Water Practice & Technology

© 2022 The Authors

Water Practice & Technology Vol 17 No 10, 2100 doi: 10.2166/wpt.2022.120

Predictive modelling physico-chemical properties groundwater in coastal plain area of Vinh Linh and Gio Linh districts of Quang Tri Province, Vietnam

Hong Giang Nguyen[a,*], Dinh Hieu Tran[a], Ngo Tu Do Hoang[b] and Tien Thinh Nguyen[c]

[a] Faculty of Architecture, ThuDauMot University, ThuDauMot 820000, Vietnam
[b] Faculty of Geology and Geography of Sciences University, Hue University, Hue 49118, Vietnam
[c] Department of International Business, National Kaohsiung University of Science and Technology, Kaohsiung 82445, Taiwan
*Corresponding author. E-mail: giangnh@tdmu.edu.vn

HGN, 0000-0002-1611-9079

ABSTRACT

This paper presents to study the performance of machine learning techniques consisting of multivariate adaptive regression spline (MARS), feed forward neural network-back propagation (FFNN-BP), and decision tree regression (DTR) for estimating the physico-chemical properties of groundwater in the coastal plain area in Vinh Linh and Gio Linh districts of Quang Tri province of Vietnam. With 290 groundwater samples collected in two districts, this study has identified three main elements CO_2, Ca, $CaCO_3$ for simulation. Quantitative analysis results have shown that these three components are such as $CaCO_3$ with from 0 to 25.8 mg/l, Ca from 0 to 87.55 mg/l and CO_2 from 0 to 12 mg/l. In the present examination, groundwater quality index (GQI) values and their representative categories have been referred by the Vietnam Groundwater Standard (QCVN01). Furthermore, the statistical accuracy parameters were used to compare among models. To deploy FFNN-BP and DTR, different types of transfer and kernel functions were tested, respectively. Determining the results of MARS, FFNN-BP and DTR showed that three models have suitable carrying out for forecasting water quality components. Comparison of outcomes of MARS model with the FFNN-BP and DTR models indicated that this model has good performance for forecasting the elements of water quality, its level of accuracy was slightly more than the other. To assess the accurate values of the models according to the measurement parameters for training phase illustrated that the order of the models was MARS to give the best result, followed by DTR and finally FFNN-BP, respectively.

Key words: groundwater, machine learning, physico-chemical properties, prediction

HIGHLIGHTS

- Machine learning methods are used for spatial modeling of physico-chemical properties of groundwater.
- MARS performances suitable precision compared to the DTR and FFNN-BP models.
- Total $CaCO_3$ value in the experiment samples adapted the regular limit of QCVN01 with 'Excellent' point.
- The quality of water parameters (i.e., $CaCO_3$, Ca, and CO_2) of the coastal plain area was predicted.
- The study results have shown that the water quality in these two districts is usable for humans, livestock, and agriculture activities.

1. INTRODUCTION

The presence of contaminants in natural freshwater is considered one of the most crucial environmental problems in many areas of developing countries, where several communities are hardly approaching a potable water supply (Organisation mondiale de la santé, Światowa Organizacja Zdrowia, World Health Organization, & World Health Organisation Staff 2004; Giang *et al.* 2021). Low-income communities, which lean on untreated surface and groundwater supplies for domestic and agricultural purposes are the most affected by poor water quality (Ayoko *et al.* 2007). Unfortunately, they also do not have adequate tools to monitor quality of water regularly (Resh 2008; Omarova *et al.* 2018; Najafzadeh & Niazmardi 2021). Thus, they are increasingly expected to obtain reliable assessments of quality of water, which can be used (Bonansea *et al.* 2015).

Climate change leads to seawater intrusion affecting groundwater resources of coastal cities (Kumar 2012; Alfarrah & Walraevens 2018). Urbanization and industrialization have caused uncontrolled over-exploitation and depletion of groundwater consequently (Kanwal *et al.* 2015; Sunardi *et al.* 2021). Furthermore, untreated wastewater from residential areas and industrial zones has seeped into the ground leading to an increasing amount and content of chemical elements in groundwater (Mukate *et al.* 2018; Khan *et al.* 2020a).

The chemical element of groundwater is considered a standard of measurement to show the capable level of groundwater for plenty of targets such as human and animal drinking, agricultural, and industrial activities. It has been shown in practice that the uses of groundwater sources require different standard indicators to distinct water quality circumstances (Loaiciga *et al.* 1992). The groundwater quality concept is an integrative index composed of chemical, physical, and biological features which maintain expected groundwater utilizations. Hence, groundwater is divided by composition as groundwater quality index (GQI) for the management and consumption of groundwater resources (Najafzadeh *et al.* 2021a).

To evaluate the quality of water for drinking and agricultural irrigation, several variables are routinely monitored. This process makes a big database, but it can be time-consuming for data acquisition while the accurate rendering of the multivariate data may be challenging.

