Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2021.27773

238 (2021) 135–141
October

Modeling and analysis of the groundwater hardness variations process using machine learning procedure

Mahmood Yousefi[a], Ali Esrafili[a,b], Mitra Gholami[a,b], Ali Akbar Mohammadi[c], Nadeem A. Khan[d], Mansour Baziar[e,*], Vahide Oskoei[f,*]

[a]*Department of Environmental Health Engineering, School of Public Health, Iran University of Medical Sciences, Tehran, Iran, emails: Mahmood_yousefi70@yahoo.com (M. Yousefi), a.esrafili@iums.ac.ir (A. Esrafili), gholamim@iums.ac.ir (M. Gholami)*
[b]*Research Center for Environmental Health Technology, Iran University of Medical Sciences, Tehran, Iran*
[c]*Department of Environmental Health Engineering, Neyshabur University of Medical Sciences, Neyshabur, Iran, email: mohammadi.eng73@gmail.com*
[d]*Civil Engineering Department, Jamia Millia Islamia, New Delhi, India, email: er.nadimcivil@gmail.com*
[e]*Ferdows School of Paramedical and Health, Birjand University of Medical Sciences, Birjand, Iran, email: baziar.ehe@gmail.com*
[f]*Department of Environmental Health Engineering, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran, email: oskoei.v@gmail.com*

Received 21 April 2021; Accepted 3 September 2021

ABSTRACT

This paper focuses on applying artificial neural network (ANN) models to predict total hardness from groundwater. The input parameters of the neural network are electrical conductivity (EC) and pH, which are considered fast, measurable water quality factors. ANN-based Levenberg–Marquardt (trainlm) training algorithm has demonstrated exceptional ability to predict all data; in parallel, the excellent prediction was displayed by a different test dataset with *R* of 0.986 and 0.98079, respectively. The mean square error and mean absolute error for all datasets were considered to be 0.0011 and 0.0265, respectively; besides, their values for the other test dataset were acquired 0.0008 and 0.0243. Sensitivity analysis represented that EC plays a catch-all role in ANN models with the relative importance of 71%, while in contrast with the less important for pH by 29%.

Keywords: Artificial neural network; Total hardness; Modeling; Water quality

1. Introduction

Water is the most critical substance on the earth for humans, animals, and plants. Water has an abundant amount on the earth, but unfortunately, its small proportions are consumable by humans [1,2]. In some cases, the resource's chemical properties (hardness, heavy metals, soluble iron, nitrate contamination, etc.) do not satisfy acceptable levels. Water hardness is one of the most significant factors of water appropriateness for consumption, either on an industrial scale or domestic scale [3]. The use of water with a hardness of higher than standard value can lead to human health problems such as cardiovascular disorders [4]. Besides, in various industries, water consumption with hardness in higher values (hard water) may cause scale formation and process malfunction. The total hardness of more than 200 mg/L was classified as poor resources, while the unacceptable amount for household consumptions was more than 300 mg/L [5–7]. The calcium, carbonate ions, and magnesium on the earth layer dramatically affect groundwater's hydrochemistry. Magnesium and calcium ions of the earth layers react with the moisture and carbon dioxide, which cause

* Corresponding authors.

water hardness. The hardness degrees have been classified as temporary and permanent. Therefore, it is necessary to perceive the hardness of the water before consumption [8,9].

A neural network can be considered the computational system of simple interconnected processing elements, called the neuron, connected and respond to the network by the set of weights. The network architecture regulates the networks, the significance of the weights, and a processing element's modes involve in operation [10–12]. Artificial neural networks (ANNs) are known as promising tools in various sciences due to their user-friendliness in simulation, higher predictive performance in modeling and prediction than standard approaches [13–16]. ANNs were applied to predict several water quality indices such as sulfate, nitrate [17], sodium adsorption ratio (SAR) [10], etc. However, there is no study on the application of ANN in the prediction of total hardness (TH) using pH and electrical conductivity (EC) from groundwater. As a result, using these water quality parameters is a simple, easy, and cost-effective way to estimate water hardness concentration. This study aimed to develop an ANN model using EC and pH of groundwater to predict total water hardness.

