Soft computing technique-based prediction of water quality index

Balraj Singh*, Parveen Sihagb, Vijay P. Singhc, Alireza Sepahvandd and Karan Singhe

a Civil Engineering Department, Panipat Institute of Engineering and Technology, Samalkha, India
b Civil Engineering Department, Shoolini University, Solan, India
c Department of Biological and Agricultural Engineering & Zachry Department of Civil Engineering, Texas A & M University, College Station, Texas, USA
d Department of Range and Watershed Management Engineering, Lorestan University, Khorramabad, Iran
e Civil Engineering Department, National Institute of Technology, Hamirpur, India

*Corresponding author. E-mail: balraj.civil@piet.co.in

ABSTRACT

Water quality plays a crucial role in management of water resources. Water quality indexes (WQIs) are frequently used methods to assess water quality for drinking purposes. A WQI can be predicted using chemical analysis which might not, however, be viable for a longer period in all country-scale rivers. Thus, in this investigation, two neural-based soft computing techniques – an artificial neural network (ANN) and a generalized regression neural network (GRNN) – and one hybrid soft computing techniques – an adaptive neuro-fuzzy interference system (ANFIS) with four membership functions were used to predict WQIs in Khorramabad, Biranshahr and Alashtar sub-watersheds in Iran. Ten distinct physiochemical parameters were used as input variables and WQI as output. Simultaneously, a correlation plot and pairs were used to ascertain the relation of input and output variables. The soft computing techniques were compared using six fitness criteria: Nash-Sutcliffe efficiency (NSE), mean absolute error (MAE), Legates-McCabe Index (LMI), root mean square error (RMSE), mean absolute percentage error (MAPE), and coefficient of correlation (CC). Results indicated that ANN better predicted WQI than did GRNN and ANFIS. Among the different membership functions of ANFIS, ANFIS_trimf was far better than were the others. Thus, it was concluded that ANN was a viable tool for the prediction of a WQI.

Key words: adaptive neuro-fuzzy interference system, artificial neural network, generalized regression neural network, soft computing, water quality index

HIGHLIGHTS

• Water quality index values were predicted by using three soft computing techniques.
• Ten distinct physiochemical parameters were utilized in modelling.
• The soft computing techniques were compared using six fitness criteria: NSE, MAE, LMI, RMSE, MAPE, CC.

INTRODUCTION

Industries, agriculture, and people pollute water resources through a variety of activities (Katyal 2011). In some watersheds, pollution has exceeded the permissible limit. There is global concern as water quality has degraded almost everywhere (Adriaenssens et al. 2004; Azad et al. 2018). Quality water is fundamental for sustainable living so the abatement of pollution and protection of water resources are necessary, which requires an assessment of water quality (Witek & Jarosiewicz 2009; Reza & Singh 2010; Tiri et al. 2018). Water quality can be determined by chemical, physical and biological analyses (Abbasi & Abbasi 2012; Ewaid & Abed 2017; Medeiro et al. 2017; Tiri et al. 2018). A Water Quality Index (WQI) is one of the commonly used methods for the assessment of water quality (Medeiros et al. 2017; Tiri et al. 2018). WQI group parameters can be utilized for grading water quality and hence used in classification of the health of water systems, such as rivers (Hasan et al. 2015; Ewaid & Abed 2017).