With regard to use machine learning for forecasting physico-chemical parameters in water, using artificial neural network (ANN)-estimated river water quality components (Niroobakhsh *et al.* 2012; Emamgholizadeh *et al.* 2014; Najah *et al.* 2014; Raheli *et al.* 2017; Haghiabi *et al.* 2018; İlhan *et al.* 2021; Najafzadeh *et al.* 2021b); employing multivariate adaptive regressive splines (MARS) to predict physico-chemicals in water (Haghiabi 2016; Bhatt *et al.* 2017; Ahmadi *et al.* 2019; Esmaeilbeiki *et al.* 2020); deploying decision tree regression (DTR) to forecast quality of water (Liao & Sun 2010; He *et al.* 2012; Jaloree *et al.* 2014; Chandanapalli *et al.* 2018; Gakii & Jepkoech 2019; Jalal & Ezzedine 2020; Lu & Ma 2020). Furthermore. MARS, feed forward neural network-back propagation (FFNN-BP), and DTR models also belong to nonparametric learning, and the model is used in those areas (Bengio *et al.* 2010; Al Iqbal *et al.* 2012; Genuer *et al.* 2017; Khaldi *et al.* 2019; Kohler *et al.* 2019; Yurochkin *et al.* 2019; Antoniadis *et al.* 2020; Devianto *et al.* 2020; Khan *et al.* 2020b; Zheng *et al.* 2020; Amiri-Ardakani & Najafzadeh 2021). Najafzadeh & Ghaemi (2019) implemented the LS-SVM and MARS models to estimate BOD_5 and COD parameters through 200 samples collected from Karoun River, in the southwest of Iran. The result showed that the MARS model has proved precise approximations compared with real data. Saghebian *et al.* (2014) applied a decision tree model to classify groundwater quality in Ardebil, Iran. Research results have proved that this model can be acceptable range of criteria for quality classification of groundwater. Khan *et al.* (2021) used FFNN-BP model to estimate *Escherichia coli* in groundwater with 1,301 groundwater samples were obtained from 348 villages and cities in from 2016 to 2019 in Rajasthan state, India. Consequently, deploying the model based on Grover's algorithm was more efficient in forecasting all patterns in the calculated *E. coli* in groundwater. Najafzadeh *et al.* (2021a) studied the groundwater quality of the Rafsanjan Plain of Iran, quantifying it using artificial intelligence (AI) to assess GQI values for 15 years. The results of the groundwater quality prediction analysis of the MARS model with $\text{RMSE} = 2.444$ and $\text{SI} = 0.0304$. In addition, this result was also compared with the World Health Organization groundwater standard which also showed that the entire area of Rafsanjan lacks water quality at the 'Excellent' level with a high probability. The chance for 'Good' water quality varies from 1% (at $\text{GQI} = 50$ worst cases) to 55% (at $\text{GQI} = 100$ best cases).

Groundwater quality prediction work has some errors for various reasons such as the quality of collected groundwater samples, measurement of variability and the subjective opinion of groundwater sample analysts, and other random parameters related to groundwater quality prediction that have not been studied yet. Therefore, the problem of assessing reliability for implicit quality classifications. In addition, using analytical methods is also subject to the bias of environmentalists, geologists, and experts.

This paper presents the prediction of the physico-chemical properties of groundwater using FFNN-BP, DTR, and MARS models. The input vectors used in the models are leaned on 290 samples that were collected from 290 wells of households in coastal plain area in Gio Linh and Vinh Linh districts of Quang Tri province. With the support of the collected data, the GQI values were analyzed based on the Vietnam Groundwater Standard (QCVN01) and their relevance for proposed use. After that, highlight comparison among three models that base on the results of statistical accuracy parameters such as mean (M), bias (bias is shown by mean error), root mean square error (RMSE), mean absolute error (MAE), standard deviation (St Dev), pearson correlation coefficient (R), skewness coefficient (Skew), minimum (Min), maximum (Max), scatter index (SI), and Nash–Sutcliffe efficiency (NSE). Finally, the collection of results of these three models may show the working efficiency of the models for predicting the quality of water.

The structure of the paper is organized as follows. Section 1 gives the paper's introduction. Section 2 presents study area, the MARS, FFNN-BP, and DTR models and explains them clearly for understanding use throughout this paper. Section 3 describes the study results. Finally, Section 4, and Section 5 introduce the discussions and conclusions.

2. STUDY AREA AND METHODOLOGY

2.1. Study area

The place of study is about 150 km^2 and covers the Gio Linh and Vinh Linh coastal plain of Quang Tri province of Vietnam. It is surrounded by Quang Binh province in the north, Thach Han River in the south, 50–150 m high hills in the west, and the East Sea (see Figure 1). The coastal plain is relatively flat with an elevation between 0 and 5 m except for coastal sand-dunes at 11–22 m high, which provide a natural embankment system for seawater protection (Krutwagen 2007). During the dry season (from June to August), the saltwater pervades (i.e. where total mineralization of water M = 1 g/l) and is often inspected at about 30 and 35 km from the main tributaries of Ben Hai, Hieu, and Thach Han Rivers from the estuary (Tam *et al.* 2014). Thus, groundwater from dug wells and shallow wells scattered in the coastal is the main water source for drinking and domestic use for residents.