2. Materials and methods

2.1. Analytical methods

In this presented study, we started to promptly measure some water parameters after testing in the region, which consists of EC and pH, by utilizing relative devices such as a conductivity meter and a portable pH meter. Furthermore, we adopted Varian flame atomic absorption spectrometer to evaluate cations' concentration like Ca^{2+} and Mg^{2+} [7,10]. The whole of 51 groundwater samples was collected and evaluated in the Meshkinshahr basin of Ardabil province.

TH in groundwater was determined by the equation (Eq. (1)) [7,10]:

$$\text{TH (as mg CaCO}_3\text{/L)} = (\text{Ca}^{+2} + \text{Mg}^{+2})\ \text{meq/L} \times 50 \qquad (1)$$

2.2. Artificial neural network

In this inquiry, a feed-forward ANN model was developed to predict the total hardness in the water resources of Meshkinshahr (Ardabil – Iran). By designing the model structure, the numbers of the neurons in output and input layers were demonstrated corresponding to the number of output and input variables. The multiple numbers of neural networks were evaluated during diversified training practices by packing the disparate number of neurons in the hidden layer of the ANN model. In the validation and testing phases, the accuracy of the models was adjusted according to the best exclusive model by mean square error (MSE) and *R* as statistical indices. In modeling, all data must also be split into two sections in the ratio 80:20; the first one with 80% was adopted to test, training and validation by portions of 15%, 70%, and 15%, respectively; and the latter with 20% was used for the supplementary test. For selecting the right algorithm as the key part of every process, we utilized seven backpropagation training algorithms such as Levenberg–Marquardt (trainlm), scaled conjugate gradient (trainscg), resilient backpropagation (trainrp), Polak–Ribière conjugate gradient (traincgp), one step secant (trainoss), Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-Newton (trainbfg) and gradient descent with momentum (traingdm) which was used with the same primary input data. Also, tangent sigmoid transfer function (tansig) and a linear transfer function (purelin) were applied for hidden and output layers, accordingly. The values of MSE were determined by utilizing the MSE equation. With several numbers series from 1 to 20, the neuron was tested to identify the most appropriate number of the hidden layer of the neural network. Our prestigious goal was minimizing the error and boosting the precision of the network weights and the output prediction; therefore, this modeling method was carried out by ten replications in three phases of validation, testing, and training.

2.3. Sensitivity analysis of the ANN model

The sensitivity analysis evaluated the significance of various effectual independent variables such as EC and pH in the ANN models of water. Sensitivity analysis is the practice of investigation in which input variables have the most significant influence on the outcomes (outputs) of the model. Determining the ANN model's most effective variable, the sensitivity analysis was administered in reliance on the Garson equation (Eq. (2)) [18]. This equation's basis depends on the obtained weights according to the best model of neural network and its partitioning.

$$I_j = \frac{\displaystyle\sum_{m=1}^{m=N_h}\left(\left(\frac{\left|W_{j_m}^{i_h}\right|}{\displaystyle\sum_{k=1}^{N_i}\left|W_{k_m}^{i_h}\right|}\right)\times\left|W_{m_n}^{h_o}\right|\right)}{\displaystyle\sum_{k=1}^{k=N_i}\left\{\sum_{m=1}^{m=N_h}\left(\frac{\left|W_{k_m}^{i_h}\right|}{\displaystyle\sum_{k=1}^{N_i}\left|W_{k_m}^{i_h}\right|}\right)\times\left|W_{m_n}^{h_o}\right|\right\}}\times 100 \qquad (2)$$

In the mentioned equation I_j is the related importance of the *j*th input variable on the output variable, N_h and N_i are the numbers of hidden and input neurons respectively, *W* is connection weight based on *h* indexes, *i*, *o*, and *h* attributed to the input, output, and hidden layers, *k*, *n*, and *m* are input, output, and hidden neurons respectively.

