover rate	0.8
Mutation function	Constraint
Number of generation	1000

The Matlab optimization toolbox has been employed to identify the CPI lowest value at the ideal points by applying the fitness function of Eq. (7), the limits of predictor variables in Table 11. The genetic algorithm generated a global minimum of -0.0159 for the CPI as shown in Figure 12. The ideal values of predictor variables (normalized values) are zero for all predictor variables. At 51 iteration of the genetic algorithm, the best solution was found. Table 12 shows the suitable combination of parameters used for the genetic algorithm that leads in the lowest fitness function value.

RESULTS OF THIS STUDY

In comparison to the results of actual data, MLR, and ANN models, the GA-MLR method is

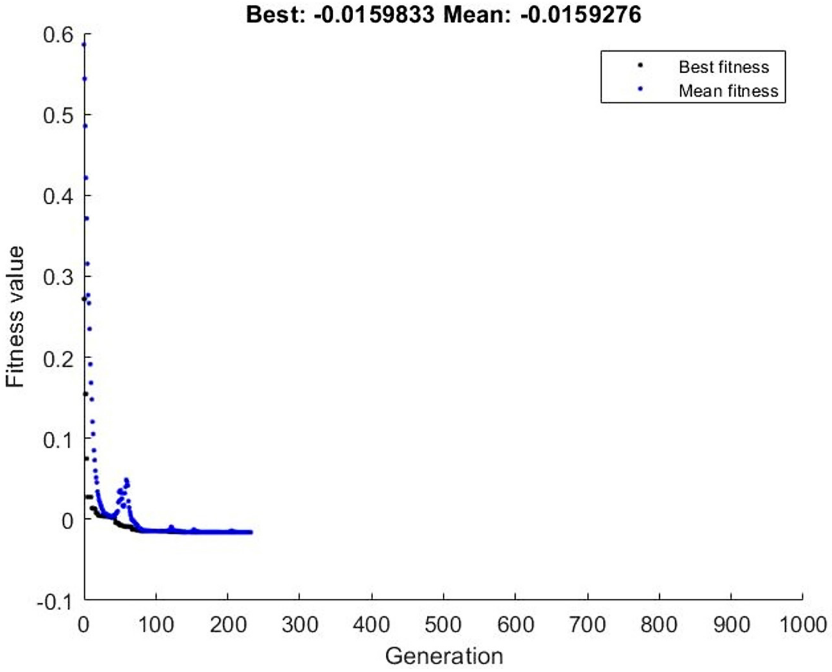

Figure 12. Best fitness value and mean fitness value

Table 13. Description of results of this research

Approach	Measured	MLR	ANN	GA
Minimum WPI (normalized value)	0	- 0.0003	0.0003	- 0.01598
Minimum WPI (real value)	0.6416	0.6361	0.6460	0.3777
Tur. (NTU)	6.9	6.9	6.9	0.6
TDS (mg/l)	584	584	584	200
EC (μs/cm)	942	942	942	871
Na (mg/l)	84	84	84	62
K (mg/l)	3.5	3.5	3.5	2.5
Ca (mg/l)	65	65	65	59
SO_4 (mg/l)	156	156	156	134
Cl (mg/l)	150	150	150	104

effective in producing the lower CPI value at the ideal values of predictor variables. As presented in Table 13, the minimum CPI value for the GA-MLR approach is 0.3777.

CONCLUSIONS

The water of the Shatt Al-Arab River is categorized as moderately polluted water and seriously polluted water by the CPI classification in this study. The performance of the MLR and ANN models for estimating CPI was evaluated and it was found the MLR and ANN models were very suitable for predicting the CPI based on the results of this study. The optimum prediction model was found in the 9-16-1 network structure. This structure produced the lowest MSE value of 9.755×10^{-7} for the training set, 1.945×10^{-7} for the validation set, and 8.388×10^{-8} for testing set, and maximum R value of 0.99996 for training set, 0.99998 for validation set, and 1 for testing set. According to the results of this study, the GA-MLR technique is capable of estimating the ideal parameters that result in the minimum CPI value. The minimum predicted CPI value recommended by the GA-MLR approach was 0.3777.

