Applied Geochemistry 132 (2021) 105054

Contents lists available at ScienceDirect

Applied Geochemistry

journal homepage: www.elsevier.com/locate/apgeochem

Prediction on the fluoride contamination in groundwater at the Datong Basin, Northern China: Comparison of random forest, logistic regression and artificial neural network

Mouigni Baraka Nafouanti[a], Junxia Li[a,b,*], Nasiru Abba Mustapha[a,c], Placide Uwamungu[d], Dalal AL-Alimi[e]

[a] State Key Laboratory of Biogeology and Environmental Geology, China University of Geosciences, Wuhan, 430074, China
[b] China Laboratory of Basin Hydrology and Wetland Eco-restoration, China University of Geosciences, Wuhan, 430074 China
[c] Department of Environmental Sciences, Federal University Dutse, Jigawa State, Nigeria
[d] State Key Laboratory of Geological Processes and Mineral Resources China University of Geosciences, Wuhan, 430074, China
[e] School of Computer Science, China University of Geosciences, Wuhan, 430074, China

ARTICLE INFO

Editorial handling by Prof. M. Kersten

Keywords:
Groundwater
Fluoride
Random forest
Artificial neural network
Logistic regression

ABSTRACT

Groundwater fluoride is posing a health risk to humans, and analyzing groundwater quality is time-wasting and expensive. Statistical methods provide a valuable approach to study the spatial distribution of groundwater fluoride. Random Forest (RF), Artificial Neural Network (ANN), and Logistic Regression (LR) were used in this study for groundwater fluoride prediction in Datong Basin. The groundwater chemistry of 482 groundwater samples was collected and used to figure out the performance of three statistical technologies and extract the main factors controlling the enrichment of fluoride in groundwater. The data was separated into two parts for the statistical analysis, 80% for training and 20% for testing. The Chi-squared was applied to select the most relevant variables, and TDS, Cl^-, NO_3^-, Na^+, HCO_3^-, SO_4^{2-}, K^+, Zn, Ca^{2+}, and Mg^{2+} were selected as best inputs for the fluoride prediction. Models were evaluated using the confusion matrix and The receiver operating characteristic area under the curve ROC (AUC). The results suggest that within ten input variables, the accuracies of RF, ANN, and LR were 0.89, 0.85, and 0.76, respectively. The mean decrease in impurity (MDI) and permutation feature demonstrates that eight of ten parameters, including TDS, Cl^-, NO_3^-, Na^+, HCO_3^-, SO_4^{2-}, Ca^{2+} and Mg^{2+} are the variables influencing the groundwater fluoride in the study area. RF exhibited the best model with high conformity and confidence in predicting groundwater fluoride contamination in the study area.

1. Introduction

Groundwater is an essential source of water supply in numerous developed and underdeveloped countries such as Germany, the United States of America, Bangladesh, and Benin (West Africa) (Houéménou et al., 2020; Khosravi et al., 2020; Sutradhar and Mondal, 2021). It is a principal resource in arid areas where precipitation and surface water are restricted and helps in the development of economic growth (Li et al., 2017). In northern China, groundwater is the main water source for domestic, agriculture, and industrial purposes (Su et al., 2013). However, there is an increasing threat to groundwater due to the presence of several chemical elements like fluoride (Su et al., 2013). Therefore, understanding the groundwater quality is necessary for adequate water management sustainability purposes.

According to the world health organization (WHO), the minimum amount of fluoride concentration in groundwater is between 0.5 and 1.0 mg/L, and the maximum range is 1.5 mg/L, while the Chinese standard is 1 mg/L (Su et al., 2013). Fluoride is a necessary element, and it is required in small amounts to maintain the development of tooth enamel and the health of bones in humans (Rafique et al., 2008; Tripathy et al., 2006). However, long-term excessive fluoride consumption causes numerous human health problems, including dental fluorosis, skeletal fluorosis, gastrointestinal disorders, and immune system disorders, which have been widely reported in several countries such as India, Korea, Pakistan, China, Mexico, and many countries in Africa (Apambire et al., 1997; Ayenew, 2008; Kim et al., 2011; Naseem et al., 2010;

[*] Corresponding author. State Key Laboratory of Biogeology and Environmental Geology, China University of Geosciences, Wuhan, 430074, China.
E-mail address: jxli@cug.edu.cn (J. Li).

https://doi.org/10.1016/j.apgeochem.2021.105054
Received 18 April 2021; Received in revised form 20 July 2021; Accepted 22 July 2021
Available online 24 July 2021

Rafique et al., 2009).

The fluoride concentration in the groundwater systems has been reported to be influenced by conditions resulting from the natural hydrogeochemical processes, such as dissolution of fluoride-containing minerals, including fluorite and biotite, precipitation of carbonate minerals, Ca–Na exchange on the clay minerals, and intense evapotranspiration (Li et al., 2020). Additionally, recent studies showed that anthropogenic activities using fertilizers and irrigation processes directly affect the groundwater fluoride concentration (Ayenew, 2008; Cairncross and Feachem, 1993). Altogether, these factors leave us with an unknown regarding the fate of fluoride in the groundwater. Therefore, the assessment of groundwater quality, monitoring, and modeling are necessary for identifying groundwater trends and groundwater sustainability.

Numerical models have been previously applied for groundwater quality modeling purposes (Rapantova et al., 2007). However, these models have limitations, such as needing a large quantity of data, considerable time, and have a complex structure that restricts their use (Alagha et al., 2014; Coppola et al., 2005). Thus to solve this issue, it is necessary to adopt a potential approach for assessing groundwater contamination.

The application of machine learning models can provide an efficient alternative in predicting groundwater contamination which have been widely used by many studies (e.g., Mohammadi et al., 2016a; Nadiri et al., 2013; Noshad et al., 2019). For example, the artificial neural network (ANN) can detect complex non-relationship between predictor variables and the dependent variable and has the ability to solve erroneous and voluminous problems in a dataset (Mohammadi et al., 2016a; Tarasov et al., 2018). The random forest (RF) model can handle high-dimensional data, continuous, missing values, and binary data. Furthermore, logistic regression (LR) is an efficient algorithm that is fast in dataset training and efficiently used to analyze binary classification (Stoltzfus, 2011). Many studies employed ANN, RF, and LR to predict groundwater contamination. For instance, ANN was applied to predict fluoride contamination in groundwater in Khaf (Mohammadi et al., 2016a). Likewise, it was used to forecast groundwater contaminated by nitrate in Iran and predict the concentration of high fluoride groundwater in the Maku area (Nadiri et al., 2019; Ostad-Ali-Askari et al., 2017). Similarly, RF was employed to predict groundwater contamination by uranium in California and predict groundwater pollution by nitrate in Southern Spain (Lopez et al., 2020; Rodriguez-Galiano et al., 2014). In addition, some existing studies used LR for predicting groundwater contamination. For example, it was employed in India to predict groundwater contamination by Fluoride (Podgorski et al., 2018). However, these algorithms were all individually applied to predict groundwater contamination, and there is a gap in identifying the best machine learning to effectively predict groundwater contamination. In this regard, the current study compares three machine learnings, RF, ANN, and LR, to predict the fluoride in groundwater using binary classification analysis.