2.4. Analogy of the ANN model

The correlation coefficient (*R*), mean absolute error (MAE) and mean squared error (RMSE) were developed to assess the goodness of fit and accuracy prediction of the model. Overall, the developed model was deemed most relevant to high values of *R* and small values of MSE and MAE. These indices of MATLAB were presented by the mathematical equations as follows (Eqs. (3)–(5):

$$R = \text{corr (real total hardness values, anticipated total hardness values)} \qquad (3)$$

MSE = mean ((real total hardness value-anticipated total hardness value)^2) (4)

MAE = mean (abs (real total hardness value-anticipated total hardness va.lue)) (5)

3. Results and discussion

3.1. ANN model of TH

3.1.1. Backpropagation training algorithm choice

Better performance for algorithms is achievable to possess the outstanding backpropagation training algorithms by considering the small value of the MSE for any number of neurons in the distinctive training algorithms. Accordingly, the trainlm was considered the first and foremost ANN model for evaluating TH (Table 1).

3.1.2. Optimization of neuron number

In ANN models for optimizing the number of neurons, we considered the neurons with the minimum MSE proportions in the three phases of validation, training, and testing [10]. Depending upon the outcome, the smallest MSE (Table 2) was observed in trainlm algorithm for the neurons with the number of 8; therefore, it is considered as the perfect choice for TH.

Hence, configuration 2-8-1 (Fig. 1) appeared to be the most optimal topology for ANN models of TH, upon which 1 and 2 are the number of neurons in output and input layers, respectively. The type of transfer functions in hidden, and output layers were respectively tansig and purelin. The data normalization range was 0.1–0.9 (Eq. (6)).

$$y = 0.8\left(\frac{x_i - x_{\min}}{x_{\max} - x_{\min}}\right) + 0.1 \tag{6}$$

3.1.3. Model validation and tests

Generically, 30% of tests were considered to be used in test and validation phases (15% by each). The prediction performance of TH data and the ANN model was evaluated using Eq. (7).

Scatter plots (Fig. 2) presented the anticipated values of the TH vs. real ones, indicating the correlation coefficient prepared models in all data and the whole three phases. Fig. 2 shows the R values of the TH in three stages of validation, training, and test by the value of 0.9602, 0.99and 0.9753 accordingly, while for all datasets R-value stands at 0.9859.

Therefore, it can be concluded that the developed model can predict the total hardness values accurately. Also, acquired results for all data set was further analyzed by the determination coefficient (R^2). The results demonstrate that the organized model has the best ability to predict TH proportions up to 97.2%. Notably, the MSE portion for TH in the three phases of testing, validation, and

Table 2
Optimization of neuron number for ANN-based trainlm

Neuron	MSE			
	All data	Training	Validation	Testing
1	0.009875	0.012784	0.004203	0.001972
2	0.009758	0.01197	0.002501	0.006693
3	0.00949	0.012242	0.004046	0.002095
4	0.002767	0.003615	0.001272	0.000303
5	0.001769	0.001831	0.001453	0.001794
6	0.002376	0.002078	0.00379	0.002353
7	0.002852	0.000996	0.003366	0.010995
8	**0.001171**	**0.000918**	**0.002777**	**0.000742**
9	0.002247	0.002402	0.00137	0.002401
10	0.001263	0.00077	0.002031	0.002791
11	0.001376	0.000747	0.002361	0.003324
12	0.00148	0.001061	0.002783	0.002132
13	0.003707	0.000687	0.003095	0.018414
14	0.001657	0.000454	0.002777	0.006148
15	0.001755	0.000316	0.001744	0.00848
16	0.002369	0.002076	0.005527	0.000583
17	0.002221	0.002302	0.001329	0.005565
18	0.002421	0.00193	0.001169	0.005961
19	0.008159	0.005993	0.018879	0.007551
20	0.001371	0.000196	0.004343	0.003884

Table 1
Backpropagation training algorithm results

Backpropagation algorithms	Total hardness			
	R	MSE	Iteration number	Best hidden layer neuron
traincgp (Polak–Ribière conjugate gradient)	0.9396	0.002	27	5
traingdm (gradient descent with momentum)	0.8924	0.0109	1,000	9
trainscg (scaled conjugate gradient)	0.9344	0.0014	30	18
trainrp (resilient backpropagation)	0.9508	0.0019	27	18
trainoss (one step secant)	0.9517	0.0026	17	17
trainbfg (Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-Newton)	0.9731	0.0019	19	8
trainlm (Levenberg–Marquardt)	**0.986**	**0.0011**	**10**	**8**