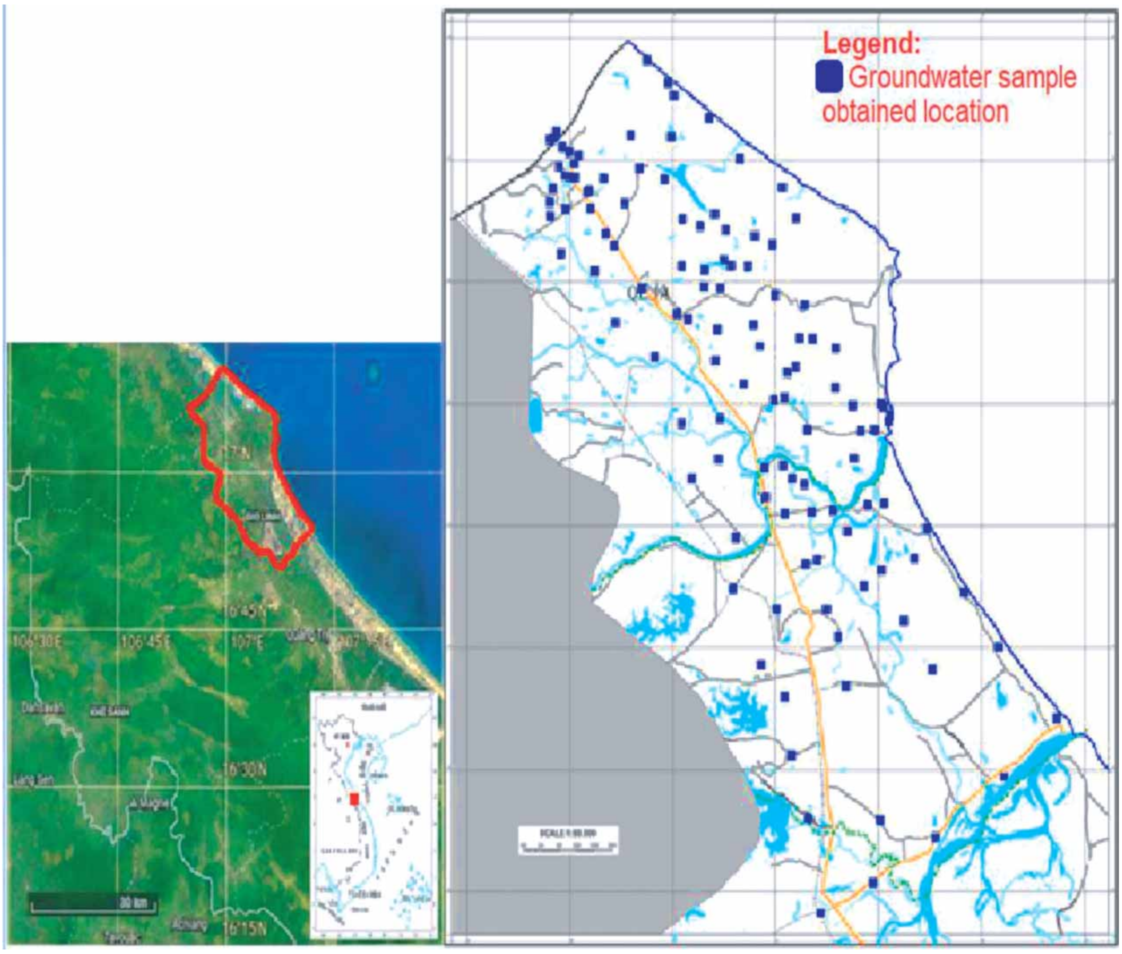

Figure 1 | The location for samples collection.

2.2. MARS

The MARS model is a novel approach in soft computing, and it is a nonparametric regression model, introduced by Friedman (Friedman *et al.* 2010). MARS seems like a method for a fitted relationship between independent and dependent variables in each desired phenomenon. MARS supports techniques for modelling systems with high accuracy, which is based on a dataset (Sekulic & Kowalski 1992; Steinberg 2001; Gutiérrez *et al.* 2009). The MARS algorithm feature is the procedure of the backward and forwards stepwise, at the same time may explain and control the complex nonlinear mapping between the inputs and output variables. MARS model highlights input variables that have a note worthy effect on the output variables. The general form of MARS is described as below:

$$y = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(x) \tag{1}$$

where, y is output variables, β_0 is constant value, M is the number of functions, $h_m(x)$ is M_{th} basis function and β_m is the corresponding coefficient of $h_m(x)$. Furthermore, $h_m(x)$ shows information about the relationship between input and output variables, and it is described as below:

$$h_m(x) = Max(0,\ C - x)\ or\ h_m(x) = Max(C - x,\ 0) \tag{2}$$

where h is the basis function, x is the input variables, and C is the threshold value of the independent (input) variables of x. It is notable that C is called '*knots*' or '*hinges*'.