The objective of the present work is to identify the most suitable predictive model that can be applied to predict fluoride contamination in groundwater in the Datong Basin. Therefore, an evaluation and comparison of three models, RF, ANN, and LR classifiers, were applied using physicochemical water parameters from the study area. Also, the determination of the variables influencing the fluoride in the study area was considered in this study. This investigation will provide insights into using classification models to predict groundwater and enhance groundwater prediction in the study area and elsewhere in the world.

2. Hydrogeological setting

Datong Basin belongs to the Shanxi rift system with around 6000 km^2 formed by Cenozoic faulted basins (Xing et al., 2013). It is situated in East Asia, characterized by a seasonal monsoon region with a semiarid climate. According to topography, the area is enclosed by mountains and slopes from the northwest to the southeast (Fig. 1). The annual precipitation is between 225 mm and 400 mm, and the evapotranspiration is over 2000 mm. The annual average air temperature is 6.5 °C (Su et al., 2015). The Sanggan and the Huangshui rivers are the main rivers running across the study area. They are used for land irrigation due to several agricultural activities developed in the area (Wang and Shpeyzer, 2000).

The outcrops for bedrock are detected in the western, eastern, and northern. The outcrops for the north are basalt and Archean gneiss. The west is constituted by Carboniferous–Permian–Jurassic sandstone, Cambrian–Ordovician limestone, and shale. In the northeast, the basin is formed of granite sparsely and Archean gneiss. The sediment in the basin is alluvial–pluvial sand and gravel. The central part of the basin is formed by Alluvial–pluvial sands, lacustrine and alluvial– lacustrine sandy loam soils. Also, silts and silty clay abundant in organic matter are reported in the central part of the basin (Guo and Wang, 2005).

Furthermore, three aquifers are beneath the flat alluvial–lacustrine plain in the basin center, the upper, middle, and lower aquifers. The upper aquifer is formed by sands, and gravel usually occurs between 5 and 60 m under the land surface with 2–10 m thick. The middle aquifer is molded by sandy gravel and sand from 60 to 160 m beneath the land surface. Finally, the lower aquifer consists of silt and fine sand observed at depths bigger than 160 m beneath the land surface (Xie et al., 2009).

The groundwater recharge is by infiltration of the basins meteoric water vertically, irrigation return flow, and bedrock fractures in the mountain front, accompanied by an outflow from non-perennial rivers laterally (Guo and Wang, 2005). Evaporation and abstraction are the two major causes of the groundwater discharged in the study area.

3. Methodology

3.1. Sampling and analytical methods

The details on the groundwater sampling and chemistry analysis can be obtained from our previous work (Li et al., 2012). Briefly, 482 samples were collected from different wells in August 2011 (Fig. 1). Quality assurance and quality control were maintained in the sampling and all analytical procedures (Li et al, 2012, 2020). All the chemical measurements were accomplished at the State Key Laboratory of Biogeology and Environmental Geology, China University of Geosciences, Wuhan.

3.2. Data preprocessing

The data comprised 16 input variables, including TDS, Cl^-, NO_3^-, Na^+, K^+, HCO_3^-, SO_4^{2-}, Ca^{2+}, Mg^{2+}, pH, Ba, Li, Mn, Pb, Sr, Zn, and the dependent variable. The data was converted into high and low classes by allocating zero (0) to all fluoride concentrations lower than 1 mg/L and assigning by one (1) for the fluoride concentrations higher than 1 mg/L. The independent variables were then scaled between 0 and 1 for the three algorithms to enhance the model speed and accuracy. The data was then randomly divided into two sections 80% for training and 20% for testing.

3.3. Selection of the relevant input

Discarding significant variables or maintaining irrelevant variables affects machine learning model performance (Gheyas and Smith, 2010). For selecting the relevant inputs, filter methods were applied in this study. These methods are rapid compared to the wrapper methods as they do not involve model training. Moreover, they can determine the relationship between the independent and the dependent variables (Hendrawan and Murase, 2011; Sánchez-Marono et al., 2007).

In the filter methods, the Chi-squared was implemented in this study as a feature selection method. The Chi-squared compares the observed distribution between various variables in the dataset and the dependent variable. It summarizes squared differences among observed and ex-

