Comparative analysis of machine learning techniques for estimating groundwater deuterium and oxygen-18 isotopes

Bilal Cemek 1 · Hakan Arslan 1 · Erdem Küçüktopcu 1 · Halis Simsek 2

Accepted: 1 June 2022 / Published online: 29 June 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
Isotope techniques are most frequently used when hydrochemical analysis are insufficient to determine the origin and quality of groundwater and reveal seawater intrusion into groundwater along coastlines. In this study, the potential of the multilayer perceptron, adaptive neuro-fuzzy inference system, generalized regression neural networks, radial basis neural networks, classification and regression tree, Gaussian process regression, multiple linear regression analysis, and support vector machines were compared using known hydrochemical properties of waters for estimating deuterium (δD) and oxygen-18 (δ18O) isotopes in groundwater of the Bafra plain, Northern Turkey. The data were divided into training (70%) and testing (30%) sets. Cluster analysis was performed to decrease the number of input variables. The data on electrical conductivity, chloride, magnesium, and sulfate were introduced into the models after examining different combinations of these variables in the studied models. The determination coefficient (R^2), mean absolute error (MAE), and root mean square error (RMSE) were used to evaluate the performances of the models. In addition, visualization techniques (Taylor diagram and heat maps) were prepared to assess the similarities between the measured and estimated δD and δ18O values.

The R^2, RMSE, and MAE for δ18O (0.98, 0.31 and 0.20%, respectively), and δD (0.95, 2.85 and 1.89%, respectively) values for the testing datasets revealed that the performance accuracy of multilayer perceptron is the best among the applied models tested. Therefore, the study suggests using data-driven methods, multilayer perceptron in this case, when lacking appropriate laboratory isotope analysis or facing high laboratory analysis costs.

Keywords Artificial intelligence · Isotope · Deuterium · Oxygen-18 · Groundwater

1 Introduction
The precipitation, evaporation, soil type, geological structure, and human activities have a significant impact on the quality of groundwaters. The water used in irrigation, residential areas, and industrial facilities substantially increased due to the constantly growing human population worldwide. Excessive use of groundwater as a water source in coastal areas disturbs the water balance between fresh and saline water and triggers seawater intrusion into groundwater. Excessive groundwater use causes groundwater to be salinized and it negatively affects plant growth and soil salinization (Libutti et al. 2018; Ding et al. 2020).

Understanding the origin of salinity and the dynamics of saline-water intrusion in aquifers, especially under the threat of saline water intrusion is of great importance for the sustainable groundwater management. In addition to classical hydrogeology techniques, isotope hydrology methods are also used to determine the regions with saline water intrusion. The distribution of deuterium (δD) and oxygen-18 (δ18O) isotopes in groundwater or surface waters is widely used to determine the origin of waters and the relationships between water resources (Jayathunga...
et al. 2020). The $\delta^{18}$O and $\delta^2$H, among the environmental isotopes, are good tracers as they are natural water components. The studies on water quality assessment revealed that high volume of groundwater abstraction, excessive pumping, and less recharge in coastal wells lead to seawater intrusion, thus increased groundwater salinization (Klassen and Allen 2017; Mohanty and Rao 2019). Many studies indicated a significant relationship between the stable isotopes and other hydrochemical properties of groundwater (Lin et al. 2011; Arslan et al. 2012; Mongelli et al. 2013; Isawi et al. 2016; Bahir et al. 2018; Jahnke et al. 2019). However, isotopes analysis is costly, and very few laboratories are equipped to analyze them.

Artificial intelligence (AI) techniques such as artificial neural networks (ANN), adaptive neuro-fuzzy inference system (ANFIS), and support vector machine (SVM) are widely used to solve complex problems in a various research field, thus gaining popularity (Odabas et al. 2017; Ahsan et al. 2022). Complex nonlinear systems can be easily modeled using AI techniques that compensate for numerical models’ weaknesses. The AI techniques have recently been employed successfully to estimate water quality parameters. Barzegar and Moghaddam (2016) evaluated the performances of three different ANN models including, multi-layer perceptron (MLP), generalized regression neural network (GRNN), and radial basis function neural network (RBNN), to estimate groundwater salinity in the Tabriz plain. Hameed et al. (2017) predicted the water quality index using a backpropagation neural network (BPNN) and RBNN. Lal and Datta (2018) employed genetic programming (GP) and gaussian process regression (GPR) models in modeling groundwater salinity. Juntakut et al. (2019) used the classification and regression tree (CART) to predict the nitrate concentration of groundwaters in Nebraska. Jafari et al. (2019) estimated the total dissolved solid contents of the groundwater aquifers in Tabriz plain using four AI techniques, including MLP, ANFIS, SVM, and gene expression programming (GEP). Noori et al. (2020) combined a process-based watershed model and an ANN to improve water quality predictions. Cahyadi et al. (2020) applied the BPNN model to estimate the hydraulic conductivity of fractured groundwater flow media.

Previous studies made significant contributions to the improve AI technology to evaluate water quality parameters. However, the AI techniques have not been used to estimate the distribution of $\delta^D$ and $\delta^{18}$O isotopes in groundwater. Hence, the estimation of $\delta^D$ and $\delta^{18}$O in groundwater using AI models will fill the gap in the literature as a novel modeling approach. Overall, the objective of this study was to develop a simple, low-cost, and accurate model to estimate the $\delta^D$ and $\delta^{18}$O isotopes in groundwater by using different data-driven models, including MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR.

2 Material and methods

2.1 Study area and dataset

The study was conducted in Kizılrımak Delta of Bafra plain, which was located in the middle part of the Black Sea region of northern Turkey and situated in $41^\circ30'–41^\circ45'$ north latitudes, $35^\circ30'–36^\circ15'$ east longitudes (Fig. 1). The soils in the study area formed over young alluviums deposited by the Kizılrımak River. Most of the soils in the area are hydromorphic, while coastal dunes are found at the seaside and colluvial-alluvial mixtures in the inland. Overall, the soils in the plain are young and have been extensively used in agricultural production.