The function of backward stepwise function relates to removing basis functions one at a time until the criterion of '*lack of fit*' is a minimum. In the deletion of backward stepwise, the last important basic functions are destroyed one at a time. The lack of used fitting measurement is based on generalized cross-validation (GCV) (Attoh-Okine *et al.* 2009):

$$CGV = A * \sum_{j=1}^{P} (y_i - \hat{f}(x)/N \tag{3}$$

where $A = \left[1 - \frac{C(M)}{N}\right]^{-2}$ and $C(M) = 1\ trace[B(B^{-}B)^{-1}B']$ are the complexity function (Friedman *et al.* 2010). The GCV criterion is considered the average of residual error multiplied by a penalty to modify for the variability associated with more parameters estimation in the model (LeBlanc & Tibshirani 1994).

2.3. Feed forward neural network-back propagation

FFNN-BP model is a member of neural network method (Khoshhal & Mokarram 2012). It may simulate arbitrarily complex nonlinear processes for any systems in terms of inputs and outputs. A FFNN-BP structure in Figure 2 demonstrates a three-layer neural network consisting of inputs layer, hidden layer (layers) and outputs layer (Ramchoun *et al.* 2016; Ashiquzzaman & Tushar 2017). The physico-chemical properties groundwater presents neural network that have trained for its FFNN-BP regression. The input layer has 290 input nodes from e_1 to e_{290}, one neuron of the output layer has deputized the values of physico-chemical properties groundwater. There are four hidden layers include the first hidden layer contains neurons from H_{11} to H_{1100}, the second layer is from H_{21} to H_{2100}, the third layer is from H_{31} to $H_{3100,}$ and the last one is from H_{41} to H_{4100}. Each neuron of the hidden and output layers get a corresponding weight and bias, as w_{11}^2, $B_1^{(2)}$ and w_{12}^2, $B_2^{(2)}$ are the weight and bias to represent for neurons of H_{11} and H_{12}, so on. The values of weight and bias can be assigned progressively and corrected during the training process in order to compare predicted outputs with known outputs. As networks are often trained using a backpropagation algorithm (Ashiquzzaman & Tushar 2017). Each neuron of the hidden layers attains the output from all neurons of the previous layers and converts these values with a weighted linear sum into the output layer. The output layer receives the values from the last hidden layer. The ReLU function is deployed as the activation function for the hidden layers. Adam method is stochastic optimizations to the solver of weight optimization.

2.4. DTR

A decision tree is a data structure that includes an arbitrary number of nodes and branches at each node (Pekel 2020). The values of the input variable(s) consider a particular function in the training stage (Loh 2011). The stimulant of the decision tree is an algorithm that generates a decision tree from given instances. The structure of decision tree regression is as below, assuming $X = X_1, X_2, \ldots, X_{mn}$, mn are estimator variables, a total number of estimator variables, respectively. At the same time, n and $Y = Y_1, Y_2, \ldots, Y_n$ describe the number of observations and a goal variable that doing continuous values, respectively. In addition, vf, th, t, and γ (γ = (vf,