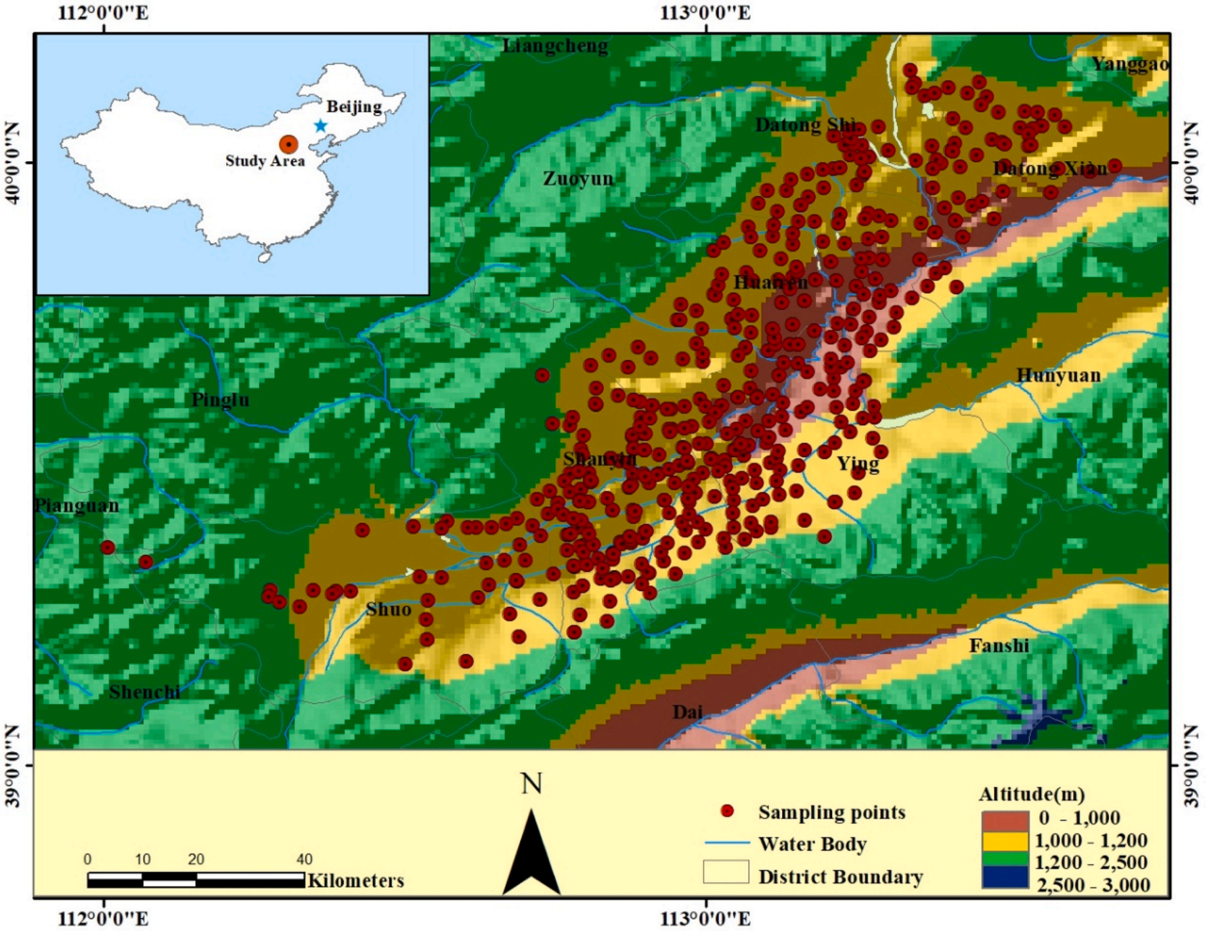

Fig. 1. Location map of the study area showing the sampling location.

pected values divided by expected values to determine the most relevant independent variables in the prediction (Lee et al., 2011). The variables are independent when the observed count is close to the expected count, and these variables will have a small Chi-squared value. Thus, a high Chi-Squared value indicates that the variable is more dependent on the output, and it can be chosen for model training. The variables were selected using the sklearn library in python using the "SelecktBest," which retained the first k (The degree of freedom which is the number of samples being summed) input variables with the highest scores (Table 1). Therefore, ten (10) variables such as TDS, Cl^-, NO_3^-, Na^+, K^+, HCO_3^-, SO_4^{2-}, Ca^{2+}, Mg^{2+}, Zn are reported with a high Chi-squared value and were selected as relevant inputs for groundwater fluoride prediction.

$$\text{The Chi} - \text{squared is defined as: } X_c^2 = \sum \frac{(Oi - Ei)^2}{Ei} \tag{1}$$

C = degree of freedom (Degree of freedom refers to the maximum number of independent values, which have the freedom to vary in the data sample).

O = Observed value(s) (They are the values that are observed in the dataset).

E = Expected value(s) (The expected value is based on the row and column totals. It is the multiplication of row total by the column total and then dividing by the total, and gives the expected value for each cell).

3.4. Random forest modeling

RF is an algorithm that can be used for regression and classification analysis. In this study, random forest classification was used in predicting groundwater fluoride contamination. The RF combined many decision trees to limit overfitting, formulates a robust model, and gives high accuracy. In the random forest, the random is presented in two ways in the trees growing. First, a random selection with the substitute of all data rows results from one-third of the data and "out-of-bag" (OBB), which are not randomly selected for a decision tree. The second is the restricted number of randomly selected variables available at each node. In the RF, the number of trees and the number of predictor variables chosen at each node are the tuning parameters determining the RF overall fit. In this work, one hundred (100) trees were grown to generate the RF model.

In addition, RF can identify significant predictor variables and efficiently describe how they affect contaminant existence in aquifers. In this study, to assess the essential variables, the mean decrease in

Table 1
Selection of Relevant Inputs by using the Chi-Squared Analysis.

Variables	TDS	Na^+	HCO_3^-	NO_3^-	SO_4^{2-}	Cl^-	Ca^{2+}	Mg^{2+}	K^+	Zn
score	20668.6	8967.3	8515.7	5226.4	2131.2	1583.5	1001.9	459.9	59.2	7.3

impurity (MDI) was applied, a measure utilized for relative importance in RF sub-nodes to create splitting on a given variable (Bylander, 2002; Han et al., 2016). In impurity, a split with a considerable decrease is considered essential. Then the higher the mean decrease in impurity, the more important the variable is.

3.5. Neural network

ANN is a model intended to simulate biological 'neurons' behavior (Ostad-Ali-Askari et al., 2017). In this study, the ANN applied is the multilayer perceptron (MLP) feedforward. The MLP is a type of neural network in which each neuron is associated with above-layer neurons (Nevtipilova, 2014). The MLP neural network used in this study was composed of three different layers (Fig. 2). It was composed of input, hidden, and output layers. The input layers were formed of 10 neurons, which are the number of predictor variables. The hidden layer where data is processed was composed of two layers, and an output layer produces the results. Each layer comprises a fundamental element named neuron, which has a threshold and an activation function essential to the training process (Dreyfus, 2008; Mohammadi et al., 2016b). In this study, the "adam" optimizer was used to update the weight in the network. The mathematical expression of the MLP defines as:

$$Xnm = \sum_i WnmXn + Wm \tag{2}$$

Xn represents the output of nodes, *i* located for any of the previous layers, W*nm* the weight associated with the link connecting nodes *n* and *m*, and *Wm* the bias of node *m*.

In this study, the activation function used to the hidden layer is relu and is defined by:

$$f(x) = \max(0, x =)f(x) = \begin{cases} xi \; if \; xi > 0 \\ 0, if \; xi < 0 \end{cases} \tag{3}$$

In the output layer, the activation function depends on the prediction of the model. For this analysis, the sigmoid activation is applied in the output layer, and it defines as

$$f(x) = \frac{1}{1 + e^{x}} \tag{4}$$

Furthermore, the permutation feature was used to determine the essential variables between the predictors and the dependent variables. It demonstrates whether eliminating a variable would affect the network accuracy.

3.6. Logistic regression

Logistic regression (LR) is mostly used for binary classification (Qian et al., 2020). It is a conversion of linear regression using the sigmoid function. In this work, LR is applied to predict the fluoride in groundwater. LR equation describes as:

$$F(x) = \frac{1}{1 + e^{-(\beta 0 + \beta 1x)}} \tag{5}$$

where β_0 and β_1 are the estimated parameters.

3.7. Model evaluation criteria

The models predictive capability in the testing stage was evaluated using the confusion matrix for each model. The accuracy, sensitivity, specificity, and error were calculated to assess the model prediction. The receiver operating characteristic area under the curve ROC (AUC) was also considered to evaluate the LR.

The evaluation of predictive performance for binary classification is mainly based on the confusion matrix. It shows how the model classified the actual values compared to predicted values (Bowes et al., 2012). The prediction was compared to observed concentrations to identify the percentages of observations that were correctly classified. The percentage of fluoride correctly classified is known as the sensitivity, and the non-fluoride that was correctly classified is known as the specificity. The three models were carried out using the Python3.7 programming language.