WQIs were proposed by Brown et al. (1970) and Horton (1965) and various methods have since been developed for calculating them (Debels et al. 2005; Tsegaye et al. 2006; Saeedi et al. 2010). Using a WQI, Kannel et al. (2007) analyzed seasonal and spatial changes of the Bagmati River. Debels et al. (2005) calculated a WQI using 9 physio-chemical parameters in the Chill’an River. Yidana & Yidana (2010) combined GIS with a multivariate statistical method to calculate a WQI.
Soft computing techniques have been devised for addressing non-stationary and non-linearity of quality of water. These techniques are attractive, because they directly and quickly model water quality (Gaya et al. 2020; Karim & Kamsani 2020; Yasin & Karim 2020; Hmoud Al-Adhaileh & Waselallah Alsaade 2021), and have a great ability to reduce errors and time of computation (Bhagat et al. 2019). M5P model tree, adaptive neuro-fuzzy interference system (ANFIS), support vector regression (SVM), random forest (RF), Gaussian process (GP), and artificial neural network (ANN), are among the most frequently used soft computing techniques (Barzegar et al. 2016). Chen & Zheng (2008) used soft computing techniques in the prediction of water quality and observed that ANN is the best model which gives the most appropriate result. Singh et al. (2009) implemented an ANN model in the modelling of water quality to compute biological oxygen and oxygen demand. Practical swarm optimization with ANN was applied to predict water quality of sewage effluent by Zheng et al. (2010). Gao et al. (2015) applied a special type of neuronal network: a back propagation neural network combined with practical swarm optimization in prediction of water quality. Nourani et al. (2013) employed ANN for computing a WQI and found that it outperformed other conventional methods. Emamgholizadeh et al. (2014) used ANN and ANFIS to estimate a WQI in the Karoom watershed and found that ANN predicted better than ANFIS. Different combinations of soft computing techniques have also been employed for the estimation of WQIs (Yaseen et al. 2018).

The reliability of soft computing techniques for WQIs has been amply demonstrated (Bui et al. 2020; Gaya et al. 2020; Najafzadeh & Lottifi-Dashbalagh 2020; Tung & Yaseen 2020; Riahi-Madvar et al. 2021). Since analysis of the water quality of all rivers might not be possible at frequent intervals on a countrywide scale for a substantial period of time (De LR Wagener et al. 2019), modeling of water quality with ease and fewer parameters provides motivation for the use of soft computing techniques. Although, a lot of literature is available in which water quality was predicted by the use of soft computing techniques, no one has predicted the water quality of three sub watersheds with a combination of ANFIS, ANN and GRNN (generalized regression neural network), which indicates its significance and novelty. The analysis of water quality in a laboratory is a very costly and time consuming process which requires collection of samples, transportation and testing. In this regard, the study presents a real-time system to evaluate an alternative approach based on soft computing techniques for predicting water quality. The objectives of this study are:

i to develop a neuro-fuzzy based model, an ANFIS, for the prediction of a WQI for Khorramabad, Biranshahr and Alashtar sub-watersheds in Iran;
ii to validate the output of ANN, GRNN and ANFIS; and
iii to compare the performances of the soft computing techniques using model fitness criteria.

The paper is organized as follows. In the second section, a description of the soft computing techniques and study area is given. Details of the methodology, data description, and model fitness criteria are also given. Results and discussion summarizing the performances of soft computing techniques are presented in section 3, followed by conclusions (section 4) and references.

**MATERIALS AND METHODS**

**Soft computing techniques**

**Artificial neural network (ANN)**

ANN, a concept taken from the human mind, is a widely used prediction technique (Sepahvand et al. 2019; Sihag et al. 2020; Singh 2020). ANN has a brain-like architecture and neuron system. It contains a single input layer, single target layer and one or multiple hidden layers. Every layer has a certain quantity of nodes, and the weighted relation between these layers depicts the node relationship. The input layer, which has the same number of nodes as the number of input parameters, delivers data to the network but does not assist in processing. The last processing unit is the target layer. Whenever an input layer receives input information that moves through the linkages among the nodes, the values are multiplied by the associated weights and added together to get the final target \( Z_d \) to the unit.

\[
Z_d = \sum (A_{cd} \times B_d)
\]

where \( A_{cd} = \) weight of interconnection from unit c to d, \( B_c = \) input value at the input layer, and \( Z_d = \) target obtained by the activation function to produce a target for unit d. Haykin (1999) has given a complete discussion
of ANN. The main advantage of this method is that it learns automatically and produces an output which is not limited to the input provided. Also, its working is not affected by loss of data as it stores the input in its own networks instead of a database.