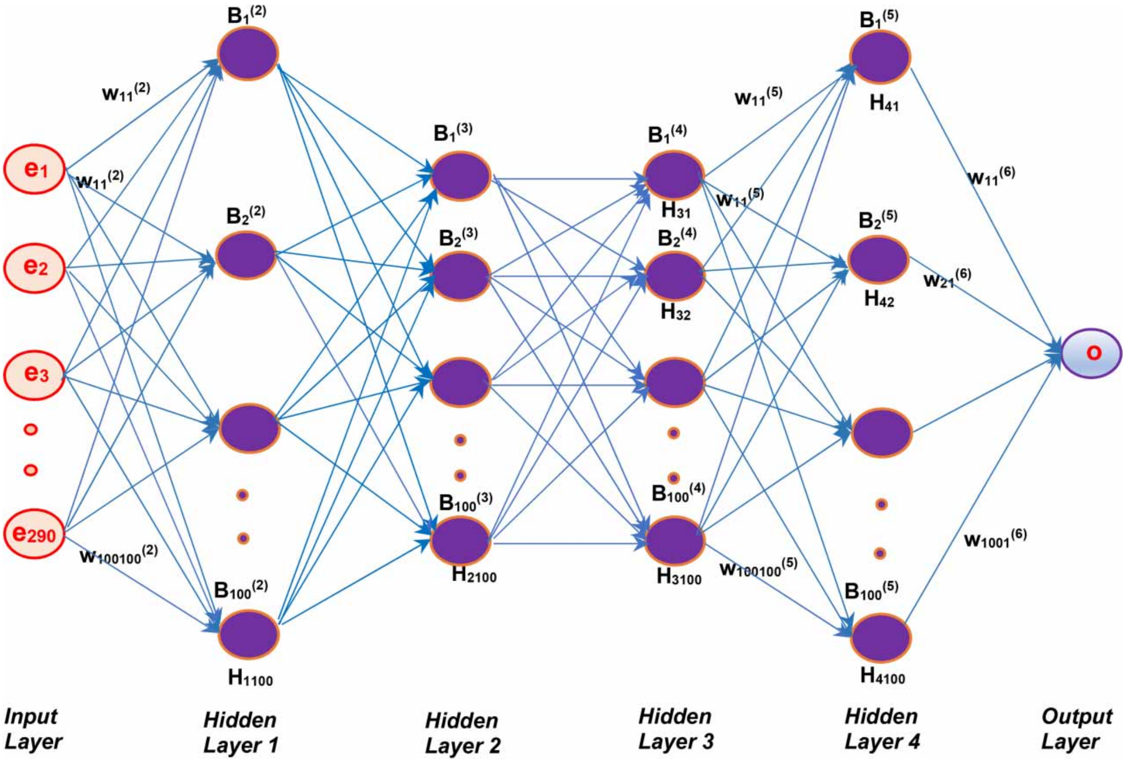

Figure 2 | Structure of FFNN-BP network for physico-chemical properties groundwater prediction.

th_t)) are a characteristic of a variable, a value of threshold, a node, and a candidate split, respectively. Q_l is realized by splitting the data into γ. Q_l and Q_r are calculated in Equations (4) and (5) declare the left and right sides in the decision tree of Q.

$$Q_l = ((x, y)|x_{vf} \leq th_t) \tag{4}$$

$$Q_r = ((x, y)|x_{vf} > th_t) \tag{5}$$

Let calling n_l and n_r are the number of a sampling of left and right sides; I is a function of impurity. Equation (6) indicates how the function of impurity is computed. The function of impurity is minimized by pondering Q and γ indicators as:

$$I(Q,\ \gamma) = \frac{n_l}{n} S(Q_l(\gamma)) + \frac{n_r}{n} S(Q_r(\gamma)) \tag{6}$$

2.5. Performance metrics

Forecasting results are based on the calculation and comparison of the actual values to the forecasted values. These metrics of the accuracy measurement parameters include MAE, RMSE, NSE, bias, SI. Furthermore, the error metrics are defined as follows (Yang & Yang 2005; Touzani *et al.* 2018; Kardani *et al.* 2020):

$$\text{MAE} = \frac{\sum_{t=1}^{n} |x_t - x'_t|}{n} \tag{7}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (x_t - x'_t)^2}{n}} \tag{8}$$

$$NSE \; = 1 - \frac{\sum_{t=1}^{n}(x_t - x'_t\,)^2}{\sum_{t=1}^{n}(x_t - \; \bar{x})^2} \tag{9}$$

$$R = \frac{\sum_{t=1}^{n}(x_t - \; \bar{x})(x'_t - \overline{x'})}{\sqrt{\sum_{t=1}^{n}(x_t - \; \bar{x})^2}\sqrt{\sum_{t=1}^{n}(x'_t - \overline{x'})^2}} \tag{10}$$

$$Bias \; = \frac{\sum_{t=1}^{n} x_t - \; x'_t}{n} \tag{11}$$

$$SI = \frac{RMSE}{average\ observed\ values} \tag{12}$$

where x_t, x'_t are the estimated value and observed value in the period time *t*, and *n* is the number of the observed values in the testing data. $\bar{x}$, $\overline{x'}$ are mean of the observed value. The NSE, R should be approaching *1* to indicate strong model performance, and the bias, MAE, and RMSE should be as close to *zero* as possible.

2.6. Data collection

Study data includes 290 groundwater samples obtained at wells of households in coastal plain area of Vinh Linh and Gio Linh districts. The predominant chemical compositions in these samples consisted of three main ingredients as calcium carbonate ($CaCO_3$) calcium (Ca), and carbon dioxide (CO_2). In addition, there were some other physico-chemical components (as ammonia, magnesium, and iron oxide), but their contents were not significant in these samples. Three input variables include Ca, CO_2, and $CaCO_3$, which was collected from 290 wells of two districts' households. The statistical characteristic results are also pointed out in Table 1. The range of the following characteristics was computed from the observation: the mean, min, and max values, St Dev, skew. The mean and standard deviation of the $CaCO_3$, Ca and CO_2 were 1.30 mg/l and 4.23 mg/l, 6.05 mg/l and 16.1 mg/l, 0.79 mg/l and 2.42 mg/l, respectively. The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero. Negative values for the skewness indicate skewed left data, and positive values for the skewness indicate skewed right data (Sahu *et al.* 2003; Brys *et al.* 2004). Hence, the skew of data fluctuating from 1.56 mg/l to 2.06 mg/l could be considered acceptable for prediction through these models. The input data patterns of 290 items were randomly selected with two parts. The first part was used for the training phase, which contained about 70% of the entire data. The second part was used for the test phase, which contained about the remaining 30%. In addition, the methodology of this study is described by the diagram in Figure 3. The process was summarized by experimental stages as below. Firstly, the collected dataset is preprocessed and tested statistical procedure, and the data is also divided into training phase and testing phase. Secondly, the FFBB-PB, MARS, and DTR models are employed based on the training samples, and to acquire the best network parameters. Finally, the performances of the algorithms are compared by using metrics of the accuracy parameters, and looking for the most suitable forecasting model is found for the study.