The climate is semi-humid with an annual average temperature of 14.40 °C. Annual precipitation ranges between 536.4 and 783.5 mm. The elevation rises from the coast to the inland and reaches 10 m within about 6 km.
The differences in elevation cause significant fluctuations in groundwater levels and drainage within the study area. Irrigation is carried out mainly by using border and furrow irrigation methods. About 75% of the agricultural fields in the plain are irrigated using surface waters, while 25% of them are using groundwater (Cemek et al. 2007).

A total of sixty-one water samples were collected between October 2007 and September 2008 in different locations in the study area. Fifty-six water samples were collected from the twenty-eight different monitoring wells in Bafra plain and five from the Black Sea. The water samples were filtered using a 0.45 μm filter, filled in polyethylene bottles, and stored at 4 °C until processing. Electrical conductivity (EC) and pH of water samples were measured in a portable handheld probe in the field. Major cation (K⁺, Na⁺, Ca²⁺, and Mg²⁺) and anion (Cl⁻ and SO₄²⁻) concentrations were analyzed using the ion chromatography (761 Compact IC, Metrohm Schweiz AG, Switzerland) method. The HCO₃⁻ concentration of water samples was determined by the titration method.

The isotopes (δD and δ¹⁸O) analysis was carried out at the General Directorate of Hydraulic Work (DSI) in the Department of Technical Research and Quality Control Laboratory in Turkey. The isotopic composition is usually expressed in δ per mil, i.e., deviation % of the isotope ratios ²H/¹H and ¹⁸O/¹⁶O from the reference Vienna Standard Mean Ocean Water (V-SMOW). Statistical summary of water analysis results is given in Table 1. The δ¹⁸O values ranged from −3.59 to −10.69%, with an average of −8.16%, and the δD values ranged from −19.56 to −72.64%, with an average of −50.42%. The EC values varied between 1.86 and 28.80 dS m⁻¹ with an average of 5.96 dS m⁻¹. The pH values were between 7.28 and 8.16, having an average of 7.65. The concentrations of Ca²⁺, Mg²⁺, K⁺, and Na⁺ ranged from 37.00 to 305.00 mg L⁻¹, 46.00 to 839.00 mg L⁻¹, 4.00 to 200.00 mg L⁻¹, and 247.70 to 6900.00 mg L⁻¹, respectively. Average concentrations of these major cations were 102.44, 170.52, 35.03, and 1170.21 mg L⁻¹, respectively. The concentrations of Cl⁻, SO₄²⁻, and HCO₃⁻ vary from 345.34 to 9800.00 mg L⁻¹ (average, 1551.02 mg L⁻¹), 40.00 to 1950.00 mg L⁻¹ (average 419.34 mg L⁻¹), and 174.44 to 1617.00 mg L⁻¹ (average 840.07 mg L⁻¹), respectively.

### 2.2 Data-driven techniques

#### 2.2.1 Multilayer perceptron (MLP)

The MLP is one of the most frequently used neural network types (Fig. 2a). This network transmits the information from the input layer neurons to the hidden layer(s) neurons and finally to the output layer neurons (Skansi 2018).

Levenberg–Marquardt (LM) learning algorithm was employed in this study (El-Bakry 2003; Cigizoglu and Kişi 2005). The hidden and output layers were constructed using tangent sigmoid (tansig) and linear transfer (purelin) functions. The number of hidden nodes was increased from three to seven to optimize the training network.

#### 2.2.2 Radial basis neural networks (RBNN)

The RBNN was introduced into the neural network literature by Broomhead and Lowe (1988) to represent an efficient mechanism for complex nonlinear functions of local approximators (Mai-Duy and Tran-Cong 2003), pattern classification and recognition (Umasankar and Kalairaisi 2014), and modeling and controlling dynamic systems (Zhao 2008) from the input–output data. The RBNN is distinguished from other neural networks due to its universal approximation and shorter training phase. The RBNN networks are special feed-forward neural networks composed of three layers (Fig. 2b). The first layer (input) represents the input data. The second layer (hidden) comprises several nodes that implement a nonlinear transformation to the input variables. The last layer corresponds to the final output of the network and uses a linear activation function.

| Table 1 |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| Descriptive statistics of some chemical parameters of groundwater in Bafra plain | | | | | | | |
| Parameters | Unit | Avg. | Max. | Min. | SD | Skew | Kurt |
| δ¹⁸O | % | −8.16 | −3.59 | −10.69 | 1.56 | 1.67 | 3.50 |
| δD | % | −50.42 | −19.56 | −72.64 | 10.89 | 1.37 | 3.07 |
| EC | dS/m | 5.96 | 28.80 | 1.86 | 6.20 | 2.49 | 5.61 |
| pH | – | 7.65 | 8.16 | 7.28 | 0.21 | 0.75 | 0.14 |
| Ca | mg/L | 102.44 | 305.00 | 37.00 | 54.46 | 1.34 | 2.26 |
| Mg | mg/L | 170.52 | 839.00 | 46.00 | 159.82 | 2.62 | 6.82 |
| K | mg/L | 35.03 | 200.00 | 4.00 | 46.59 | 2.27 | 4.67 |
| Na | mg/L | 1170.21 | 6900.00 | 37.00 | 1534.13 | 2.52 | 5.72 |
| Cl | mg/L | 1551.02 | 9800.00 | 345.34 | 2147.42 | 2.61 | 6.14 |
| SO₄ | mg/L | 419.34 | 1950.00 | 40.00 | 400.85 | 2.20 | 5.02 |
| HCO₃ | mg/L | 840.07 | 1617.00 | 174.44 | 399.34 | 0.30 | −0.78 |
function. Extensive detail on the RBNN method can be obtained from Haykin (2001).

2.2.3 Generalized regression neural networks (GRNN)

The GRNN, initially proposed and developed by Specht (1991), comprises four layers: input, pattern, summation and output (Fig. 2c). The total number of variables is equal to the number of input units in the first layer. Each unit in the pattern layer represents a training pattern. The S and D-summation neurons are connected to each pattern layer in the summation layer. The S-summation neuron adds the product of the pattern layer’s weight outputs, whereas the D-summation neuron adds all weights. The final layer divides the output of the S-summation neuron by the output of the D-summation neuron and generates the predicted output. The distinguishing feature of GRNN is its use of a smoothing factor, which affects the generalization of the network. Low smoothing factors impair the generalizability of the network, whereas high smoothing factors frequently increase the prediction error while are helpful in generalization. Detailed information on GRNN can be obtained from Haykin (2001).