REFERENCES

1. Abbas A.H.A., Dawood A.S., Al-Hasan Z.M. 2017. Evaluation of groundwater quality for drinking purpose in basrah governorate by using application of water quality index. Kufa Journal of Engineering, 8, 65-78.
2. Abdullah E.J. 2013. Quality assessment for Shatt Al-Arab River using heavy metal pollution index and metal index. Journal of Environment and Earth Science, 3, 114-120.
3. Abuiziah I., Nidal S. 2013. A review of genetic algorithm optimization: operations and applications to water pipeline systems. International Journal of Physical, Natural Science and Engineering. 7, 341-347.
4. Abyaneh H.Z. 2014. Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. Journal of Environmental Health Science and Engineering, 12, 1-8.
5. Al Tobi M.A.S., Bevan G., Wallace P., Harrison D., Ramachandran K. 2016. A review on applications of genetic algorithm for artificial neural network. International Journal of Advanced Computational Engineering and Networking, 4, 50-54.
6. Al-Adhab H., Salman A., Sagban A. 2019. Using multivariate statistical methods to Evaluate water quality in some of Basrah province locations. Proceedings of the ICCEET.
7. Al-Asadi S.A., Al-Qurnawi W.S., Al Hawash A.B., Ghalib H.B., Alkhlifa, N. H. A. 2020. Water quality and impacting factors on heavy metals levels in Shatt Al-Arab River, Basra, Iraq. Applied Water Science, 10, 1-15.
8. Al-Asadi S.A., Alhello A.A. 2019. General assessment of Shatt Al-Arab River, Iraq. International Journal of Water, 13, 360-375.
9. Allafta H., Opp C. 2020. Spatio-temporal variability and pollution sources identification of the surface sediments of Shatt Al-Arab River, Southern Iraq. Scientific reports, 10, 1-16.
10. Almuktar S., Hamdan A.N.A., Scholz M. 2020. Assessment of the effluents of Basra City main water treatment plants for drinking and irrigation purposes. Water, 12, 1-26.
11. Azad A., Farzin S., Mousavi S.F., Firoozbakht A., Ghorbani S., Heravi F. 2016. The use of optimized artificial neural network model by the Genetic Algorithm in estimating water salinity parameters (Case

study: Gorganrood River). International Congress on Civil Engineering, Architecture and Urban Development.

12. Banejad H., Olyaie E., Application of an artificial neural network model to rivers water quality indexes prediction–a case study. Journal of American science, 7, 60-65.
13. Chen Y., Fang X., Yang L., Liu Y., Gong C., Di Y. 2019. Artificial Neural Networks in the Prediction and Assessment for Water Quality: A Review. In Journal of Physics: Conference Series, IOP Publishing, 1237, 1-8.
14. Chopra S., Yadav D., Chopra A.N. 2019. Artificial neural networks based indian stock market price prediction: before and after demonetization. International Journal of Swarm Intelligence and Evolutionary Computation, 8, 1-7.
15. Dawood A.S. 2017. Using multivariate statistical methods for the assessment of the surface water quality for a river: A case study. International Journal of Civil Engineering and Technology (IJCIET), 8, 588-597.
16. Dawood A.S., Faroon M.A., Yousif, Y.T. 2020. The use of multivariate statistical techniques in the assessment of river water quality. Anbar Journal of Engineering Sciences, 8, 93-203.
17. Dawood A.S., Hamdan A.N., Khudier A.S. 2018. Assessment of water quality index with analysis of physiochemical parameters. Case study: The Shatt Al-Arab River, Iraq. International Energy and Environment Foundation, 5, 93-106.
18. Dawood A.S., Hussain H.K., Hassan A. 2016. Modeling of river water quality parameters using artificial neural network-a case study. Proceedings of 40th ISERD International Conference, Cairo, Egypt.
19. Ewaid S.H., Abed S.A. 2017. Water quality index for Al-Gharraf river, southern Iraq. The Egyptian Journal of Aquatic Research, 43, 117-122.
20. Ezzat S.M., Elkorashey R.M. 2020. Wastewater as a Non-conventional Resource: Impact of Trace Metals and Bacteria on Soil, Plants, and Human Health. Human and Ecological Risk Assessment: An International Journal, 26, 1-21.
21. Fisz J.J. 2006. Combined genetic algorithm and multiple linear regression (GA-MLR) optimizer: Application to multi-exponential fluorescence decay surface. The Journal of Physical Chemistry, 110, 12977-12985.
22. Gallo C. 2015. Artificial neural networks tutorial. Encyclopedia of Information Science and Technology, Third Edition, IGI Global, 179-189.
23. Gazzaz N.M., Yusoff M.K., Aris A.Z., Juahir H., Ramli M.F. 2012. Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. Marine pollution bulletin, 64, 2409-2420.
24. Ghalib H. B. 2017. Groundwater chemistry evaluation for drinking and irrigation utilities in east Wasit province, Central Iraq. Applied Water Science, 7.
25. Ghose D. K., Samantaray S. 2018. Modelling sediment concentration using back propagation neural network and regression coupled with genetic algorithm. Procedia Computer Science, 125, 85-92.
26. Goudarzi N., Goodarzi M., Chen, T. 2012. QSAR prediction of HIV inhibition activity of styrylquinoline derivatives by genetic algorithm coupled with multiple linear regressions. Medicinal Chemistry Research, 21, 437-443.
27. Guidea A., Sarbu C. 2019. Modeling and prediction of amino acids lipophylicity using multiple linear regression coupled with genetic algorithm. Studia Universitatis Babes-Bolyai, Chemia, 64, 243-254.
28. Guo Z. X., Wong W.K. 2013. Optimizing Decision Making in the Apparel Supply Chain Using Artificial Intelligence (AI): from Production to Retail. Woodhead Publishing, 1st Edetion, ISBN: 978-0-85709-779-8.
29. Hamdan A.N.A., Dawood A.S. 2016. Neural Network Modelling of Tds Concentrations in Shatt Al-Arab River Water. Engineering and Technology Journal, 34, 334-345.
30. Hamdan A., Dawood A., Naeem D. 2018. Assessment study of water quality index (WQI) for Shatt Al-arab River and its branches, Iraq. In MATEC Web of Conferences.
31. Khudhur Z.A., Arab S.A., Dawood A.S. 2020. Assessment of dissolved oxygen in Shatt Al-Arab River by other quality parameters of water using Artificial Neural Networks. Muthanna Journal of Engineering and Technology (MJET), 8, 42-50.
32. Kulisz M., Kujawska J., Przysucha B., Cel W. 2021. Forecasting water quality index in groundwater using artificial neural network. Energies, 14.
33. Mahmood W., Ismail A.H., Shareef M.A. 2019. Assessment of potable water quality in Balad city, Iraq. In IOP conference series: materials science and engineering, IOP Publishing.
34. Manroo S., Ganiny S. A. 2020. Coupled Multi-linear Regression and Genetic Algorithm Based Modeling and Optimization of Surface Roughness in Machining of Brass. International Conference on Advances in Systems, Control and Computing.
35. Matta G., Naik, P., Kumar A., Gjyli L., Tiwari A.K., Machell J. 2018. Comparative study on seasonal variation in hydro-chemical parameters of Ganga River water using comprehensive pollution index (CPI) at Rishikesh (Uttarakhand) India. Desalination and Water Treatment, 118, 87-95.
36. Mijwel M.M. 2016. Genetic algorithm optimization by natural selection. Department of computer science, college of science, Baghdad University, Baghdad, Iraq.