Table 1 | Statistical characteristics of physico-chemical components data

Item	St Dev	Mean	Min	Max	Skew
$CaCO_3$	4.23	1.30	0	25.80	2.06
Ca	16.1	6.05	0	87.55	1.57
CO_2	2.42	0.79	0	12	1.56

Unit: mg/l.

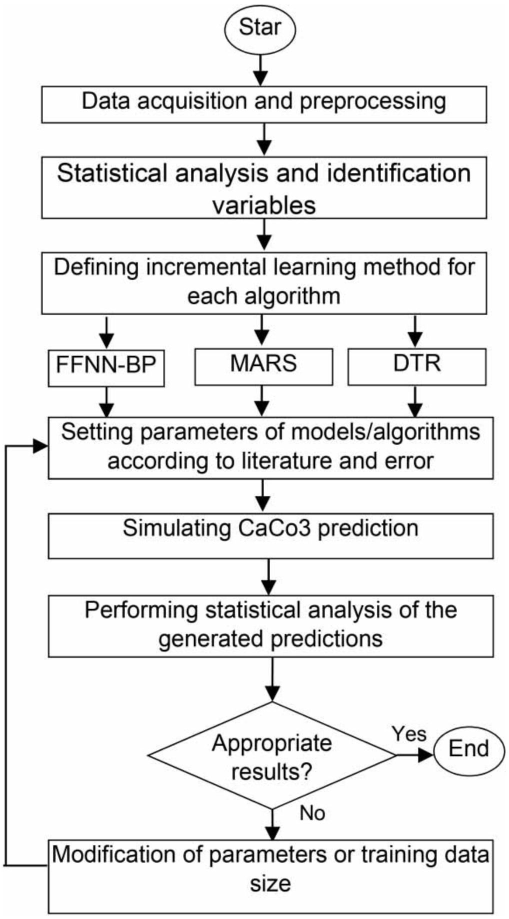
Star
Data acquisition and preprocessing
Statistical analysis and identification variables
Defining incremental learning method for each algorithm
FFNN-BP
MARS
DTR
Setting parameters of models/algorithms according to literature and error
Simulating CaCo3 prediction
Performing statistical analysis of the generated predictions
Appropriate results?
Yes
End
No
Modification of parameters or training data size

Figure 3 | Flowchart of the experimental steps conducted in this study.

3. RESULTS

The output function of MARS is presented as below:

MARS = 7.907–0.249F_1–0.129F_2, where F_1 = max(0, Ca – 55.79), F_2 = max(0, 55.79 – Ca).

F_i is the basis function. F_1 may be explained as the maximum value of 0 and Ca – 55.79. The minus sign ahead of the maximum value is equivalent to a minimum value. In addition, the MARS analysis indicates that the most important is Ca. Furthermore, the output function for FFNN-BP and DTR do not occur.

The data in Figure 4(a)–4(c) shows the relationship between the three variables. The content of CO_2 and Ca increase lead to the content of $CaCO_3$ increase. The FFNN-BP model makes a forecasting form that resembles

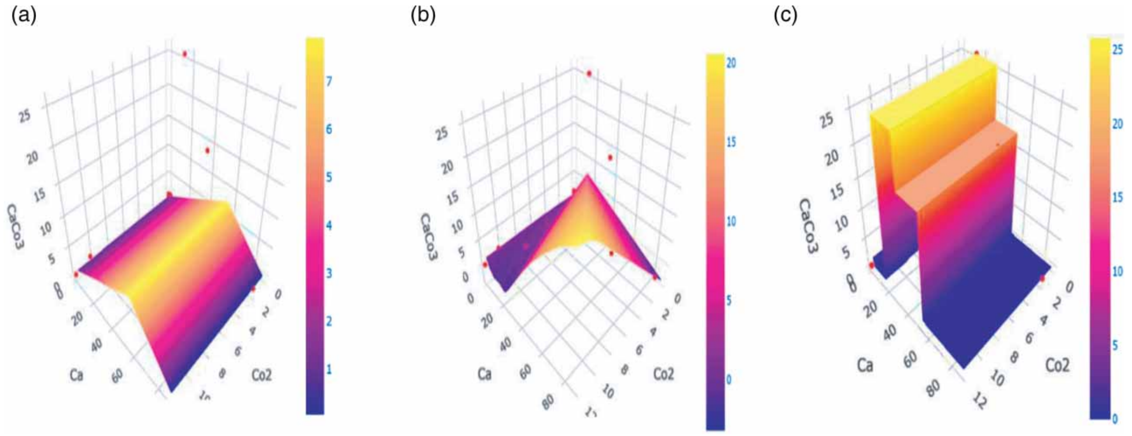

Figure 4 | Physico-chemical properties prediction with (a) MARS model, (b) FFNN-BP model, and (c) DTR model (Unit: mg/l).