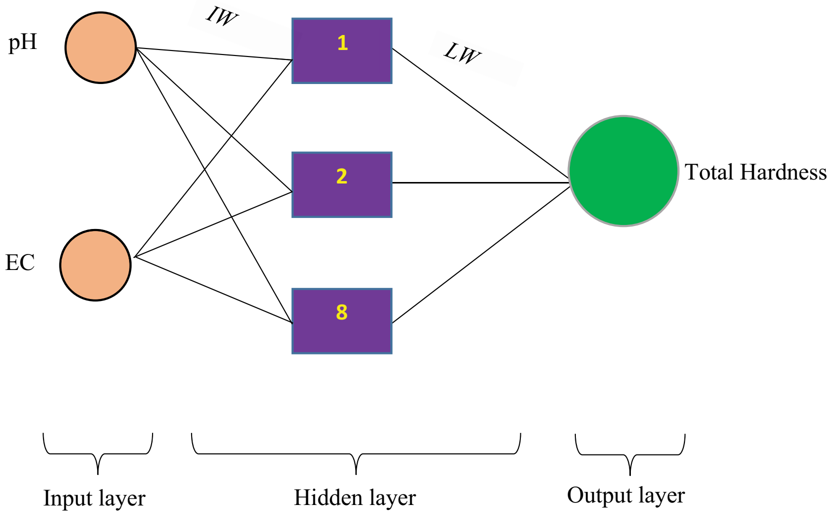

Fig. 1. Structure of 2-8-1 designated for the ANN model.