37. Mijwel M.M., Alsaadi A. 2019. Overview of Neural Networks. Department of Computer Science, College of Science, Baghdad University, Baghdad, Iraq.
38. Mirjalili S. 2019. Evolutionary algorithms and neural networks. Springer, Cham, 1st Edition.
39. Mishra S., Kumar A., Shukla P. 2016. Study of water quality in Hindon River using pollution index and environmetrics, India. Desalination and Water Treatment, 57, 1-10.
40. Najah A., El-Shafie A., Karim O.A., and El-Shafie A.H. 2013. Application of artificial neural networks for water quality prediction. Neural Computing and Applications, 22, 187-201.
41. Nwobi-Okoye C.C., Ochieze B.Q. 2018. Age hardening process modeling and optimization of aluminum alloy A356/Cow horn particulate composite for brake drum application using RSM, ANN and simulated annealing. Defence Technology, 14, 336-345.
42. Salihu M., Shawai S.A.A., Shamsuddin I.M. 2017. Effect and control of water pollution a panacea to national development. International Journal of Environmental Chemistry, 1, 23-27.
43. Singh, J., Yadav P., Pal A.K., and Mishra V. 2020. Water Pollutants: origin and status. In Sensors in Water Pollutants Monitoring: Role of Material, Springer, Singapore, 5-20
44. 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, 888-895.
45. Sivanandam S.N., Deepa, S.N. 2008. Introduction to genetic algorithms. Springer, Berlin, Heidelberg, 1st Edition.
46. Son C.T., Giang N.T.H., Thao T.P., Nui N.H., Lam, N.T., Cong, V.H. 2020. Assessment of Cau River water quality assessment using a combination of water quality and pollution indices. Journal of Water Supply: Research and Technology-Aqua, 69, 160-172.
47. Tabassum M., Mathew K. 2014. A genetic algorithm analysis towards optimization solutions. International Journal of Digital Information and Wireless Communications, 4, 124-142.
48. Yan C.A., Zhang W., Zhang Z., Liu Y., Deng C., Nie N. 2015. Assessment of water quality and identification of polluted risky regions based on field observations and GIS in the honghe river watershed, China. PLoS ONE, 10, 1-13.
49. Zain A.M., Haron H., Sharif S. 2010. Application of GA to optimize cutting conditions for minimizing surface roughness in end milling machining process. Expert Systems with Applications, 37, 4650-4659.