a cone shape. In the meantime, the DTR and MARS charts look like the image of papers with some folds. Through these three images, it is hard to judge which model gives the best estimating. Hence, the values of performance metrics of the three models are presented in Table 2, Figure 5. The NSE of the three models for the training and testing phases are from 0.89 to 0.95 and are closer to 1. The MAE, RMSE, and bias values are also from −0.12 mg/l to −0.09 mg/l and are closer to 0. In addition, the values of the SI statistical indicator are simulated fluctuation from 0.21 mg/l to 2.23 mg/l. These show that the forecast results are very consistent compared with the actual data. As for the experimental results for each specific model, it indicates that the MARS model for training phase with NSE, MAE, RMSE, bias, and SI values are 0.95, 0.14 mg/l, 0.24 mg/l, 0.00 mg/l, and 0.26 mg/l respectively, and these properties are better than DTR and FFNN-BP models. Regarding the testing phase results, the highest accuracy for forecast is MARS model with NSE = 0.95, the second-highest is FFNN-BP model with NSE = 0.91, and the lowest is DTR model with NSE = 0.89.

The Taylor charts check the performance of estimated and actual values based on the standard deviation and Pearson Correlation Coefficient (Qin & Xiao 2018), which are contained simultaneously in assessing the

Table 2 | Accuracy parameters for physico-chemical components prediction

	DTR		**FFNN-BP**		**MARS**	
Parameter	**Testing**	**Training**	**Testing**	**Training**	**Testing**	**Training**
MAE (mg/l)	0.25	0.25	0.50	0.46	0.21	0.14
RMSE (mg/l)	0.34	0.33	0.99	0.90	0.41	0.24
Bias (mg/l)	−0.09	−0.19	−0.12	0.01	−0.04	0.00
SI (mg/l)	3.23	3.10	1.53	1.19	0.21	0.26
R	0.91	0.93	0.92	0.91	0.93	0.94
NSE	0.89	0.94	0.91	0.90	0.95	0.95
GCV (mg/l)					0.14	0.14

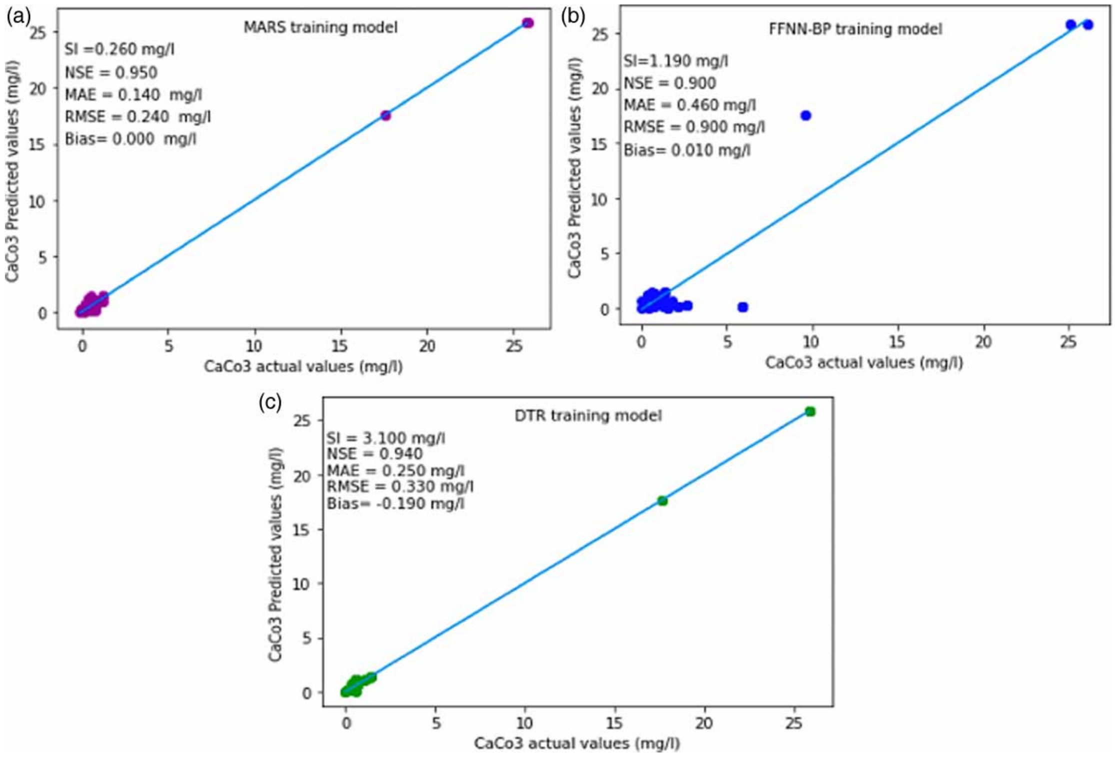

Figure 5 | The best performance indicators for $CaCO_3$ prediction (a) MARS training model, (b) FFNN-BP training model, (c) DRT training model.