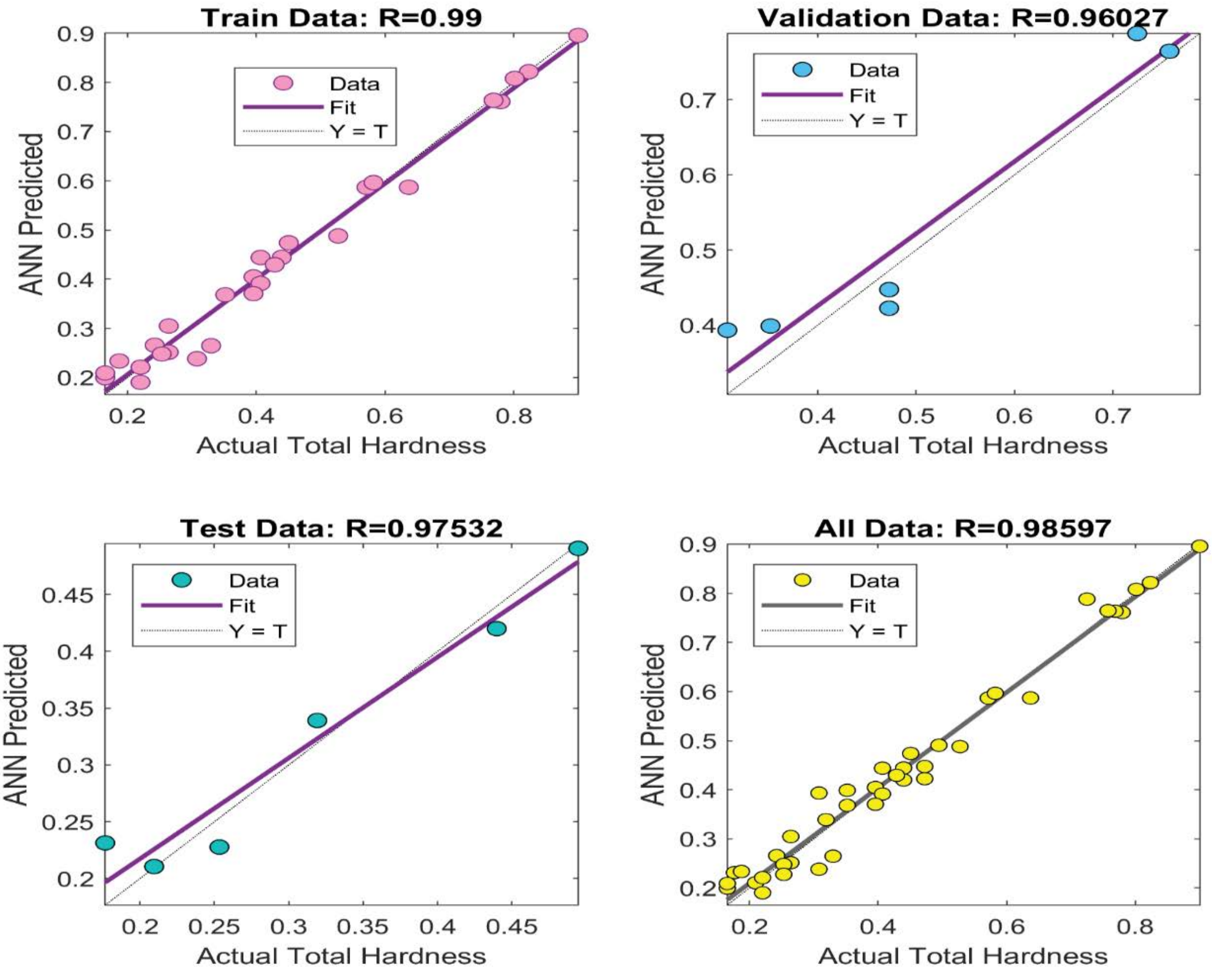

Fig. 2. Obtained scatter plots for training, validation, testing, and all datasets.

training was obtained to be 0.000742, 0.002777, and 0.000918 (Fig. 3). Furthermore, this presented picture has shown the MAE proportions and residual error.

The ANN model for TH prediction is presented in Eq. (7) as follows:

$$\text{ANN equation} = \text{purelin}\ \{W_2 \times \text{tansig}\ (W_1 \times [\text{pH; EC}] + b_1) + b_2\} \quad (7)$$

While Figs. 2 and 3 present the terrific linear fit and R^2 concerning the created-ANN model for all datasets successively.

$$y = 0.9726x + 0.0149 \quad (8)$$

$$R^2 = 0.9721 \quad (9)$$

where x and y are the real and estimated concentrations of TH, accordingly.

Also, in parallel, further analysis was organized to predict TH by adopting the ANN model. Based on the achieved outcome (Figs. 4 and 5), the MSE, MAE, and R values are 0.000871, 0.0243, and 0.9808, respectively. As claimed by Figs. 4 and 5, you can show the equations of the most exceptional linear fit (Eq. (10)) and R^2 (Eq. (11)) concerning the built-ANN model for additional datasets.

$$y = 0.9705x + 0.0265 \quad (10)$$

$$R^2 = 0.962 \quad (11)$$

Sulfate and SAR study in the aquifer of Southeastern Turkey by Yesilnacar and Sahinkaya [10] illustrated that the developed-ANN models had the highest prediction potential at the R-value of 0.956 and 0.98, respectively. A functioning study which was conducted by Balkaya et al. [19] appeared to illustrate the relationship between hardness and groundwater quality and hardness output as well. In the final developed ANN model, R-value was at the point of 0.591.

The study covers many applications of ANN models in various disciplines, which include effectively modeling a wide range of processes as their strong points.

3.2. Sensitivity analysis

Sensitivity analysis in the ANN models of TH investigated the significance of different useful independent