The metrics equation for the confusion matrix are described as:

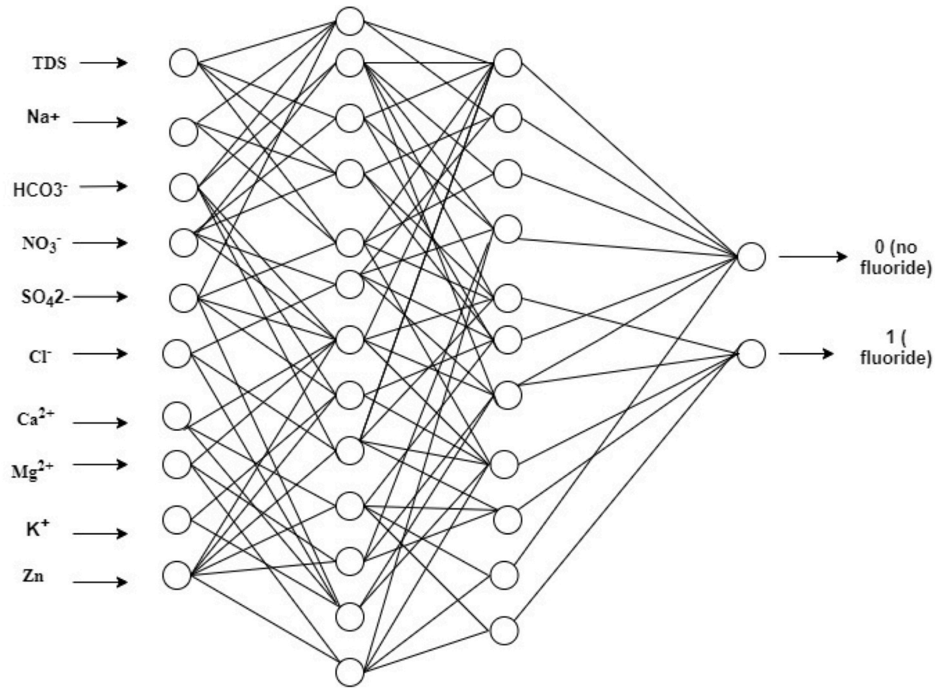

Fig. 2. Structure of Artificial Neural Network with the Inputs variables of the study area.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{7}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{8}$$

$$Error\ rate = \frac{FP + FN}{TP + TN + FP + FN} \tag{9}$$

TP: The true positive is when an item is predicted as fluoride, and it is correct is fluoride.
TN: The true negative is when an item is predicted as non-fluoride, and it is correct is non-fluoride.
FP: False positive when an item is predicted as fluoride, and it is not fluoride (It presents the error in binary classification).
FN: False-negative when an item is predicted as non-fluoride, and it is fluoride (It is the opposite error in binary classification).
Error rate: Number of instances misclassified over the whole set of instances (FP and FN).

4. Results and discussion

4.1. Hydrochemical characteristics

Understanding groundwater hydrochemistry is indispensable for knowing the different chemical compositions related to the various aquifer. The mean fluoride concentration is up to 1.7 mg/L in groundwater and ranges from 0.01 to 66.7 mg/L (Table 2). Thus, 294 out of 482 groundwater samples were detected as high fluoride groundwater in the study area. The total dissolved solids (TDS) ranges from 289.3 to 20588 mg/L. The concentration of HCO_3^- is between 159.2 and 1786 mg/L with an average of 476.2 mg/L. The concentration of Ca^{2+} is varied from 3.2 to 716.6 mg/L with a mean value of 55.7 mg/L. Cl^-, NO_3^-, SO_4^{2-}, K^+, Na^+, Mg^{2+}, and Zn were also identified in groundwater samples (Table 2).

Fluoride concentration in groundwater is mainly controlled by different processes from natural to anthropogenic activities. A previous study proposed that at the Datong Basin, the main geochemical processes influencing the mobilization of groundwater fluoride include the precipitation and dissolution of carbonate, gypsum, halite, silicate weathering, hydrolysis, and evapotranspiration (Su et al., 2015).

4.2. Model evaluation and comparison

Table 3
Statistical metrics for the Models Evaluation using Physico-Chemical Water Parameters for the three algorithms Random Forest, Neural Networks, and Logistic Regression.

Metrics	RF	ANN	LR
Accuracy	0.89	0.85	0.76
Sensitivity	0.98	0.89	0.82
Error rate	0.10	0.14	0.23

After model building and training, the models received test predictors data to evaluate their performance in predicting the fluoride in groundwater. The evaluation metrics for RF, ANN, and LR were extracted from their confusion matrix, and details are described in Table S1, S2, and S3 in the supporting material section. The Metrics employed for evaluating the three models (Table 3) revealed that the accuracy, sensitivity, specificity, and error rate for the RF model were 0.89, 0.98, 0.76, and 0.10, respectively. The high sensitivity above specificity in binary classification demonstrates a less false negative, suggesting a good prediction model. The ability of RF in predicting water blooms contamination and groundwater contamination by nitrate has been previously explored with reported accuracies of 0.87, and 0.77 respectively, which are slightly lower than the accuracy in our study (Liu and Wu, 2018; Tesoriero et al., 2017). The performance accuracy for RF in this study is enhanced by selecting the relevant inputs and applying many trees leading to a good performance model.

For the ANN accuracy, sensitivity, specificity, and error rate were 0.85, 0.89, 0.80, and 0.14, respectively. This result is consistent with the finding of a previous study in predicting water quality using different physicochemical water parameters (Ahmed et al., 2019). Similarly, the performance of the ANN was also demonstrated in a previous study in the prediction of water pollution with an accuracy of 0.80 (Keskin et al., 2015). The ANN performance in this study was improved by the number of hidden layers in the network training. The selection of more than one hidden layer to improve accuracy was also suggested in previous studies (Awan et al., 2018; Uzair and Jamil, 2020). In the ANN, an adequate number for training the network can be achieved with a maximum of two hidden layers.

For the LR, the accuracy, sensitivity, specificity, and error rate were 0.76, 0.82, 0.67, and 0.23, respectively. The LR was further evaluated using the ROC (AUC) to determine the ability of the model (Fig. 3). The result revealed that the LR performed with an AUC of 0.83, slightly higher than the performances reported in previous studies to predict groundwater spring and groundwater fluoride with reported AUC of 0.82 and 0.78, respectively (Ozdemir, 2011; Podgorski et al., 2018).

The statistical evaluations revealed that the three models used in this study to predict groundwater fluoride yielded satisfactory results with high sensitivity and specificity, which means a lower error in predicting groundwater fluoride. The better performance accuracies observed in this study for three models might be attributed to the relevant parameters selected by using the Chi-squared analysis.

Amongst the three algorithms, the results suggested that RF has shown a high performance for predicting fluoride in groundwater in the study area, which outperforms the performance of ANN and LR (Table 3). The dissimilarity in predictive ability can be attributed to variations in the algorithm structure. The high performance of the RF is attributed to the fact that the model does not consider an easy interpretation of a single independent variable. However, a random subset of the independent variables is used for each tree at each node. In this regard, it avoids the overfitting problem and enhances the prediction accuracy (Al-Mukhtar, 2019; Francke et al., 2008).