**Generalized regression neural network (GRNN)**

Specht (1991) was the first to introduce GRNN, which uses a normalized radial basis function (RBF) network with a single hidden component based at each training example. The kernel function, also known as the RBF, is a probability density function that includes neural networks, Gaussian processes, and support vector machines. The target values are the hidden-to-output weights, so the output is a weighted average of the target values of training bags near the specified input bags. The widths of the RBF components are the only weights which need to be investigated. There are only four levels in a GRNN structure. The input values are in the first level, the pattern elements are in the second level, the targets from this level are crisscrossed to the summation elements in the third level, and the output elements are in the final level. The first level is completely linked to the second, pattern level, where each element shows a training pattern and its output is a measure of the distance of input from the stored pattern. The optimal value of the user-defined parameter known as spread (s) is determined experimentally. For more information about GRNN, readers are referred to Specht (1991) and Wasserman (1993). The advantage of GRNN is that it can handle the noises in the input easily and use single-pass learning so no back proportion is needed.

**Adaptive neuro-fuzzy interference system (ANFIS)**

The configuration of ANFIS is displayed in Figure 1. There are five layers to this system (Sihag et al. 2019) and details of these five layers are as follows:

1. The membership degree is measured in the first layer. A membership degree is produced by every other node. The membership functions are being used in fuzzy sets.

\[ O_{i,j} = \mu_{A_i}(x) \quad i = 1, 2, \ldots \]  
\[ O_{i,j} = \mu_{B_{i-2}}(y) \quad i = 3, 4, \ldots \]

where \( x \) and \( y \) are the outputs, \( A_i \) and \( B_i \) are the linguistic labels, and \( \mu_{A_i} \) and \( \mu_{B_{i-2}} \) are the degrees of membership functions for \( A_i \) and \( B_{i-2} \), correspondingly.

![Figure 1 | Details of the ANFIS model.](http://iwaponline.com/ws/article-pdf/doi/10.2166/ws.2021.157/911180/ws2021157.pdf)
2. Based on the calculated membership degree, the performance of the second layer (fire strings) is derived. To produce the firing power, the previous layer's membership functions are multiplied together:

\[ O_{2i} = w_i = \mu_{A_i}(x) \cdot \mu_{B_i}(y) \]  

(4)

where \( O_{2i} \) is the output of this layer called the fire string.

3. At this stage, the firing capabilities are standardized. The steady nodes evaluate the contribution of the firing capabilities:

\[ O_{3i} = w_i = \frac{W_i}{w_1 + w_2} \]  

(5)

4. In this row, the ratio of the \( i \)th rule to the final result is evaluated using subsequent parameters:

\[ O_{4i} = \bar{w}_i = \bar{w}_i (p_i x + q_i y + r_i) \]  

(6)

where \( p_i, q_i, \) and \( r_i \) are the consequent parameters.

5. To obtain the total output, this layer uses a concise summary of input signals:

\[ O_{5i} = \sum \bar{w}_i \sum w_i f_i = \sum w_i \]  

(7)

It has been shown that a Gaussian membership function can lead to accurate outputs (Azimi et al. 2019; Ebt-e-haj et al. 2019; Gholami et al. 2019). Thus, the current study uses this membership function:

\[ \mu_{A_i}(x) = \left( \frac{(x - c_i)^2}{2\sigma_i^2} \right) \]  

(8)

where \( c_i \) and \( \sigma_i \) are the parameters for the membership function. The ANFIS model has the advantage of having both numerical and linguistic knowledge.

Area of the research work

The research project was done using flow and water quality data measured in a watershed consisting of three sub-watersheds: Khorramabad, Biranshahr, and Alashtar sub-watersheds, from Lorestan province, Iran, located between 48°03'10" E and 48°59'07" E, and between 33°11'47" N and 34°03'27" N with an area of 3,562.1 km². The elevation of the catchment area varies from 1,158 to 3,646 m above sea level. The observations were recorded from September 2014 to August 2017. Most of the rainfall occurs from November to May in a calendar year. The average rainfall is 442 mm, 484 mm, and 556 mm for Khorramabad, Biranshahr, and Alashtar sub-watersheds, respectively. Figure 2 shows the details of the study area and the black dots represent the sampling sites for the three watersheds.