2.2.4 Adaptive neuro-fuzzy inference system (ANFIS)

This technique combines the learning ability of ANN and relational structure with the decision-making mechanism of the fuzzy inference system (FIS). Similar to the ANN, the ANFIS learns with samples from a training set. This method generates the most appropriate ANN structure for resolving the problem at hand. The test process is used to determine the effect of structure on previously unobserved samples. The low error values demonstrate the applicability of a model. However, one of the primary shortcomings of the ANN is the lack of justification for weight values acquired. The FIS, which is integrated into the structure of the ANFIS, addresses this shortcoming. The structure of ANFIS allows to be applied in many real-world issues. Daneshmand et al. (2015) provided the details on ANFIS.
2.2.5 Support vector machines (SVM)

The SVM, a binary learning machine, can be used for pattern recognition and nonlinear regression problems. The SVM is a two-layer network composed of nonlinear weights in the first layer and linear weights in the second (Bray and Han 2004). The SVM can efficiently handle nonlinear classification/modeling using the kernel method.

Given a training data of \( n \) samples \( \{z_i = (x_i, y_i), \quad i = 1, 2, 3, \ldots, n\} \), where \( x_i \in \mathbb{R}^d \) is the training data and \( y_i = R \) is the response for \( x_i \). The following quadratic optimization problem can be written as follow by solving the Kuhn-Tucker conditions.

\[
\begin{align*}
\text{Min} & \quad \frac{1}{2} ||\omega||^2 + C \sum_{i=1}^{n} (\xi_i^+ + \xi_i^-) \\
\text{Subject to} & \quad a_i \phi(x_i) + b - y_i \leq \epsilon + \xi_i^-, \quad \xi_i^+ \geq 0 \\
& \quad \forall i, \quad i \in (1, 2, \ldots, n)
\end{align*}
\]

where \( 1/2o^2 \) is the regularization term, \( \phi(x_i) \) is the nonlinear mapping function, \( C \) is the error penalty factor to regulate the difference between the empirical error and the linear mapping function, \( \epsilon \) is the loss function, \( b \) is scalar, and \( \omega \) represents a normal vector.

The best decision function \( f(x) \) can be stated as;

\[
f(x) = \sum_{i=1}^{N} (a_i - a_i^+)k(x_i, x) + b
\]

where \( a_i, a_i^+ (i = 1, 2, 3, \ldots, n) \) are the Lagrange multipliers, and \( k(x_i, x) \) is the kernel function.

Different kernel functions (i.e., linear, polynomial, Gaussian, exponential) were employed in the SVM. A polynomial kernel function was used in this study (Achirul G. et al. 2019). The readers may refer to Vapnik (1995) and Vapnik et al. (1997) for more information on SVM theoretical basis.

2.2.6 Gaussian process regression (GPR)

The GPR is a widely used nonparametric kernel-based machine learning technique that clarifies responses by presenting latent variables \( f(x_i), i = 1, 2, 3, \ldots, n \) derived from a Gaussian process (GP) and explicit basis functions. A GP is a finite number of random variables with a joint Gaussian distribution.

A GP \( f(x) \) is specified by a mean \( m(x) \) function and covariance function which are defined as follows (Williams and Rasmussen 1996):

\[
m(x) = E(f(x))
\]

\[
k(x_i, x_j) = E\left[\{f(x_i) - m(x_i)\}\{f(x_j) - m(x_j)\}\right]
\]

when \( f(x) \sim GP(0, k(x_i, x_j)) \), the GPR model can be defined as:

\[
P(f|x) \sim N(f|0, K(X,X))
\]

The covariance function is described by different kernel functions, which can be configured in kernel parameters in vector \( \theta \); therefore, the covariance function is stated as \( k(x_i, x_j|\theta) \). This study employed four kernel functions, rational quadratic, exponential, squared exponential, and matern 5/2, to estimate the \( \deltaD \) and \( \delta^{18}O \). Detailed information on the GPR can be obtained from Rasmussen (2003).

2.2.7 Classification and regression tree (CART)

Breiman et al. (1984) developed the CART, a nonparametric modeling approach. This technique is one of most commonly used methods in groundwater research management (Naghibi et al. 2016; Zhao et al. 2016; Choubin et al. 2019; Juntakut et al. 2019; Knoll et al. 2019) as it is easy to understand and interpret (Genç et al. 2015). The process begins by subdividing the data into subgroups, resulting in more homogeneous samples in the child nodes compared to the parent nodes. This process is repeated until the CART reaches a steady state in which the parent nodes do not significantly improve the capacity of the entire tree. The CART can effectively estimate the outcome in regression problems by using the least-square deviation criteria. The CART model is not a black box model, therefore, the effect of each variable on the desired output can be visualized clearly in the corresponding tree structure. Detailed information on the CART model can be obtained from Breiman et al. (1984) and Chipman et al. (1998).

2.2.8 Multiple linear regression analysis (MLR)

Multiple linear regression (MLR) is a statistical technique that predicts the outcome of a response variable by combining several explanatory variables. The MLR attempts to model the linear relationship between independent and dependent variables. General equation of a MLR model is as follows:

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon
\]

where \( y \) is the dependent variable; \( x_1, x_2, \ldots, x_n \) are the independent variables; \( \beta_1, \beta_2, \ldots, \beta_n \) are the slope
coefficients for each independent variable; $\beta_0$ is the constant term; $\epsilon$ is the error term of the model.

### 2.3 Cluster analysis

Cluster analysis is a technique to group related observations into a set of clusters based on the observations of multiple variables for each individual. The cluster technique has effectively been used to classify water samples in many studies (Kazakis et al. 2017; Rao and Chaudhary 2019; Egbueri 2020; Yang et al. 2020). In this study, the cluster analysis was used to group groundwater samples for $\delta^{18}$O, $\delta$D, EC, pH, Ca, Mg, K, Na, Cl, SO$_4$, and HCO$_3$ content.