The lower performance observed for the ANN compared with the RF model can be attributed to the fact that ANN models are incapable of extrapolating beyond the data used for training. Therefore, overfitting is a complex problem in the training data for the ANN (Al-Mukhtar, 2019; Minns and Hall, 1996). These problems can yield a lower performance for the ANN since the RF model does not suffer from overfitting but

Table 2
Statistical Analysis of Physico-Chemical Parameters for Groundwater samples for the study area.

Variables	TDS	Cl^-	NO_3^-	SO_4^{2-}	HCO_3^-	K^+	Na^+	Ca^{2+}	Mg^{2+}	Zn	F^-
Minimum	289.3	5.3	0.01	0.01	159.2	0.01	5.8	3.2	4.3	0.01	0.01
Maximum	20588	8032	3855	6688	1786	326.7	2895	716.6	1913	18.2	66.7
Mean	1458	250.4	65.9	274.9	476.2	6.9	251.2	55.7	74.1	0.7	1.7
Standard deviation	1880	608.7	212.9	553.7	264.1	25	391.2	59.8	128.5	0.8	3.4

(unit: mg/L).

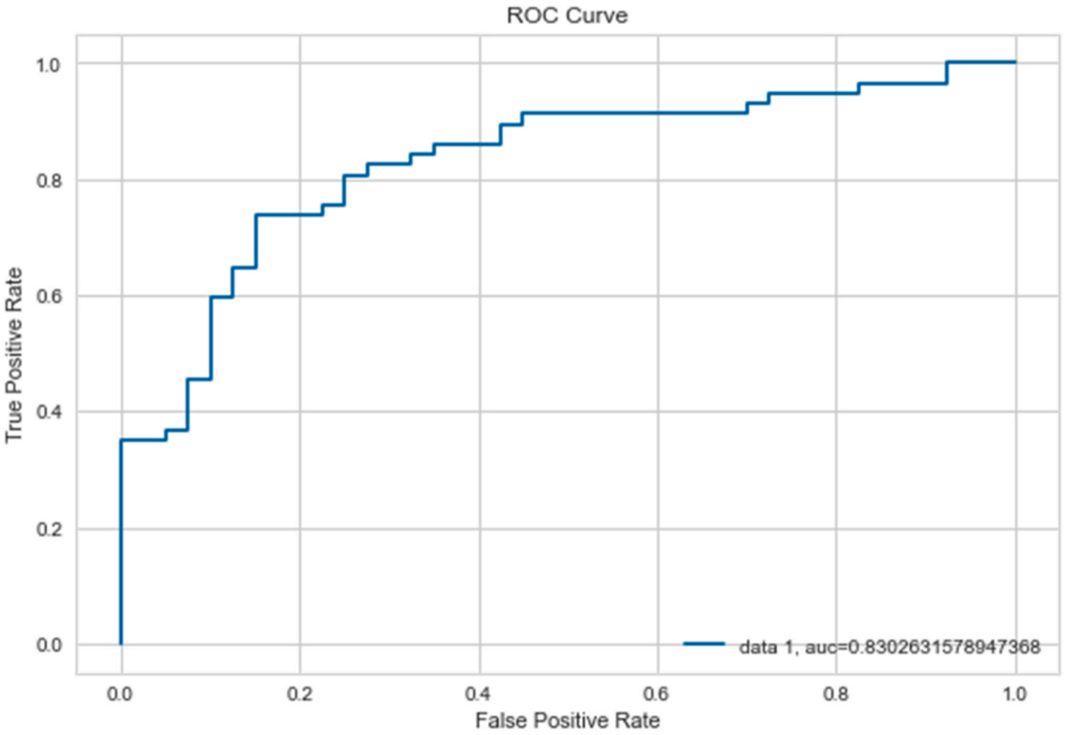

Fig. 3. Performance for Logistic Regression using ROC (AUC) curve.

instead combine many trees to produce the prediction that increases the model performance.

The LR demonstrated the lowest performance amongst the three models regarding accuracy, sensitivity, and specificity (Table 3). The lower performance of LR can occur in high dimensional data in the training data set, and the model may overfit and might not be accurate on the test data set. Despite the weak performance of ANN and LR in the current study, they are advantages in using them in other studies to predict groundwater contamination.

The process of groundwater contamination is complicated to understand due to the presence of several fluctuating variables. Consequently, the more flexible the algorithm, the greater the predictive and more reliable model (De'ath and Fabricius, 2000). An algorithm performance depends on the algorithm structure, the data nature, and the parameter selection (Asim et al., 2018). For statistical analysis, feature selection (e.g., Filter methods) should be considered to obtain an excellent predictive model in such classification tasks.

4.3. Identification of the variables influencing the fluoride mobilization

The relationship between predictors with fluoride was determined using the mean decrease in impurity (MDI), a measure used for variable importance in RF (Calle and Urrea, 2011). It is a tree-specific feature importance measure computed by the feature importance implemented in the "skirt-learn library" for RF in python. The sum of MDI for each feature across every forest tree is accumulated each time a variable is

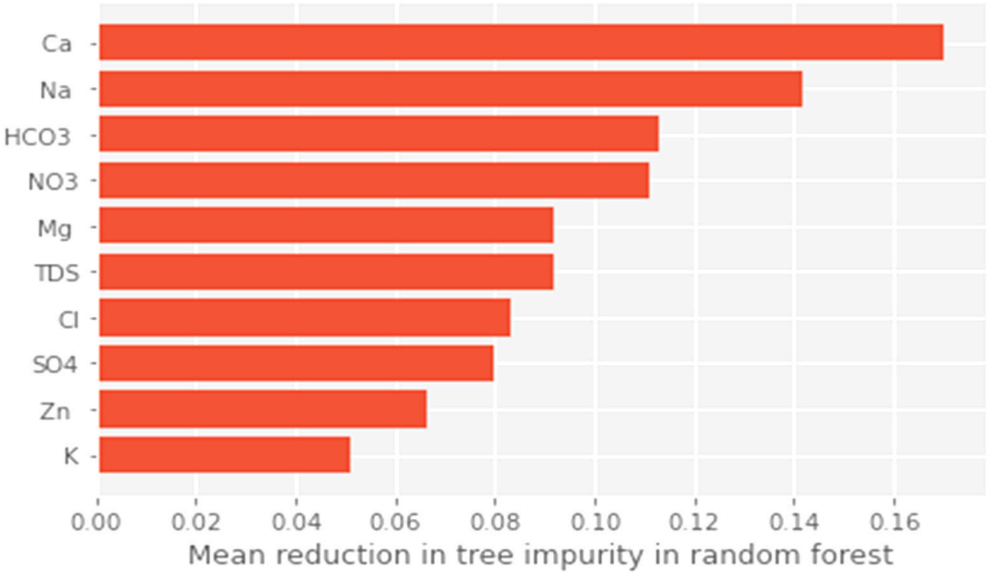

Fig. 4. Important Features to the Fluoride using Mean Decrease in Impurity in Random forest.

chosen to split a node. As demonstrated in (Fig. 4), the variables that tend to split nodes closer to tree root will have a more significant value. Thus, the essential variables of the model will be the highest in the plot and have the most significant MDI values, which are the cases of TDS, Cl^{-}, NO_3^{-}, Na^{+}, HCO_3^{-}, SO_4^{2-}, Ca^{2+}, Mg^{2+} in the plot. The variables K^{+} and Zn are lower in the plot, which means they have a small MDI and suggesting that they do not influence the fluoride in the study area.