Methodology and data descriptions

Three types of parameters – biological, physical, and chemical – were used to analyze the water quality by which the WQI was determined (Dogan et al. 2009). These parameters were total dissolved solids (TDS), sodium (Na), sulfate (SO₄), electrical conductivity (EC), calcium (Ca), the potential of hydrogen (pH), bicarbonate ions (HCO), chlorides (Cl), magnesium (Mg), and potassium (K). With the use of these parameters, the WQI was calculated as follows:

1. A weight (\( z_i \)) was assigned to each parameter on the scale of one to five according to its significance for drinking suitability and human health. The \( z_i \) values are given by Table 1 (Yidana et al. 2010; Varol & Davraz 2014; Vasant et al. 2016; Şener et al. 2017; Wagh et al. 2017).
2. The relative weight ($Z_i$) was determined for each of the parameters. Details of the $Z_i$ are tabulated in Table 1 and its formula is:

$$Z_i = \frac{z_i}{\sum_{i=1}^{n} z_i}$$

\hspace{1cm} (9)

**Table 1**  | Details of weight and relative weight of parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Weight ($z_i$)</th>
<th>Relative Weight ($Z_i$)</th>
<th>Standard values from WHO</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDS</td>
<td>5</td>
<td>0.15625</td>
<td>500</td>
</tr>
<tr>
<td>EC</td>
<td>2</td>
<td>0.0625</td>
<td>500</td>
</tr>
<tr>
<td>pH</td>
<td>4</td>
<td>0.125</td>
<td>6.5–8.5</td>
</tr>
<tr>
<td>HCO</td>
<td>3</td>
<td>0.09375</td>
<td>500</td>
</tr>
<tr>
<td>Cl</td>
<td>4</td>
<td>0.125</td>
<td>250</td>
</tr>
<tr>
<td>SO$_4$</td>
<td>4</td>
<td>0.125</td>
<td>250</td>
</tr>
<tr>
<td>Ca</td>
<td>3</td>
<td>0.09375</td>
<td>75</td>
</tr>
<tr>
<td>Mg</td>
<td>3</td>
<td>0.09375</td>
<td>50</td>
</tr>
<tr>
<td>Na</td>
<td>2</td>
<td>0.0625</td>
<td>200</td>
</tr>
<tr>
<td>K</td>
<td>2</td>
<td>0.0625</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>32</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

All values are in mg/L except pH (on a scale) and EC ($\mu$S/cm).

2. The relative weight ($Z_i$) was determined for each of the parameters. Details of the $Z_i$ are tabulated in Table 1 and its formula is:
3. A scale for the quality rating \( (s_i) \) was calculated for each of the parameters as:

\[
s_i = \left( \frac{\text{Con}_i}{\text{Std}_i} \right) \times 100
\]  

(10)

4. The sub-index level number \( (\text{SIL}_i) \) was calculated as:

\[
\text{SIL}_i = Z_i \times s_i
\]

(11)

5. The water quality index \( (\text{WQI}) \) was determined as:

\[
\text{WQI} = \sum \text{SIL}_i
\]

(12)

where \( \text{Con}_i \) is the concentration of parameters in mg/l, \( \text{Std}_i \) is the standard value of each parameter as per WHO, and \( n \) is the number of parameters.

A total of 124 observations were used which were observed from September 2014 to August 2017. The dataset consists of 10 input variables, pH, Na, Mg, SO4, K, TDS, K, Cl, HCO and EC, and 1 output variable, WQI. The pairs of all variables are represented in Figure 3 which shows the interrelation of all variables with each other and also gives information about the outliers which were not in large quantities. 70 percent of the entire dataset was used in a training stage of the soft computing techniques and 30 percent was used for testing the techniques. The characteristics of water quality parameters for the sub-watersheds are tabulated in Table 2. The characteristics of the three watersheds are similar except for some values which were higher in the Khorammabad watershed. Figure 4 shows the flow chart as well as the architecture of soft computing used in the investigation which suggests the implementation of ANN, GRNN and ANFIS in the study area.

**Correlation plot**

This investigation used the 10 input variables to estimate the output variable in the three watersheds. The correlation coefficients were obtained to evaluate the correlation among the output and input variables. Figure 5 shows the correlation plot of training and testing datasets. The WQI yielded a good correlation with TDS and EC (0.97
Table 2 | Characteristics of water quality parameters. All values are in mg/L, except pH (on a scale), EC (μS/cm) and WQI.