### 2.4 Data pre-processing

Before developing the model structure, the input and output data were standardized between 0 and 1 to certify the equal handling of all variables and enhance the efficiency of the training network. The following equation is used to normalize data:

$$Z_{\text{norm}} = \frac{(Z_{\text{mea}} - Z_{\text{min}})}{(Z_{\text{max}} - Z_{\text{min}})}$$  \hspace{1cm} (8)

where $Z_{\text{norm}}$ is the standardized value, $Z_{\text{mea}}$ is the measured value, and $Z_{\text{min}}$ and $Z_{\text{max}}$ are the minimum and maximum values.

The k-fold cross-validation technique was used to select a model from a 70% training data set and a 30% testing data set. This technique randomly divides the initial data into k equal-sized subsets (k-folds). A single subset of the k partitions is designated as testing data to evaluate the model’s ability, and the remaining k-1 subsets are used to train the network. This procedure is repeated k times, and the k fold results can then be averaged to produce a single estimate. In this study, the k value was taken as 10. The specific information about this procedure can be obtained from Cemek et al. (2020).

### 2.5 Performance criteria of model

The coefficient of determination ($R^2$), mean absolute error (MAE) and root mean square error (RMSE) were used to evaluate the performance of models (Waller 2003).

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (Z_{\text{mea}} - Z_{\text{pre}})^2}{\sum_{i=1}^{n} (Z_{\text{mea}} - Z_{\text{avg}})^2}$$ \hspace{1cm} (9)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (Z_{\text{pre}} - Z_{\text{mea}})^2}{n}}$$ \hspace{1cm} (10)

$$MAE = \frac{\sum_{i=1}^{n} |Z_{\text{pre}} - Z_{\text{mea}}|}{n}$$ \hspace{1cm} (11)

where $Z_{\text{mea}}$ is the measured value; $Z_{\text{pre}}$ is the predicted value; $Z_{\text{avg}}$ is the average value measured; and $n$ is the number of data.

In addition, Taylor diagrams were used to analyze the standard deviation (SD) and correlation coefficients (R) between the model-predicted and observed values.

### 3 Results and discussion

Extensive groundwater extraction causes salinization of groundwater and soils in the areas where seawater intrusion occurs, potentially resulting in significant reductions in plant yields (Alfarrah and Walraevens 2018). Generally, hydrochemical properties alone are insufficient to determine seawater intrusion. Thus, combining hydrochemical properties and isotopes can provide additional information about seawater intrusion into the aquifer (Pennisi et al. 2006; Mahlknecht et al. 2017).

The isotopes of $\delta^{18}$O and $\delta$D are two of the most critical isotopes for determining the source of water (Maurya et al. 2019). In this study, four different ANN models—MLP, GRNN, RBNN, and ANFIS—and MLR were used to estimate the $\delta$D and $\delta^{18}$O isotopes in groundwater Bafra plain. The prediction performance of AI models varies depending on the input parameters (Küçüktopcu and Cemek 2021). In AI studies, appropriate variable selection is accomplished with multivariate statistical analysis methods (Lu and Ma 2020). For example, Shah et al. (Shah et al. 2021) used principal component analysis, a type of multivariate statistical analysis, to select the most influential input parameters for estimating the concentrations of dissolved oxygen and total dissolved solids in waters.

In this study, Hierarchical cluster analysis was applied to reduce the number of input variables before developing the ANN and MLR models, and similarities between the variables were determined. Three groups were generated from the hierarchical cluster analysis (Fig. 3). Group A includes EC, Cl, Na, Mg, SO$_4$, $\delta$D, and $\delta^{18}$O. The variables of K, Ca, and pH composed group B, and HCO$_3$ was constituted group C.

Increased EC and Cl values in coastal groundwater may indicate seawater intrusion (Arslan et al. 2012). Cluster analysis revealed that EC and Cl variables were in the same group as $\delta$D and $\delta^{18}$O isotopes, meaning that a seawater intrusion occurred, and these variables could be successfully used to estimate $\delta$D and $\delta^{18}$O isotopes. The
evaluation indicated a similarity between the isotopes ($\delta D$ and $\delta^{18}O$) and the variables of EC, Cl, Na, Mg, and SO$_4$. Therefore, $\delta D$ and $\delta^{18}O$ isotopes were estimated using these variables in the same group as isotopes.

Different input combinations were tested to assess the degree of effect of each variable on $\delta D$ and $\delta^{18}O$ values. The input combinations tested were (i) EC; (ii) EC and Cl; (iii) EC and SO$_4$; (iv) EC, Cl and SO$_4$; (v) EC, Cl, SO$_4$, and Mg.

In the MLP models, the different numbers of hidden nodes were examined, and the optimal number was generated. The lowest RMSE value in the testing phase was selected. In the ANFIS technique, the most appropriate outputs were tested for Gaussian, trapezoidal, and triangle membership functions (MFs) with varying numbers of MFs. In the RBNN model, the optimum spread and hidden node numbers were determined using an approach to trial and error. The GRNN application found optimal models by a trial-and-error method using different spreads. In the SVM technique, optimal parameters are selected using the rule and the stopping criteria. Several studies have been used machine learning models to predict water quality parameters (Ahmed et al. 2019; Bui et al. 2020; Lu and Ma 2020).

In the MLR analysis, the $\delta D$ and $\delta^{18}O$ were used as dependent variables, and EC, Cl, SO$_4$, and Mg were assumed to be independent variables for the training data set. After deriving the regression equations, the values of $\delta D$ and $\delta^{18}O$ for the testing data set were determined.