The application of the MDI to determine the important variables in a dataset to the dependent variable was quoted in previous studies (Breiman, 2001; Meinshausen, 2007; Zhao, 2000). In addition, the MDI was used to identify significant predictors to the dependent variables in microarray and facies prediction studies and has shown the ability to identify the important variables related to the dependent variable (Archer and Kimes, 2008; Bhattacharya and Mishra, 2018). The result of the MDI demonstrates that TDS, Cl^{-}, NO_3^{-}, Na^{+}, HCO_3^{-}, SO_4^{2-}, Ca^{2+}, and Mg^{2+} are the variables influencing the fluoride in the study area, which is consistent with the findings of previous studies (Chae et al., 2007; Dhiman and Keshari, 2006; Guo et al., 2007).

The permutation feature was adopted to assess the variable importance of ANN to know the most influential variables on the output. The permutation decreases the definitive model score when eliminating a single variable (Maier and Dandy, 1996; Wen et al., 2013). Overall, eleven (11) networks were evaluated to determine the most significant variables to the output. Each one demonstrated the change observed in network accuracy variation after removing a variable (Table 4).

In the observation, after eliminating the variables K^{+} and Zn, the accuracy is 0.85 same as the original model accuracy. Therefore K^{+} and Zn could be excluded from the model as they do not affect the network accuracy and suggest that K^{+} and Zn do not enhance the fluoride in the study area. Conversely, with the elimination of other variables such as TDS, Cl^{-}, NO_3^{-}, Na^{+}, HCO_3^{-}, SO_4^{2-}, Ca^{2+}, and Mg^{2+}, the model accuracy decrease confirming their importance to the fluoride. Previous studies used the permutation feature to determine the most important variables to dissolved oxygen and learning event data (Matayoshi et al., 2019; Wen et al., 2013).

Therefore, in this study, the permutation feature and the mean decrease in impurity suggest the same results as TDS, Cl^{-}, NO_3^{-}, Na^{+}, HCO_3^{-}, SO_4^{2-}, Ca^{2+}, and Mg^{2+}, the variables influencing the fluoride in the study area. However, the permutation feature is applicable compared to the MDI to determine the relationship between the input and output variables. Thus, the permutation is appropriate to any algorithms to assess the essential variables to the output, but the MDI is an important measure feature limited to the RF algorithm.

Previous studies have demonstrated different chemicals and processes that influence fluoride in the study area (Su et al, 2013, 2015). The high fluoride in groundwater was generally characterized by the water type of HCO_3–Na(Mg), $HCO_3.SO_4$–Na(Mg) and $SO_4.Cl$–Na(Mg) (Su et al., 2013). Moreover, it stated that the increase in groundwater HCO_3^{-} concentration facilitates the fluorite dissolution, thereby promoting the release of fluoride into groundwater. The enrichment mechanism for fluoride concentration in groundwater is also related to cation exchange on the clay minerals, which causes the removal of Ca^{2+} by replacing it with Na $^{+}$ favoring the enrichment of groundwater fluoride (Rango et al., 2009; Saxena and Ahmed, 2003).

5. Conclusion

Globally, groundwater is an essential source of drinking water, especially in arid areas. The present study investigated and compared three algorithms, Random Forest, Logistic Regression, and Artificial Neural Networks, to predict groundwater fluoride in the Datong Basin. Our findings revealed that among the three algorithms implemented, RF suggests a significant prediction modeling to predict the fluoride in the study area with an accuracy of 0.89 and an error rate of 0.10. The mean decrease in impurity and the permutation feature were applied to determine the variables influencing the fluoride in the study area. These methods find similar variables related to fluoride, including TDS, Cl^{-}, NO_3^{-}, Na^{+}, HCO_3^{-}, SO_4^{2-}, $Ca^{2+,}$ and Mg^{2+}.

Table 4
Importance Features using Permutation Feature for ANN showing the change of the Accuracy after a variable is eliminated.

Variables	Accuracy variation for ANN
All variables	0.85
Eliminated TDS	0.77
Eliminated Cl^{-}	0.82
Eliminated NO_3^{-}	0.79
Eliminated SO_4^{2-}	0.78
Eliminated HCO_3^{-}	0.82
Eliminated K^{+}	0.85
Eliminated Na^{+}	0.78
Eliminated Ca^{2+}	0.77
Eliminated Mg^{2+}	0.80
Eliminated Zn	0.85

According to many model evaluation criteria, the RF algorithm outperformed the ANN and LR when predicting groundwater fluoride contamination. These results suggest that the RF model can be used as a consistent algorithm to predict groundwater fluoride in the Datong Basin and can be applied to other study areas in predicting groundwater contamination. However, for the consistent performance of RF to predict groundwater fluoride in the study area, future research should be focused on developing other models that should be more flexible in predicting groundwater contamination.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The research work was financially supported by the National Natural Science Foundation of China (4202010400, 41521001, and 41502230), the Ministry of Education of China (111 projects), and the Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) (No.CUGGC07).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apgeochem.2021.105054.

References

Al-Mukhtar, M., 2019. Random forest, support vector machine, and neural networks to modeling suspended sediment in Tigris River-Baghdad. Environ. Monit. Assess. 191, 673. https://doi.org/10.1007/s10661-019-7821-5.

Alagha, J.S., Said, M.A.M., Mogheir, Y., 2014. Modeling of nitrate concentration in groundwater using artificial intelligence approach—a case study of Gaza coastal aquifer. Environ. Monit. Assess. 186, 35–45.

Apambire, W.B., Boyle, D.R., Michel, F.A., 1997. Geochemistry, genesis, and health implications of fluoriferous groundwaters in the upper regions of Ghana. Environ. Geol. 33, 13–24.

Archer, K.J., Kimes, R.V., 2008. Empirical characterization of random forest variable importance measures. Comput. Stat. Data Anal. 52, 2249–2260.

Asim, Y., Shahid, A.R., Malik, A.K., Raza, B., 2018. Significance of machine learning algorithms in professional blogger's classification. Comput. Electr. Eng. 65, 461–473.

Awan, S.M., Riaz, M.U., Khan, A.G., 2018. Prediction of heart disease using artificial neural network. VFAST Trans. Softw. Eng. 6, 51–61.

Ayenew, T., 2008. The distribution and hydrogeological controls of fluoride in the groundwater of central Ethiopian rift and adjacent highlands. Environ. Geol. 54, 1313–1324.

Bhattacharya, S., Mishra, S., 2018. Applications of machine learning for facies and fracture prediction using Bayesian Network Theory and Random Forest: case studies from the Appalachian basin, USA. J. Petrol. Sci. Eng. 170, 1005–1017. https://doi.org/10.1016/j.petrol.2018.06.075.