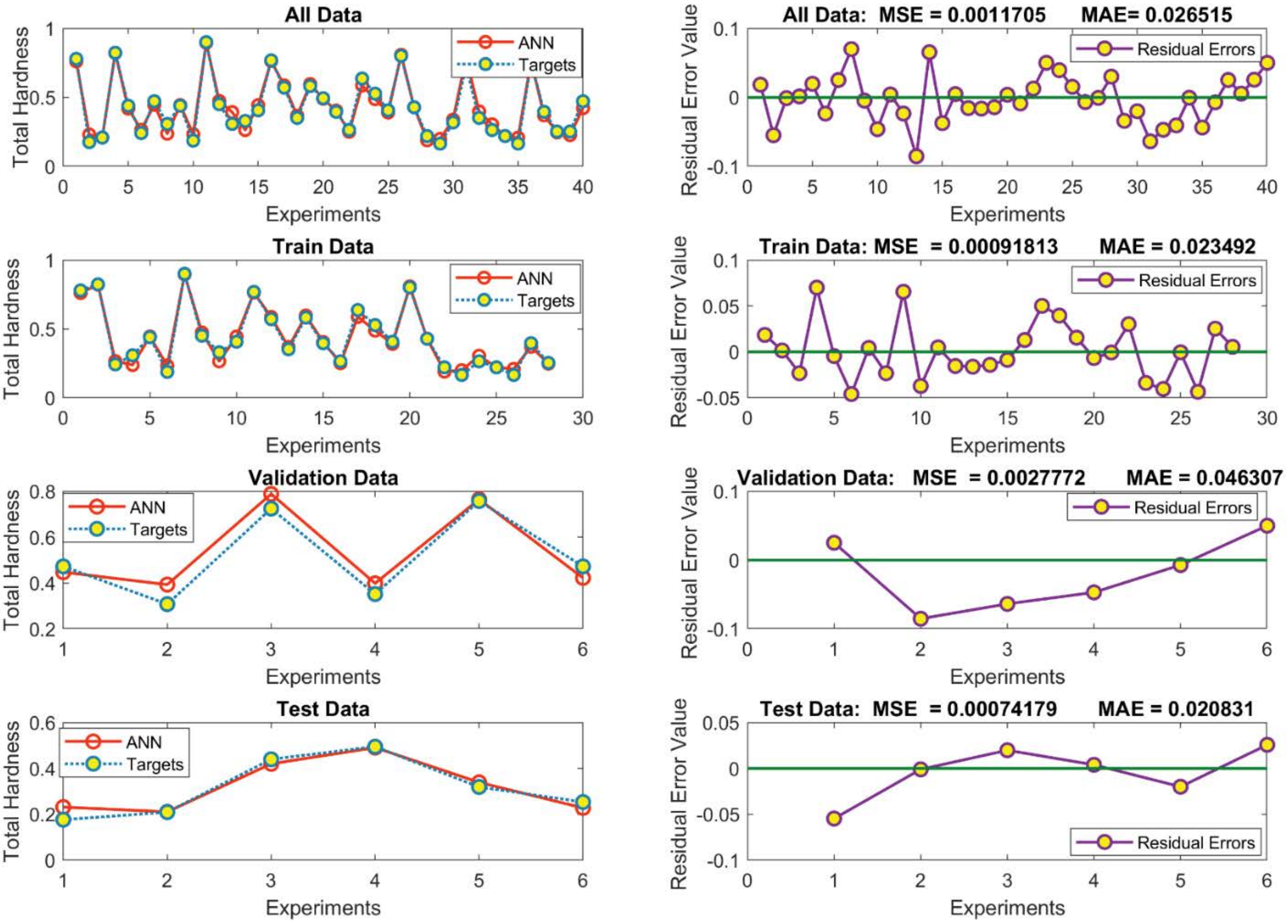

Fig. 3. Results of the best-developed model for training, validation, testing, and all dataset as well as their corresponding residual error values.

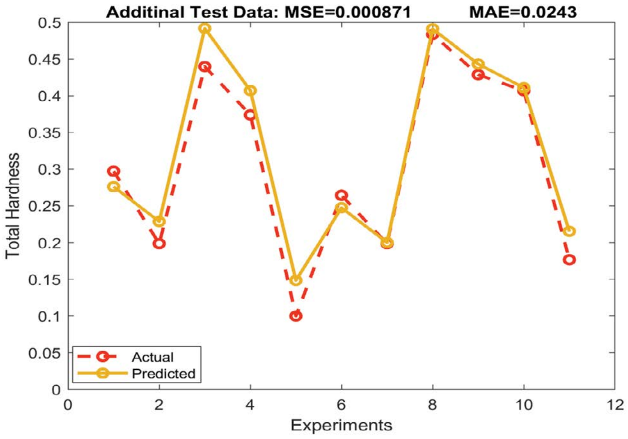

Fig. 4. Additional tests for TH using the ANN model.

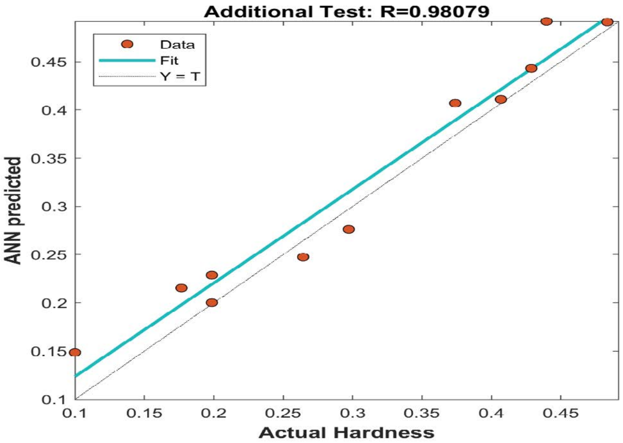

Fig. 5. Regression plot for additional test data.