The training and testing results for the MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR models in the estimation of $\delta^{18}O$ values are given in Table 2. In the training stage, the RMSE values for MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR ranged between 0.19–0.47, 0.34–0.47, 0.34–0.47, 0.50–0.72, 0.30–0.74, 0.21–0.35, 0.28–0.35, and 0.55–0.70%, respectively. The lowest RMSE value (0.19%) was obtained for the MLP5 (4,5,1) model in which the inputs were EC, Cl, SO$_4$, and Mg; however, the highest value (0.72%) was obtained for the SVM2 (Gaussian) model in which the inputs were EC and SO$_4$. Similarly, the lowest value of MAE (0.10%) was calculated for the MLP5 (4,5,1) model, whereas the highest value (0.56%) was for the MLR1 model. The $R^2$ values between the measured and estimated $\delta^{18}O$ were between 0.91 and 0.98 for MLP models and between 0.94 and 0.98 for GPR models, while the $R^2$ value varied between 0.77 and 0.97 for the other models.

In the testing stage, the RMSE values for MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART and MLR ranged between 0.31–0.54, 0.35–0.70, 0.32–0.68, 0.44–0.74, 0.60–0.85, 0.33–0.44, 0.39–0.41 and 0.48–0.70%, respectively. The lowest MAE (0.20%) was attained for the MLP5 (4,5,1) model, whereas the highest value (0.56%) was for the MLR1 model. For the MLP, the $R^2$ values of ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR models were between 0.93–0.98, 0.86–0.97, 0.87–0.97, 0.83–0.95, 0.85–0.94, 0.95–0.98, 0.96–0.97 and 0.86–0.93%, respectively.

The results of training and testing for the MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR models in the estimation of $\delta D$ values are given in Table 3. In the training stage, the RMSE values for MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR ranged between 2.05–2.64, 2.35–5.23, 3.46–9.03, 4.14–5.59, 2.36–5.96, 2.63–3.19, 2.97–3.24, and 4.56–5.69%, respectively. Like the RMSE criterion, the minimum MAE value (1.36%) was determined using the SVM5 (cubic) model. The maximum MAE value (4.23%) was found for RBNN3 (2,0.6310), of which the input variables were spread parameter values. The hidden nodes were 2, 0.6, and 10, respectively. Similarly, the $R^2$ between the measured and estimated $\delta D$ values for the MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR models ranged between 0.94–0.96, 0.86–0.95, 0.84–0.89, 0.71–0.85, 0.68–0.95, 0.90–0.94, 0.90–0.91, and 0.70–0.81, respectively.

In the testing stage, the lowest and highest RMSE values for MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR were 0.19% and 0.72%, respectively.
MLR models were 2.77–2.97, 3.02–4.76, 3.92–4.89, 4.04–5.29, 4.14–4.99, 3.17–3.57, 3.50–3.96, and 4.24–4.75\% respectively. The lowest MAE value (1.89\%) was obtained for the MLP5(4,5,1) model, while the highest MAE value (4.03\%) was found for the RBNN2(2,0.5,10) model. The $R^2$ values ranged from 0.94 to 0.95 for MLP models and from 0.80 to 0.95 for all other models. Ghobadi et al. (2021) obtained a similar result when examining

| Inputs       | Model               | Training | Testing |
|--------------|---------------------|----------|----------|
|              |                     | MAE (%)  | RMSE (%) | $R^2$    | MAE (%)  | RMSE (%) | $R^2$    |
| EC           | MLP1(1,5,1)         | 0.36     | 0.47     | 0.91     | 0.46     | 0.54     | 0.93     |
| EC and Cl    | MLP2(2,5,1)         | 0.24     | 0.34     | 0.95     | 0.37     | 0.45     | 0.94     |
| EC and SO4   | MLP3(2,5,1)         | 0.25     | 0.31     | 0.96     | 0.25     | 0.35     | 0.97     |
| EC, Cl and SO4 | MLP4(3,5,1)     | 0.17     | 0.23     | 0.98     | 0.25     | 0.32     | 0.97     |
| EC, Cl, SO4, and Mg | MLP5(4,5,1) | 0.10     | 0.19     | 0.98     | 0.20     | 0.31     | 0.98     |
| EC           | ANFIS1(3,trimf)     | 0.36     | 0.47     | 0.91     | 0.46     | 0.54     | 0.93     |
| EC and Cl    | ANFIS2(5,trimf)     | 0.39     | 0.48     | 0.90     | 0.31     | 0.39     | 0.95     |
| EC and SO4   | ANFIS3(4,trimf)     | 0.27     | 0.39     | 0.94     | 0.28     | 0.35     | 0.97     |
| EC, Cl and SO4 | ANFIS4(3,trimf)   | 0.30     | 0.40     | 0.93     | 0.56     | 0.70     | 0.86     |
| EC, Cl, SO4, and Mg | ANFIS5(3,trimf) | 0.25     | 0.34     | 0.95     | 0.35     | 0.45     | 0.95     |
| EC           | RBNN1(1,0,1,8)      | 0.24     | 0.34     | 0.95     | 0.46     | 0.66     | 0.87     |
| EC and Cl    | RBNN2(2,0,5,10)     | 0.38     | 0.45     | 0.91     | 0.28     | 0.37     | 0.96     |
| EC and SO4   | RBNN3(2,0,1,10)     | 0.36     | 0.47     | 0.90     | 0.26     | 0.32     | 0.97     |
| EC, Cl and SO4 | RBNN4(3,0,1,10)   | 0.33     | 0.41     | 0.93     | 0.31     | 0.41     | 0.95     |
| EC, Cl, SO4, and Mg | RBNN5(4,0,1,10) | 0.34     | 0.43     | 0.92     | 0.50     | 0.68     | 0.87     |
| EC           | GRNN1(1,0,1,1)      | 0.43     | 0.59     | 0.85     | 0.39     | 0.56     | 0.91     |
| EC and Cl    | GRNN2(2,0,1)        | 0.52     | 0.72     | 0.78     | 0.54     | 0.74     | 0.83     |
| EC and SO4   | GRNN3(2,0,1)        | 0.42     | 0.55     | 0.87     | 0.36     | 0.51     | 0.93     |
| EC, Cl and SO4 | GRNN4(3,0,1)      | 0.41     | 0.54     | 0.88     | 0.35     | 0.50     | 0.93     |
| EC, Cl, SO4, and Mg | GRNN5(4,0,1) | 0.38     | 0.50     | 0.90     | 0.32     | 0.44     | 0.95     |
| EC           | SVM1(Linear)        | 0.55     | 0.72     | 0.76     | 0.59     | 0.75     | 0.87     |
| EC and Cl    | SVM2(Gaussian)      | 0.47     | 0.74     | 0.81     | 0.64     | 0.85     | 0.90     |
| EC and SO4   | SVM3(Cubic)         | 0.50     | 0.72     | 0.85     | 0.62     | 0.84     | 0.88     |
| EC, Cl and SO4 | SVM4(Linear)      | 0.44     | 0.56     | 0.85     | 0.53     | 0.78     | 0.85     |
| EC, Cl, SO4, and Mg | SVM5(Cubic)   | 0.20     | 0.30     | 0.97     | 0.37     | 0.60     | 0.94     |
| EC           | GPR1(Exponential)   | 0.27     | 0.35     | 0.94     | 0.29     | 0.42     | 0.96     |
| EC and Cl    | GPR2(matern 5/2)    | 0.26     | 0.35     | 0.94     | 0.29     | 0.42     | 0.96     |
| EC and SO4   | GPR3(Exponential)   | 0.22     | 0.27     | 0.97     | 0.28     | 0.44     | 0.95     |
| EC, Cl and SO4 | GPR4(Exponential)  | 0.21     | 0.26     | 0.97     | 0.26     | 0.42     | 0.96     |
| EC, Cl, SO4, and Mg | GPR5(Exponential) | 0.16     | 0.21     | 0.98     | 0.23     | 0.33     | 0.98     |
| EC           | CART1               | 0.27     | 0.35     | 0.94     | 0.30     | 0.41     | 0.96     |
| EC and Cl    | CART2               | 0.26     | 0.31     | 0.95     | 0.30     | 0.40     | 0.97     |
| EC and SO4   | CART3               | 0.27     | 0.35     | 0.94     | 0.30     | 0.41     | 0.96     |
| EC, Cl and SO4 | CART4              | 0.22     | 0.28     | 0.96     | 0.30     | 0.39     | 0.97     |
| EC, Cl, SO4, and Mg | CART5              | 0.23     | 0.28     | 0.96     | 0.30     | 0.40     | 0.96     |
| EC           | MLR1                | 0.56     | 0.70     | 0.77     | 0.52     | 0.70     | 0.86     |
| EC and Cl    | MLR2                | 0.46     | 0.55     | 0.86     | 0.38     | 0.52     | 0.91     |
| EC and SO4   | MLR3                | 0.51     | 0.62     | 0.82     | 0.42     | 0.54     | 0.91     |
| EC, Cl and SO4 | MLR4              | 0.50     | 0.61     | 0.83     | 0.40     | 0.52     | 0.91     |
| EC, Cl, SO4, and Mg | MLR5              | 0.48     | 0.59     | 0.84     | 0.36     | 0.48     | 0.93     |
MLP, GRNN, and MLR models to forecast groundwater quality parameters (dissolved oxygen, total dissolved solids, and turbidity) in Iran. They concluded that the MLP model was the best of the models studied.