<table>
<thead>
<tr>
<th>Watershed</th>
<th>Parameter</th>
<th>TDS</th>
<th>EC</th>
<th>PH</th>
<th>HCO</th>
<th>Cl</th>
<th>SO₄</th>
<th>Ca</th>
<th>Mg</th>
<th>Na</th>
<th>K</th>
<th>WQI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biranshahr watershed</td>
<td>Mean</td>
<td>269.21</td>
<td>419.91</td>
<td>7.94</td>
<td>102.15</td>
<td>16.08</td>
<td>18.14</td>
<td>55.50</td>
<td>14.17</td>
<td>5.81</td>
<td>1.10</td>
<td>31.81</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>484.0</td>
<td>745.0</td>
<td>8.39</td>
<td>128.14</td>
<td>92.17</td>
<td>55.71</td>
<td>76.15</td>
<td>26.74</td>
<td>4.98</td>
<td>3.91</td>
<td>49.57</td>
</tr>
<tr>
<td></td>
<td>Min</td>
<td>170.0</td>
<td>275.0</td>
<td>7.05</td>
<td>73.22</td>
<td>6.03</td>
<td>4.32</td>
<td>26.05</td>
<td>2.43</td>
<td>0.69</td>
<td>0.00</td>
<td>25.42</td>
</tr>
<tr>
<td></td>
<td>Standard Deviation</td>
<td>48.74</td>
<td>75.15</td>
<td>0.31</td>
<td>12.81</td>
<td>12.72</td>
<td>11.08</td>
<td>10.35</td>
<td>6.04</td>
<td>6.69</td>
<td>0.85</td>
<td>4.20</td>
</tr>
<tr>
<td>Khorammabad watershed</td>
<td>Mean</td>
<td>413.24</td>
<td>651.55</td>
<td>7.46</td>
<td>140.93</td>
<td>52.04</td>
<td>22.79</td>
<td>70.28</td>
<td>26.89</td>
<td>16.95</td>
<td>5.46</td>
<td>44.17</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>573.00</td>
<td>977.0</td>
<td>8.53</td>
<td>201.37</td>
<td>116.99</td>
<td>77.33</td>
<td>110.2</td>
<td>69.28</td>
<td>64.37</td>
<td>44.1</td>
<td>64.24</td>
</tr>
<tr>
<td></td>
<td>Min</td>
<td>247.00</td>
<td>386.0</td>
<td>6.52</td>
<td>82.38</td>
<td>24.82</td>
<td>0.00</td>
<td>24.05</td>
<td>10.94</td>
<td>3.45</td>
<td>0.78</td>
<td>33.00</td>
</tr>
<tr>
<td></td>
<td>Standard Deviation</td>
<td>82.55</td>
<td>132.49</td>
<td>0.55</td>
<td>27.40</td>
<td>16.51</td>
<td>22.95</td>
<td>15.12</td>
<td>10.42</td>
<td>13.89</td>
<td>6.50</td>
<td>7.05</td>
</tr>
<tr>
<td>Alashtar Watershed</td>
<td>Mean</td>
<td>279.9</td>
<td>436.5</td>
<td>7.70</td>
<td>107.4</td>
<td>16.28</td>
<td>15.88</td>
<td>54.61</td>
<td>16.52</td>
<td>5.74</td>
<td>2.24</td>
<td>32.59</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>417.0</td>
<td>634.0</td>
<td>8.36</td>
<td>149.5</td>
<td>24.82</td>
<td>83.57</td>
<td>76.15</td>
<td>48.62</td>
<td>22.53</td>
<td>7.82</td>
<td>47.15</td>
</tr>
<tr>
<td></td>
<td>Min</td>
<td>150.0</td>
<td>241.0</td>
<td>6.47</td>
<td>56.4</td>
<td>3.55</td>
<td>2.40</td>
<td>24.05</td>
<td>3.65</td>
<td>0.69</td>
<td>0.00</td>
<td>23.65</td>
</tr>
<tr>
<td></td>
<td>Standard Deviation</td>
<td>64.2</td>
<td>101.9</td>
<td>0.49</td>
<td>23.6</td>
<td>4.81</td>
<td>14.13</td>
<td>15.46</td>
<td>8.21</td>
<td>4.87</td>
<td>2.28</td>
<td>4.98</td>
</tr>
</tbody>
</table>
and 0.97), good correlation with HCO, Cl and Mg (0.8, 0.86 and 0.77), average correlation with SO₄, Ca and Na (0.45, 0.58 and 0.54), and poor correlation with pH (−0.18) for the training dataset. Again, a good correlation with TDS and EC (0.97 and 0.97), good correlation with HCO, Cl and Mg (0.84, 0.72 and 0.87), average correlation with Ca, Na, K and pH (0.44, 0.47, 0.57 and −0.41), and poor correlation with SO₄ (0.22) were obtained for the testing dataset.
The performances of ANFIS, GRNN, ANN, and ANN-FFA were compared by model fitness criteria (MFC). These MFC were defined as:

- **Nash-Sutcliffe Efficiency**: Nash & Sutcliffe (1970) introduced Nash-Sutcliffe efficiency (NSE) which was used to evaluate the working of soft computing models. The range of NSE lies between $-\infty$ to 1. If NSE is equal to 1, it shows a perfect result of the model. Efficiency equal to 0 means the model is as accurate as the mean of the experimental value and a negative value indicates a better prediction than the model (Wilcox et al. 1990; Legates & McCabe 1999). NSE can be computed as:

$$NSE = 1 - \frac{\sum_{i=1}^{F} (D - E)^2}{\sum_{i=1}^{F} (D - D)^2}$$  \hspace{1cm} (13)$$

- **Mean Absolute Error**: The mean absolute error (MAE) is the mean of the absolute difference between observed and predicted values. The range of MAE is 0 to 1. The formula for MAE is:

$$MAE = \frac{1}{F} \sum_{i=1}^{F} |D - E|$$  \hspace{1cm} (14)$$

- **Legates-McCabe Index**: The Legates-McCabe Index (LMI) was introduced by Legates & McCabe (1999). The values of the LMI lie between 0 and 1. A lesser value indicates a bad result and vice versa. The LMI can be calculated as:

$$LMI = 1 - \frac{\sum_{i=1}^{z} |D - F|}{\sum_{i=1}^{z} |D - D|}$$  \hspace{1cm} (15)$$

- **Root Mean Square Error**: The root mean square error (RMSE) is one of the most used error methods to assess model fitness. As the name indicates, it is the square root of the mean square error. A zero value indicates the

---

**Figure 5** | Correlations plot of (a) Training and (b) Testing dataset. Please refer to the online version of this paper to see this figure in colour: [http://dx.doi.org/10.2166/ws.2021.157](http://dx.doi.org/10.2166/ws.2021.157).
best prediction and 1 indicates the worst case. The equation of RMSE is:

\[ RMSE = \sqrt{\frac{1}{F} \left( \sum_{i=1}^{F} (D - E)^2 \right)} \]  

(16)

• Mean Absolute Percentage Error: The mean absolute percentage error (MAPE) is also known as mean absolute percentage deviation. It can be formulated as:

\[ MAPE = \frac{100}{F} \sum_{i=1}^{F} \left| \frac{D - E}{D} \right| \]  

(17)

• Coefficient of Correlation: The coefficient of correlation (CC) shows the interrelation between observed and measured values. The range of CC is from -1 to +1. It can be calculated as:

\[ CC = \frac{F \sum DE - (\sum D)(\sum E)}{\sqrt{F(\sum D^2) - (\sum D)^2} \sqrt{F(\sum E^2) - (\sum E)^2}} \]  

(18)

where D represents the experimentally obtained values, E denotes the values obtained from the soft computing models, \( D \) denotes the mean of the experimentally observed values, and \( F \) is the number of the dataset.

RESULTS AND DISCUSSIONS

Three soft computing techniques, namely, ANN, GRNN, and ANFIS with four membership functions (ANFIS_trmf, ANFIS_trapmf, ANFIS_gbellmf, and ANFIS_gaussianmf), were used in this study. Analysis of soft computing techniques is a trial and error process. ANFIS and GRNN were executed by MatLab, while ANN used Weka (3.9). The soft computing techniques employ regulator parameters, and the model accuracy can increase or decrease by changing these parameters. Therefore, to improve accuracy, the optimal parameter values were determined by trial and error, and modeling was done thereafter. The optimal values of hyperparameters were as follows:

1. For ANN: Momentum = 0.2, learning rate = 0.1, hidden layer = 01, neuron per hidden layer = 08, iteration = 1,500.
2. For GRNN: Spread = 0.2.