Bowes, D., Hall, T., Gray, D., 2012. Comparing the Performance of Fault Prediction Models Which Report Multiple Performance Measures : Recomputing the Confusion Matrix 109–118.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Bylander, T., 2002. Estimating generalization error on two-class datasets using out-of-bag estimates. Mach. Learn. 48, 287–297.

Calle, M.L., Urrea, V., 2011. Letter to the editor: stability of random forest importance measures. Briefings Bioinf. 12, 86–89. https://doi.org/10.1093/bib/bbq011.

Chae, G.-T., Yun, S.-T., Mayer, B., Kim, K.-H., Kim, S.-Y., Kwon, J.-S., Kim, K., Koh, Y.-K., 2007. Fluorine geochemistry in bedrock groundwater of South Korea. Sci. Total Environ. 385, 272–283.

Coppola, L., Roy-Barman, M., Mulsow, S., Povinec, P., Jeandel, C., 2005. Low particulate organic carbon export in the frontal zone of the Southern Ocean (Indian sector) revealed by 234Th. Deep-Sea Res. Part I Oceanogr. Res. Pap. 52, 51–68.

De`ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81, 3178–3192.

Dhiman, S.D., Keshari, A.K., 2006. Hydrogeochemical evaluation of high-fluoride groundwaters: a case study from Mehsana District, Gujarat, India. Hydrol. Sci. J. 51, 1149–1162.

Dreyfus, G., 2008. Apprentissage statistique. Editions Eyrolles.

Francke, T., López-Tarazón, J.A., Schröder, B., 2008. Estimation of suspended sediment concentration and yield using linear models, random forests, and quantile regression forests. Hydrol. Process. An Int. J. 22, 4892–4904.

Gheyas, I.A., Smith, L.S., 2010. Feature subset selection in large dimensionality domains. Pattern Recogn. 43, 5–13.

Guo, H., Wang, Y., 2005. Geochemical characteristics of shallow groundwater in Datong basin, northwestern China. J. Geochem. Explor. 87, 109–120.

Guo, Q., Wang, Y., Ma, T., Ma, R., 2007. Geochemical processes controlling the elevated fluoride concentrations in groundwaters of the Taiyuan Basin, Northern China. J. Geochem. Explor. 93, 1–12.

Han, H., Guo, X., Yu, H., 2016. Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In: 2016 7th Ieee International Conference on Software Engineering and Service Science (Icsess). IEEE, pp. 219–224.

Hendrawan, Y., Murase, H., 2011. Neural-Intelligent Water Drops algorithm to select relevant textural features for developing precision irrigation system using machine vision. Comput. Electron. Agric. 77, 214–228. https://doi.org/10.1016/j.compag.2011.05.005.

Houéménou, H., Tweed, S., Dobigny, G., Mama, D., Alassane, A., Silmer, R., Babic, M., Ruy, S., Chaigneau, A., Gauthier, P., Socohou, A., Dossou, H.-J., Badou, S., Leblanc, M., 2020. Degradation of groundwater quality in expanding cities in West Africa. A case study of the unregulated shallow aquifer in Cotonou. J. Hydrol 582, 124438. https://doi.org/10.1016/j.jhydrol.2019.124438.

Keskin, T.E., Düğenci, M., Kaçaroğlu, F., 2015. Prediction of water pollution sources using artificial neural networks in the study areas of Sivas, Karabük, and Bartın (Turkey). Environ. Earth Sci. 73, 5333–5347. https://doi.org/10.1007/s12665-014-3784-6.

Khosravi, K., Barzegar, R., Miraki, S., Adamowski, J., Daggupati, P., Alizadeh, M.R., Pham, B.T., Alami, M.T., 2020. Stochastic modeling of groundwater fluoride contamination: introducing lazy learners. Groundwater 58, 723–734. https://doi.org/10.1111/gwat.12963.

Kim, Y., Kim, J.-Y., Kim, K., 2011. Geochemical characteristics of fluoride in groundwater of Gimcheon, Korea: lithogenic and agricultural origins. Environ. Earth Sci. 63, 1139–1148.

Lee, I.-H., Lushington, G.H., Visvanathan, M., 2011. A filter-based feature selection approach for identifying potential biomarkers for lung cancer. J. Clin. Bioinf. 1, 11. https://doi.org/10.1186/2043-9113-1-11.

Li, J., Wang, Y., Xie, X., Su, C., 2012. Hierarchical cluster analysis of arsenic and fluoride enrichments in groundwater from the Datong Basin, Northern China. J. geochemical Explore 118, 77–89.

Li, J., Wang, Yuting, Zhu, C., Xue, X., Qian, K., Xie, X., Wang, Yanxin, 2020. Hydrogeochemical processes controlling the mobilization and enrichment of fluoride in groundwater of the North China Plain. Sci. Total Environ. 730, 138877. https://doi.org/10.1016/j.scitotenv.2020.138877.

Li, P., Tian, R., Xue, C., Wu, J., 2017. Progress, opportunities, and key fields for groundwater quality research under the impacts of human activities in China with a special focus on western China. Environ. Sci. Pollut. Res. 24, 13224–13234.

Liu, Y., Wu, H., 2018. Water bloom warning model based on random forest. ICIIBMS 2017 - 2nd Int. Conf. Intell. Informatics Biomed. Sci. 2018-Janua 45–48. https://doi.org/10.1109/ICIIBMS.2017.8279712.

Lopez, A.M., Wells, A., Fendorf, S., 2020. Soil and aquifer properties combine as predictors of groundwater uranium concentrations within the central valley, California. Environ. Sci. Technol. https://doi.org/10.1021/acs.est.0c05591.

Maier, H.R., Dandy, G.C., 1996. The use of artificial neural networks for the prediction of water quality parameters. Water Resour. Res. 32, 1013–1022.

Matayoshi, J., Uzun, H., Cosyn, E., 2019. Deep (un) learning: using neural networks to model retention and forgetting in an adaptive learning system. In: International Conference on Artificial Intelligence in Education. Springer, pp. 258–269.

Meinshausen, N., Meinshausen, M.N., 2007. The quantregForest Package.

Minns, A.W., Hall, M.J., 1996. Artificial neural networks as rainfall-runoff models. Hydrol. Sci. J. 41, 399–417.

Mohammadi, A.A., Ghaderpoori, M., Yousefi, M., Rahmatipoor, M., Javan, S., 2016a. Prediction and modeling of fluoride concentrations in groundwater resources using an artificial neural network: a case study in Khaf. Environ. Heal. Eng. Manag. J.

Mohammadi, A.A., Ghaderpoori, M., Yousefi, M., Rahmatipoor, M., Javan, S., 2016b. Prediction and modeling of fluoride concentrations in groundwater resources using an artificial neural network: a case study in Khaf. Environ. Heal. Eng. Manag. 3, 217–224. https://doi.org/10.15171/ehem.2016.23.