The $R^2$, MAE and RMSE values of the best models for different input combinations in the testing dataset were shown in Fig. 4, which facilitates the graphical comparison of the tested models. The models with input combination...
(v) generally have the highest $R^2$ and the lowest MAE and RMSE values compared to the other input combinations. In other words, the models with EC, Cl, SO$_4$ and Mg inputs yielded the most reliable results compared to the other input combinations tested. Therefore, these data were incorporated into the estimation of $\delta^{18}$O and $\delta$D values.

The measured and estimated $\delta^{18}$O and $\delta$D values by the MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART and MLR models were shown in Figs. 5 and 6. The MLP5(4,5,1) models seem to have better results than the other models tested for $\delta^{18}$O and $\delta$D estimation. The scatter plots of the MLP5(4,5,1) models for the measured and estimated $\delta^{18}$O and $\delta$D values in the testing stage is shown in Fig. 7.

The $\delta^{18}$O and $\delta$D estimation models were evaluated by using a Taylor diagram (Fig. 8). The MLP5 (4,5,1) models yielded a lower SD and RMSE values and a higher correlation coefficient compared to the other studied models.
Therefore, a comparison for the findings of models shows that the MLP5(4,5,1) models provided the most accurate results in the prediction of \( \delta^{18}O \) and \( \delta^D \) values.

Studies on the estimation of isotopes using artificial intelligence models are rare. However, different researchers have estimated the hydrochemical properties of waters with AI models (El Bilali et al. 2021; Singha et al. 2021; Egbueri and Agbasi 2022). The researchers stated that different AI models could make the most accurate predictions in evaluating different water properties.

Fig. 5 The measured and estimated \( \delta^{18}O \) values by the optimal models for MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR.
Haghiabi et al. (Haghiabi et al. 2018) compared the ANN, SVM, and group method of data handling models to determine the pH, SO₄, Na, Cl, Ca, Mg, HCO₃ values of the waters and obtained the most accurate results with the SVM model. Kouadri et al. (Kouadri et al. 2021) predicted the irrigation water quality indices using different AI models, including ANN, MLR, and long short-term memory models.

4 Conclusions

Excessive groundwater extraction in coastal areas degrades and pollutes water sources. When these waters are used for irrigation, they significantly reduce plant yields and contribute to soil salinization. Thus, continuous monitoring of water’s hydrochemical and isotope composition is required in coastal areas with excessive groundwater withdrawal, particularly in paddy cultivation areas.
Hydrochemical properties of waters are generally used in seawater intrusion studies due to the low cost and ease of obtaining. However, environmental conditions can affect the hydrochemical properties very quickly; therefore, the isotope analysis provides significant benefits in seawater intrusion studies.