Table 3
Weights and biases (of best developed ANN) for sensitivity analysis

IW (Input weights)		LW (layer weights)	Bias	
pH	EC	Total hardness	b_1	b_2
–2.14611	3.463785	–0.083719318	3.785308	–0.145
–2.27431	–2.77142	–1.080027204	4.219216	
–0.13604	–3.19683	–0.515247266	–1.95893	
–1.83045	4.537248	–0.719177722	–0.22429	
0.544139	–4.09444	–1.347067897	0.499874	
0.681557	–3.9582	1.126799185	2.342311	
1.83134	–3.39312	–0.339328606	2.968117	
3.001223	1.453424	0.565696333	4.649225	

variables (pH and EC). This approach was conducted by the connection weights of the created models. Table 3 illustrates the matrix of the neural network weights for TH.

Based on the obtained results, it is evident that EC and pH possess an essential role in predicting TH. By and large, the maximum effectiveness belongs to the EC with a value of 71%. While in contrast with the less important for pH by 29%.

4. Conclusions

As demonstrated in the study, the ANN as the best model offered the most promising results in predicting TH concentrations of groundwater wells located at the Meshkinshahr basin of Ardabil province. ANN was utilized to generate a model for predicting the proportion of TH. Experimental datasets designed the developed computer models to evaluate the various levels of pH and EC. Based on the presented sensitivity analysis results, it was evident that we can characterize EC as the significant parameter. Moreover, MSE, MAE, *R*, and R^2 are the statistical indices we used to predict the developed models.

The submitted ANN-trainlm model determines excellent prediction capability with high precision for forecasting TH than that other BP-ANN models.

Ethical approval

The authors of this article have covered all the ethical points, including non-plagiarism, duplicate publishing, data distortion, and data creation in this article. This project has been registered in Birjand University of Medical Sciences with the ethics code of IR.BUMS.REC.1400.125.

Conflict of interest

The authors of this article declare that they have no conflict of interest.

References

[1] H. Rezaei, A. Zarei, B. Kamarehie, A. Jafari, Y. Fakhri, F. Bidarpoor, M.A. Karami, M. Farhang, M. Ghaderpoori, H. Sadeghi, N. Shalyari, Levels, distributions and health risk assessment of lead, cadmium and arsenic found in drinking groundwater of Dehgolan's villages, Iran, Toxicol. Environ. Health Sci., 11 (2019) 54–62.

[2] A. Badeenezhad, H.R. Tabatabaee, H.-A. Nikbakht, M. Radfard, A. Abbasnia, M.A. Baghapour, M. Alhamd, Estimation of the groundwater quality index and investigation of the affecting factors their changes in Shiraz drinking groundwater, Iran, Groundwater Sustainable Dev., 11 (2020) 100435, doi: 10.1016/j.gsd.2020.100435.

[3] M. Yousefi, H.N. Saleh, M. Yaseri, M. Jalilzadeh, A.A. Mohammadi, Association of consumption of excess hard water, body mass index and waist circumference with risk of hypertension in individuals living in hard and soft water areas, Environ. Geochem. Health, 41 (2019) 1213–1221.

[4] S. Marque, H. Jacqmin-Gadda, J.-F. Dartigues, D. Commenges, Cardiovascular mortality and calcium and magnesium in drinking water: an ecological study in elderly people, Eur. J. Epidemiol., 18 (2003) 305–309.

[5] Z.Q. Lateef, A.-S.T. Al-Madhhachi, D.E. Sachit, Evaluation of water quality parameters in Shatt AL-Arab, Southern Iraq, using spatial analysis, Hydrology, 7 (2020) 79, doi: 10.3390/hydrology7040079.

[6] N. Khatri, S. Tyagi, D. Rawtani, M. Tharmavaram, R.D. Kamboj, Analysis and assessment of ground water quality in Satlasana Taluka, Mehsana district, Gujarat, India through application of water quality indices, Groundwater Sustainable Dev., 10 (2020) 100321, doi: 10.1016/j.gsd.2019.100321.