Nadiri, A.A., Fijani, E., Tsai, F.T.C., Moghaddam, A.A., 2013. Supervised committee machine with artificial intelligence for prediction of fluoride concentration. J. Hydroinf. 15, 1474–1490. https://doi.org/10.2166/hydro.2013.008.

Nadiri, A.A., Naderi, K., Khatibi, R., Gharekhani, M., 2019. Modeling groundwater level variations by learning from multiple models using fuzzy logic. Hydrol. Sci. J. 64, 210–226. https://doi.org/10.1080/02626667.2018.1554940.

Naseem, S., Rafique, T., Bashir, E., Bhanger, M.I., Laghari, A., Usmani, T.H., 2010. Lithological influences on occurrence of high-fluoride groundwater in Nagar Parkar area, Thar Desert, Pakistan. Chemosphere 78, 1313–1321.

Nevtipilova, V., 2014. Testing artificial neural network (ANN) for spatial interpolation. J. Geol. Geosci. 1–9. https://doi.org/10.4172/2329-6755.1000145, 03.

Noshad, Z., Javaid, N., Saba, T., Wadud, Z., Saleem, M.Q., Alzahrani, M.E., Sheta, O.E., 2019. Fault detection in wireless sensor networks through the random forest classifier. Sensors 19, 1–21. https://doi.org/10.3390/s19071568.

Ostad-Ali-Askari, K., Shayannejad, M., Ghorbanizadeh-Kharazi, H., 2017. Artificial neural network for modeling nitrate pollution of groundwater in marginal area of Zayandeh-rood River, Isfahan, Iran. KSCE J. Civ. Eng. 21, 134–140. https://doi.org/10.1007/s12205-016-0572-8.

Ozdemir, A., 2011. Using a binary logistic regression method and GIS for evaluating and mapping the groundwater spring potential in the Sultan Mountains (Aksehir, Turkey). J. Hydrol 405, 123–136. https://doi.org/10.1016/j.jhydrol.2011.05.015.

Podgorski, J.E., Labhasetwar, P., Saha, D., Berg, M., 2018. Prediction modeling and mapping of groundwater fluoride contamination throughout India. Environ. Sci. Technol. 52, 9889–9898. https://doi.org/10.1021/acs.est.8b01679.

Qian, L., Zhang, R., Bai, C., Wang, Y., Wang, H., 2020. An Improved Logistic Probability Prediction Model for Water Shortage Risk in Situations with Insufficient Data 1 Introduction 1–31.

Rafique, T., Naseem, S., Bhanger, M.I., Usmani, T.H., 2008. Fluoride ion contamination in the groundwater of Mithi sub-district, the Thar Desert, Pakistan. Environ. Geol. 56, 317–326.

Rafique, T., Naseem, S., Usmani, T.H., Bashir, E., Khan, F.A., Bhanger, M.I., 2009. Geochemical factors controlling the occurrence of high fluoride groundwater in the Nagar Parkar area, Sindh, Pakistan. J. Hazard Mater. 171, 424–430.

Rango, T., Bianchini, G., Beccaluva, L., Ayenew, T., Colombani, N., 2009. Hydrogeochemical study in the Main Ethiopian Rift: new insights to the source and enrichment mechanism of fluoride. Environ. Geol. 58, 109–118.

Rapantova, N., Grmela, A., Vojtek, D., Halir, J., Michalek, B., 2007. Ground water flow modeling applications in mining hydrogeology. Mine Water Environ. 26, 264–270.

Rodriguez-Galiano, V., Mendes, M.P., Garcia-Soldado, M.J., Chica-Olmo, M., Ribeiro, L., 2014. Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: a case study in an agricultural setting (Southern Spain). Sci. Total Environ. 476–477, 189–206. https://doi.org/10.1016/j.scitotenv.2014.01.001.

Sánchez-Marono, N., Alonso-Betanzos, A., Tombilla-Sanromán, M., 2007. Filter methods for feature selection–a comparative study. In: International Conference on Intelligent Data Engineering and Automated Learning. Springer, pp. 178–187.

Saxena, V., Ahmed, S., 2003. Inferring the chemical parameters for the dissolution of fluoride in groundwater. Environ. Geol. 43, 731–736.

Stoltzfus, J.C., 2011. Logistic regression: a brief primer. Acad. Emerg. Med. 18, 1099–1104.

Su, C., Wang, Y., Xie, X., Li, J., 2013. Aqueous geochemistry of high-fluoride groundwater in Datong Basin, northern China. J. Geochem. Explor. 135, 79–92. https://doi.org/10.1016/j.gexplo.2012.09.003.

Su, C., Wang, Y., Xie, X., Zhu, Y., 2015. An isotope hydrochemical approach to understand fluoride release into groundwaters of the Datong Basin, Northern China. Environ. Sci. Process. Impacts 17, 791–801. https://doi.org/10.1039/c4em00584h.

Sutradhar, S., Mondal, P., 2021. Groundwater suitability assessment based on water quality index and hydrochemical characterization of Suri Sadar Sub-division, West Bengal. Ecol. Inf. 101335 https://doi.org/10.1016/j.ecoinf.2021.101335.

Tarasov, D.A., Buevich, A.G., Sergeev, A.P., Shichkin, A.V., 2018. High variation topsoil pollution forecasting in the Russian Subarctic: using artificial neural networks combined with residual kriging. Appl. Geochem. 88, 188–197.

Tesoriero, A.J., Gronberg, J.A., Juckem, P.F., Miller, M.P., Austin, B.P., 2017. Predicting redox-sensitive contaminant concentrations in groundwater using random forest classification. Water Resour. Res. 53, 7316–7331.

Tripathy, S.S., Bersillon, J.-L., Gopal, K., 2006. Removal of fluoride from drinking water by adsorption onto alum-impregnated activated alumina. Separ. Purif. Technol. 50, 310–317.

Uzair, M., Jamil, N., 2020. Effects of hidden layers on the efficiency of neural networks 1–6. https://doi.org/10.1109/INMIC50486.2020.9318195.

Wang, Y.X., Shpeyzer, G., 2000. Hydrogeochemistry of Mineral Waters from Rift Systems on the East Asia Continent: Case Studies in Shanxi and Baikal. China Environ. Sci. Press, Beijing (in Chinese with English Abstr.

Wen, X., Fang, J., Diao, M., Zhang, C., 2013. Artificial neural network modeling of dissolved oxygen in the Heihe River, Northwestern China. Environ. Monit. Assess. 185, 4361–4371. https://doi.org/10.1007/s10661-012-2874-8.

Xie, X., Ellis, A., Wang, Y., Xie, Z., Duan, M., Su, C., 2009. Geochemistry of redox-sensitive elements and sulfur isotopes in the high arsenic groundwater system of Datong Basin. China. Sci. Total Environ. 407, 3823–3835.

Xing, L., Guo, H., Zhan, Y., 2013. Groundwater hydrochemical characteristics and processes along flow paths in the North China Plain. J. Asian Earth Sci. 70, 250–264.

Zhao, G., 2000. A New Perspective on Classification. Utah State University.