This study employed different data-driven models called MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR to estimate the $\delta^{18}$O and $\delta^D$ values using the known hydrogeochemical properties of groundwaters in the Bafra plain of Turkey. The estimated values obtained with the models were statistically compared using $R^2$, RMSE, and MAE. Taylor diagram was also employed to assess the performance of the investigated models. Comparative analysis of the results obtained by the models indicated that MLP5 (4,5,1) generated the most accurate models for all estimations based on $R^2$, RMSE, and MAE values for the estimation of $\delta^{18}$O (0.98, 0.31 and 0.20%, respectively) and $\delta^D$ (0.95, 2.85 and 1.89% respectively) values in the testing datasets. In the case of lack of data availability, the MLP1 (1,5,1) model, which only used the EC values generated satisfactory results in estimating the $\delta^D$ and $\delta^{18}$O values. Overall, the study suggests using data-driven methods, especially MLP, when lacking the appropriate isotope analysis and facing high costs.

**Funding** The authors have not disclosed any funding.

**Declarations**

**Conflict of interest** The authors declare that they have no conflict of interest.
References

Achirul Nanda M, Boro Seminar K, Nandika D et al (2018) A comparison study of kernel functions in the support vector machine and its application for termite detection. Information 9:5

Ahmed AN, Othman FB, Afan HA et al (2019) Machine learning methods for better water quality prediction. J Hydrol 578:124084

Ahsan M, Eskhabilov S, Cemek B et al (2022) Deep learning models to determine nutrient concentration in hydroponically grown lettuce cultivars (Lactuca sativa L.). Sustainability 14(1):416

Alfarrah N, Walraevens K (2018) Groundwater overexploitation and seawater intrusion in coastal areas of arid and semi-arid regions. Water 10:143

Arslan H, Cemek B, Demir Y (2012) Determination of seawater intrusion via hydrochemicals and isotopes in Bafra Plain, Turkey. Water Resour Manag 26:3907–3922

Bahir M, Ouhamdouch S, Carreira PM (2018) Geochemical and isotopic approach to decrypt the groundwater salinization origin of coastal aquifers from semi-arid areas (Essaouira basin, Western Morocco). Environ Earth Sci 77:485

Barzegar R, Moghaddam AA (2016) Combining the advantages of neural networks using the concept of committee machine in the groundwater salinity prediction. Model Earth Syst Environ 2:26

Bray M, Han D (2004) Identification of support vector machines for runoff modelling. J Hydronformics 6:265–280

Breiman L, Friedman J, Olshen R, et al. (1984) Classification and regression trees, complex systems, vol. 2

Buı DT, Khosravi K, Tiefenbacher J et al (2020) Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Sci Total Environ 721:137612

Cahyadi TA, Syihab Z, Widodo LE et al (2020) Analysis of hydraulic conductivity of fractured groundwater flow media using artificial neural network back propagation. Neural Comput Appl. https://doi.org/10.1007/s00521-020-04970-z

Cemek B, Güler M, Kilic K et al (2007) Assessment of spatial variability in some soil properties as related to soil salinity and alkalinity in Bafra plain in northern, Turkey. Environ Monit Assess 124:223–234

Cemek B, Ünlükara A, Kuruç A et al (2020) Leaf area modeling of bell pepper (Capsicum annuum L.) grown under different stress conditions by soft computing approaches. Comput Electron Agric 174:105514

Chipman HA, George EI, McCulloch RE (1998) Bayesian CART model search. J Am Stat Assoc 93:935–948

Choubin B, Rahmati O, Soleimani F et al (2019) Regional groundwater potential analysis using classification and regression trees. Spatial modeling in GIS and R for earth and environmental sciences. Elsevier, Armsterdam, pp 485–498

Cigizoglu HK, Kişi Ö (2005) Flow prediction by three back propagation techniques using k-fold partitioning of neural network training data. Hydrol Res 36:49–64

Daneshmand H, Tavousi T, Khosravi M et al (2015) Modeling minimum temperature using adaptive neuro-fuzzy inference system based on spectral analysis of climate indices: a case study in Iran. J Saudi Soc Agric Sci 14:33–40

Ding Z, Koriem MA, Ibrahim SM et al (2020) Seawater intrusion impacts on groundwater and soil quality in the northern part of the Nile Delta, Egypt. Environ Earth Sci 79:1–11

Egbueri JC (2020) Groundwater quality assessment using pollution index of groundwater (PIG), ecological risk index (ERI) and hierarchical cluster analysis (HCA): a case study. Groundw Sustain Dev 10:100292

Egbueri JC, Aghasi JC (2022) Data-driven soft computing modeling of groundwater quality parameters in southeast Nigeria: comparing the performances of different algorithms. Environ Sci Pollut Res. https://doi.org/10.1016/S0960-0779(23)00068-7

El Bilali A, Moukhllis M, Taleb A, et al. (2021) Predicting daily pore pressure in embankment dam using machine learning models and hydrostatic seasonal time approach

El-Bakry O (2003) Feed forward neural networks modeling for K-P nutrient concentration in hydroponically grown lettuce cultivars (Lactuca sativa L.). Sustainability 14(1):416

Genc O, Gonen B, Ardıçgil MA (2014) An evaluation of shear stress modeling based on machine learning methods in small streams. J Hydroinform 17:805–816

Ghobadi A, Cheraghi M, Sobhanardakani S, et al. (2021) Groundwater quality modeling using a novel hybrid data-intelligence model based on gray wolf optimization algorithm and multi-layer perceptron artificial neural network: a case study in Asadabad Plain, Hamedan, Iran. Environ Sci Pollut Res, 1–15

Haghiabi AH, Nasrolahi AH, Parsaie A (2018) Water quality prediction using machine learning methods. Water Qual Res J 53:3–13

Hameed M, Shafiee SS, Yaseen ZM et al (2017) Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia. Neural Comput Appl 28:893–905

Haykin S (2001) Neural networks: a comprehensive foundation.