Based on the optimal values of the hyperparameters, the soft computing techniques yielded MFC results for the prediction of WQI, and details of the MFC are summarized in Table 3. Table 3 suggests that all soft computing techniques produced good results for the training dataset, while ANN was the best in the case of testing the dataset with values of CC, LMI, NSE, RMSE, MAPE and MAE equal to 0.999, 0.9810, 0.9996, 0.1324, 0.2853, and 0.0987, respectively. The worst soft computing technique was ANFIS with trapmf membership function and with values of CC, LMI, NSE, RMSE, MAPE, and MAE of 0.3974, -0.3788, 2.6871, 12.6774, 20.5235, and 7.1767 respectively. Thus, the MFC values showed that ANN was the most accurate technique for the prediction of WQI, ahead of GRNN and ANFIS with four different membership functions.

The performance of WQI with soft computing techniques is plotted in Figure 6. For the training dataset, all the soft computing techniques worked well which is clearly shown in Figure 6 as only a few are outliers from the best agreement zone. Also, it followed the same trends as the actual WQI, except for some outliers. But, in the training dataset, ANN gave the best results, all of the ANN plot was in the best agreement zone and followed the same trend as the actual line. The plots of ANFIS_trapmf in the graphs and the line diagram did not follow the trends of actual WQI. Thus, Figure 6 indicates that ANN worked superbly in the prediction of the WQI.

The Taylor diagram, developed by Taylor in 2001 (Taylor 2001), is one of the most used and modern methods which shows how closely a pattern or set of patterns is summarized. It quantifies a set of patterns in terms of correlation, standard variation, and root mean square error. It is useful for solving the multi-aspects problem related to complex models. In this investigation, Taylor diagrams were constructed in ‘R’ language using the ‘plothrix’ package and Figure 7 shows the plot of a Taylor diagram with different soft computing techniques for the prediction of WQI. It also gives the same results as provided by the other plots. In training, all the soft computing
techniques worked well in the prediction of WQI but, in testing, ANN was the most accurate in prediction. Also, ANFIS\_trimf was the most inaccurate with a correlation of 0.42 (Figure 7). Thus, the Taylor diagram (Figure 7) also concluded that ANN was the most reliable soft computing technique for the prediction of WQI.

**Table 3** | Detail of MFC for training and testing data sets

<table>
<thead>
<tr>
<th>MFC</th>
<th>ANN</th>
<th>GRNN</th>
<th>ANFIS_trimf</th>
<th>ANFIS_trapmf</th>
<th>ANFIS_gbellmf</th>
<th>ANFIS_gaussianmf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>0.99996</td>
<td>0.99990</td>
<td>0.9999999</td>
<td>0.9964</td>
<td>0.999991</td>
<td>0.999999</td>
</tr>
<tr>
<td>LMI</td>
<td>0.98403</td>
<td>0.98400</td>
<td>0.9995241</td>
<td>0.9556</td>
<td>0.997197</td>
<td>0.998800</td>
</tr>
<tr>
<td>NSE</td>
<td>0.99976</td>
<td>0.99960</td>
<td>0.99999997</td>
<td>0.9928</td>
<td>0.999982</td>
<td>0.999997</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.08530</td>
<td>0.08962</td>
<td>0.0021884</td>
<td>0.2856</td>
<td>0.012357</td>
<td>0.006735</td>
</tr>
<tr>
<td>MAPE</td>
<td>0.29108</td>
<td>0.30010</td>
<td>0.0081858</td>
<td>0.9336</td>
<td>0.055585</td>
<td>0.022178</td>
</tr>
<tr>
<td>MAE</td>
<td>0.09498</td>
<td>0.09899</td>
<td>0.0028298</td>
<td>0.2641</td>
<td>0.016666</td>
<td>0.007138</td>
</tr>
<tr>
<td>Testing Dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>0.9999</td>
<td>0.9733</td>
<td>0.9795</td>
<td>0.3974</td>
<td>0.7986</td>
<td>0.9291</td>
</tr>
<tr>
<td>LMI</td>
<td>0.9810</td>
<td>0.7856</td>
<td>0.8900</td>
<td>−0.3788</td>
<td>0.6598</td>
<td>0.8104</td>
</tr>
<tr>
<td>NSE</td>
<td>0.9996</td>
<td>0.9452</td>
<td>0.9523</td>
<td>−2.6871</td>
<td>0.5687</td>
<td>0.8533</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.1324</td>
<td>1.5457</td>
<td>1.4414</td>
<td>12.6774</td>
<td>4.3361</td>
<td>2.5285</td>
</tr>
<tr>
<td>MAPE</td>
<td>0.2853</td>
<td>3.0443</td>
<td>1.4092</td>
<td>20.5235</td>
<td>4.3998</td>
<td>2.4634</td>
</tr>
<tr>
<td>MAE</td>
<td>0.0987</td>
<td>1.1161</td>
<td>0.5725</td>
<td>7.1767</td>
<td>1.7707</td>
<td>0.9871</td>
</tr>
</tbody>
</table>