[7] M. Mirzabeygi, M. Naji, N. Yousefi, M. Shams, H. Biglari, A.H. Mahvi, Evaluation of corrosion and scaling tendency indices in water distribution system: a case study of Torbat Heydariye, Iran, Desal. Water Treat., 57 (2016) 25918–25926.

[8] M.H. Paller, S.M. Harmon, A.S. Knox, W.W. Kuhne, N.V. Halverson, Assessing effects of dissolved organic carbon and water hardness on metal toxicity to *Ceriodaphnia dubia* using diffusive gradients in thin films (DGT), Sci. Total Environ., 697 (2019) 134107, doi: 10.1016/j.scitotenv.2019.134107.

[9] C. Merz, G. Lischeid, Multivariate analysis to assess the impact of irrigation on groundwater quality, Environ. Earth Sci., 78 (2019) 1–11.

[10] M.I. Yesilnacar, E. Sahinkaya, Artificial neural network prediction of sulfate and SAR in an unconfined aquifer in southeastern Turkey, Environ. Earth Sci., 67 (2012) 1111–1119.

[11] D.P. Strik, A.M. Domnanovich, L. Zani, R. Braun, P. Holubar, Prediction of trace compounds in biogas from anaerobic digestion using the MATLAB Neural Network Toolbox, Environ. Modell. Software, 20 (2005) 803–810.

[12] E. Reyes-Télleza, A. Parralesb, G. Ramírez-Ramosa, J. Hernándezc, G. Urquizac, M. Herediaa, F. Sierrac, Analysis of transfer functions and normalizations in an ANN model that predicts the transport of energy in a parabolic trough solar collector, Desal. Water Treat., 200 (2020) 23–41.

[13] D. Millán-Ocampo, J. Porcayo-Calderón, A. Álvarez-Gallegos, J. Solís-Pérez, J. Hernández-Pérez, S. Silva-Martínez, Electrochemical deposition of copper using a modified electrode with polyaniline film: experimental analysis and ANN-based prediction, J. Taiwan Inst. Chem. Eng., 123 (2021) 272–283.

[14] M. Baziar, R. Nabizadeh, A.H. Mahvi, M. Alimohammadi, K. Naddafi, A. Mesdaghinia, Application of adaptive neural fuzzy inference system and fuzzy C-means algorithm in simulating the 4-chlorophenol elimination from aqueous solutions by persulfate/nano zero valent iron process, Eur. J. Anal. Chem., 13 (2018) em03, doi: 10.12973/ejac/80612.

[15] M. Baziar, R. Nabizadeh, A.H. Mahvi, M. Alimohammadi, K. Naddafi, A. Mesdaghinia, H. Aslani, Effect of dissolved oxygen/nZVI/persulfate process on the elimination of 4-chlorophenol from aqueous solution: Modeling and optimization study, Korean J. Chem. Eng., 35 (2018) 1128–1136.

[16] H.R. Zakeri, M. Yousefi, A.A. Mohammadi, M. Baziar, S.A. Mojiri, S. Salehnia, A. Hosseinzadeh, Chemical coagulation-electro Fenton as a superior combination process for treatment of dairy wastewater: performance and modelling, Int. J. Environ. Sci. Technol., (2021) 1–14, doi: 10.1007/s13762-021-03149-w.

[17] M.I. Yesilnacar, E. Sahinkaya, M. Naz, B. Ozkaya, Neural network prediction of nitrate in groundwater of Harran Plain, Turkey, Environ. Geol., 56 (2008) 19–25.

[18] M. Baziar, R. Nabizadeh, A.H. Mahvi, K. Naddafi, A. Mesdaghinia, M. Alimohammadi, H. Aslani, Sensitivity analysis and modeling of 4-chlorophenol degradation in aqueous solutions by an nZVI-sodium persulfate system, Desal. Water Treat., 112 (2018) 292–302.

[19] N. Balkaya, H. Kurtulus Ozcan, O. Nuri Ucan, Determination of relationship between hardness and groundwater quality parameters by neural networks, Desal. Water Treat., 11 (2009) 258–263.