Tsinghua University Press, Beijing

Isawi H, El-Sayed MH, Essa M et al (2016) Integrated geochemistry, isotopes, and geostatistical techniques to investigate groundwater sources and salinization origin in the Sharm El-Shiekh Area, South Sinai, Egypt. Water, Air, and Soil Poll 227:151

Jafari R, Torabian A, Ghorbani MA et al (2019) Prediction of groundwater quality parameter in the Tabriz plain, Iran using soft computing methods. J Water Supply Res Tech AQUA 68:573–584

Jahneke C, Wannous M, Troeger U et al (2019) Impact of seawater intrusion and disposal of desalinization brines on groundwater quality in El Gouna, Egypt, Red Sea Area. Process analyses by means of chemical and isotopic signatures. Appl Geochem 100:64–76

Jayathunga K, Diyabanange S, Frank AH et al (2020) Influences of seawater intrusion and anthropogenic activities on shallow coastal aquifers in Sri Lanka: evidence from hydrogeochemical and stable isotope data. Environ Sci Pollut Res. https://doi.org/10.1007/s11356-020-08759-4

Juntakut P, Snow DD, Haacker EM et al (2019) The long term effect of agricultural, vadose zone and climatic factors on nitrate contamination in Nebraska’s groundwater system. J Contam Hydro 220:33–48

Kazakis N, Mattas C, Pavlou A et al (2017) Multivariate statistical analysis for the assessment of groundwater quality under different hydrogeochemical regimes. Environ Earth Sci 76:349

Klassen J, Allen D (2017) Assessing the risk of saltwater intrusion in coastal aquifers. J Hydrol 551:730–745

Knoll L, Breuer L, Bach M (2019) Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. Sci Total Environ 668:1317–1327

Kouadri S, Pande CB, Panneerselvam B et al (2021) Prediction of irrigation groundwater quality parameters using ANN, LSTM, and MLR models. Environ Sci Pollut Res 28(14):21067–21091

Küçüksözcü E, Cemek B (2021) Comparison of neuro-fuzzy and neural networks techniques for estimating ammonia concentration in poultry farms. J Environ Chem Eng 9:105699

Lal A, Datta B (2018) Genetic programming and gaussian process regression models for groundwater salinity prediction: machine learning methods for better water quality prediction. J Hydrol 578:124084

Ahmed AN, Othman FB, Afan HA et al (2019) Machine learning methods for better water quality prediction. J Hydrol 578:124084
learning for sustainable water resources management. Pages 1–7 in Proc. 2018 IEEE conference on technologies for sustainability (SusTech). IEEE

Libutti A, Cammerino ARB, Monteleone M (2018) Risk assessment of soil salinization due to tomato cultivation in Mediterranean climate conditions. Water 10:1503

Lin I-T, Wang C-H, Lin S et al (2011) Groundwater–seawater interactions off the coast of southern Taiwan: evidence from environmental isotopes. J Asian Earth Sci 41:250–262

Lu H, Ma X (2020) Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere 249:126169

Mahlknecht J, Merchán D, Rosner M et al (2017) Assessing seawater intrusion in an arid coastal aquifer under high anthropogenic influence using major constituents, Sr and B isotopes in groundwater. Sci Total Environ 587:282–295

Mai-Duy N, Tran-Cong T (2003) Approximation of function and its derivatives using radial basis function networks. Appl Math Model 27:197–220

Maurya P, Kumari R, Mukherjee S (2019) Hydrochemistry in integration with stable isotopes (δ18O and δD) to assess seawater intrusion in coastal aquifers of Kachchh district, Gujarat, India. J Geochem Explor 196:42–56

Mohanty AK, Rao VG (2019) Hydrogeochemical, seawater intrusion and oxygen isotope studies on a coastal region in the Puri District of Odisha, India. CATENA 172:558–571

Mongelli G, Monni S, Oggiorno G et al (2013) Tracing groundwater salinization processes in coastal aquifers: a hydrogeochemical and isotopic approach in Na-Cl brackish waters of north-western Sardinia, Italy. Hydrod Earth Syst Sci 17(7):2917–2928

Naghibi SA, Pourghasemi HR, Dixon B (2016) GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. Environ Monit Assess 188:44

Noori N, Kalin L, Isik S (2020) Water quality prediction using SWAT-ANN coupled approach. J Hydrod 590:125220

Odabas MS, Simsek H, Lee CW et al (2017) Multilayer perceptron neural network approach to estimate chlorophyll concentration index of lettuce (Lactuca sativa L.). Commun Soil Sci Plant Anal 48:162–169

Pennisi M, Bianchini G, Muti A et al (2006) Behaviour of boron and strontium isotopes in groundwater–aquifer interactions in the Cornia Plain (Tuscany, Italy). Appl Geochem 21:1169–1183

Rao NS, Chaudhary M (2019) Hydrogeochemical processes regulating the spatial distribution of groundwater contamination, using pollution index of groundwater (PIG) and hierarchical cluster analysis (HCA): a case study. Groundw Sustain Dev 9:100238

Rasmussen CE (2003) Gaussian processes in machine learning. Proc. summer school on machine learning. Springer, Berlin, pp 63–71

Shah MI, Javed MF, Aqhahtani A et al (2021) Environmental assessment based surface water quality prediction using hyper-parameter optimized machine learning models based on consistent big data. Process Saf Environ Prot 151:324–340

Singha S, Pasupuleti S, Singha SS et al (2021) Prediction of groundwater quality using efficient machine learning technique. Chemosphere 276:130265

Skansi S (2018) Introduction to deep learning: from logical calculus to artificial intelligence. Springer, Cham, Switzerland

Specht DF (1991) A general regression neural network. In: Proc. Advances in neural information processing systems, pp 514–520

Vapnik V (1995) The nature of statistical learning theory. Springer, New York, USA

Vapnik V, Golowich SE, Smola AJ (1997) Support vector method for function approximation, regression estimation and signal processing. In: Proc. Advances in neural information processing systems, pp 281–287

Waller DL (2003) Operations management: a supply chain approach. Cengage Learning Business Press, Boston

Williams CK, Rasmussen CE (1996) Gaussian processes for regression. In: Proc. Advances in neural information processing systems, pp. 514–520

Yang J, Ye M, Tang Z et al (2020) Using cluster analysis for understanding spatial and temporal patterns and controlling factors of groundwater geochemistry in a regional aquifer. J Hydrod 583:124594

Zhao T (2008) RBFN-based decentralized adaptive control of a class of large-scale non-affine nonlinear systems. Neural Comput Appl 17:357–364

Zhao Y, Li Y, Zhang L et al (2016) Groundwater level prediction of landslide based on classification and regression tree. Geod Geodyn 7:348–355

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.