**Figure 6** | The performance of the WQI with soft computing techniques. Please refer to the online version of this paper to see this figure in colour: http://dx.doi.org/10.2166/ws.2021.157.
The Violin cum box plot is a combination of the violin plot and box plot. In the violin cum box plot, the box plot is built between violin plots, giving the benefits of both of the plots. Just like the violin plot and box plot, the violin cum box plot also has multiple layers and marks the mean (horizontal bar in a box plot). The plot was created in the ‘R’ language environment using the ‘dplyr’ package. Figure 8 shows the violin cum box plot for the soft computing techniques for the prediction of WQI for training and testing datasets. Figure 8 shows that the violin...
cum box plot was approximately symmetrical with actual values except for ANFIS_trapmf for the training part but, for the testing part, all the plots of different soft computing techniques were different, with only the ANN plot being symmetrical with the actual one. Even the mean bar in the ANN plot was the same as for the actual WQI. Thus, the violin cum box plot (Figure 8) also supported the conclusion that ANN had an edge on GRNN and ANFIS in the prediction of WQI.

All three soft computing techniques predicted WQI, but in the case of ANN, greater points were placed in the agreement zone (Figure 6), the point of standard deviation was in an approximate overlap with the point of actual data (Figure 7), exactly the same multiple layers and mean bar with the actual one (Figure 8), and most accurate values of MFC (Table 3), and its reliability and accuracy were higher than that of the other two models. Therefore, this method can be used to predict WQI in the study area.

CONCLUSIONS

Water quality determination entails a number of parameters that have to be measured for calculating a WQI. The use of soft computing techniques in the prediction of WQI can reduce the time and make the process easy. Three soft computing techniques, ANN, GRNN, and ANFIS with four membership functions, were used for the prediction of WQI in Khorram Abad, Biranshahr, and Alashtar sub-watersheds in Iran. The effectiveness of these soft computing techniques was determined using scatter (x-y) plot, Taylor diagram, violin cum box plot, and model fitness criteria which showed that ANN produced a good agreement of predicted values with LMI and RMSE equal to 0.9810 and 0.1324. The effectiveness of ANN was also suggested by the Taylor diagram, violin cum box plot, and scatter plot. The mean bar and multiple layers of the violin cum box plot and standard variation of the Taylor diagram were the same as the actual values. Between the membership functions of ANFIS, ANFIS_trimf predicted WQI more accurately than other membership functions, while ANFIS_trapmf gave the worst results. Results indicated that with the sampling and measurement of several hydro-chemical parameters and the use of soft computing techniques, WQI can be predicted with high accuracy. Because the construction and maintenance of water quality measurement stations are expensive, their construction and maintenance as well as monitoring cannot be justified economically. In such a case, soft computing techniques can be used to assess water quality.

CONFLICT OF INTEREST

The authors have no conflict of interest.

DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories. See: https://drive.google.com/file/d/15nLuKd2gy1_usQhzpVooxg9tEyAoc-m/view?usp=sharing.

REFERENCES


First received 4 February 2021; accepted in revised form 15 May 2021. Available online 1 June 2021