respective models (Taylor 2001; Ghorbani *et al.* 2018). The standard deviation and CC between the actual and predicted datasets for the models are present in the Taylor diagram, and it also can be seen overall consistency between observed and estimated values when the CC value is approaching up to 1, as pointed in Figure 6. This can be considered for the MARS model with $CC_{training\ phase} = 0.94$, $CC_{testing\ phase} = 0.93$, FFNN-BP model with $CC_{training\ phase} = 0.91$, $CC_{testing\ phase} = 0.92$, and DTR model with $CC_{training\ phase} = 0.93$, $CC_{testing\ phase} = 0.91$. The large number of correlation coefficients indicate that there is a strong relationship. The Taylor plot also shows that these models are optimal with the highest accuracy (Taylor 2001). In other words, if the standard deviation of the predicted value of the higher standard deviation of the observed value, it will lead to an over estimation and vice versa (Abba *et al.* 2020). Furthermore, the GCV indicator of MARS brings about equilibrium between flexibility and generalization ability of the function of MARS model (Deo *et al.* 2016).

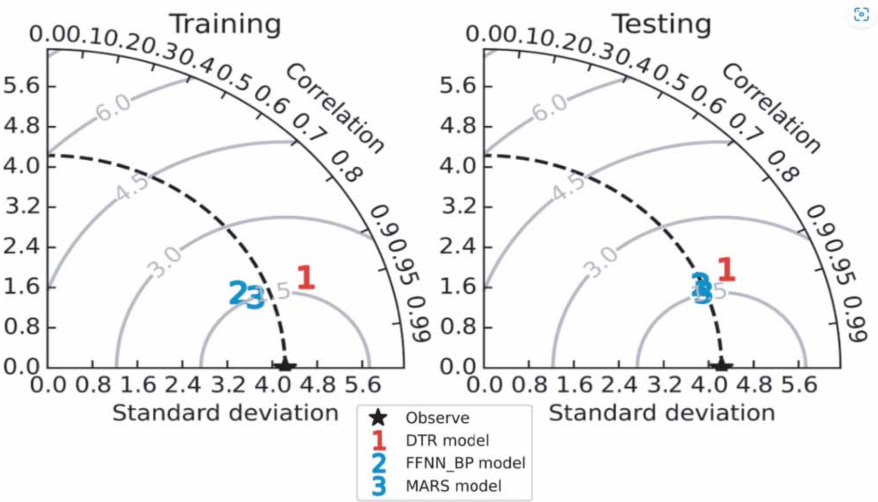

Figure 6 | The best performance indicators for $CaCo_3$ prediction for Training, and Testing.

4. DISCUSSIONS

Invasive seawater, untreated wastewater from residential areas and industrial zones, and over-exploitation of groundwater have been seriously affecting the quality and quantity of the underground water system. Therefore, a water quality assessment is a regular and continuous work to help people and authorities show solutions to treat groundwater to serve daily life. Hence, this paper described a comparative study and analysis of MARS, FFNN-BP, and DTR models in estimating physico-chemical properties of groundwater. The different circumstances, influential factors, and indicators have been observed for the experimentation. The following key findings are as the predictive errors in the case of the models decreased if the testing set decreased; MARS was the highlight in comparison to other models. Furthermore, to compare RMSE and MAE using MARS model of this study result with the study result of Najafzadeh *et al.* (2021a) about the groundwater quality indicate that their RMSE = 0.55 mg/l, and MAE = 0.00 mg/l are equivalent this study with RMSE = 0.41 mg/l, MAE = 0.21 mg/l.

According to QCVN01, the groundwater index of $CaCO_3$ is from 0 to 300 mg/l (National technical regulation on domestic water quality of Vietnam, QCVN 01: 2009/BYT 2021). The indicators in Table 3 show the range of water quality categorization based on the quality index of weight groundwater for human consumption. However, the total $CaCO_3$ value in the experiment samples ranged from 0 to 25.8 mg/l, and the mean value was 1.3 mg/l, this point illustrated that adapted the regular limit of QCVN01 with 'Excellent'. These points indicate that the

Table 3 | Water quality categorization based on the quality index of weight groundwater (GWQI)

GWQI	< 50	$50 - 100$	**100–200**	**200–300**	> 300
Quality categorization	Excellent	Good	Bad	Very bad	Unsuitable for drinking

Unit: mg/l.

water quality in these two districts is usable for humans, livestock, and agriculture activities. Rapid urbanization, industrialization, and climate change will negatively affect the groundwater resource in the areas. Therefore, households need to build a water purification system for drinking; at the same time, the local government needs to supply a clean water system so that people have clean water for daily living. In addition, local authorities and households should install early warning sensors for changes in chemical content in groundwater at some wells. The work periodically checks and advises on pollution levels of groundwater.

The groundwater samples for the study were mainly collected at the beginning of the dry season, so the hydrodynamic coefficient is dissimilar from the rainy season. The groundwater analysis equipment had an effect on the study because the equipment did not detect any more chemicals in the water that affect public health in these two areas. However, the results of this study also contribute to supporting local authorities to have appropriate solutions to help households use clean water.

5. CONCLUSIONS

Assessing and estimating water quality is a difficult and complex task, this result will give warnings to users and authorities to have appropriate treatment solutions. Hence, this study has deployed the MARS, DTR, and FFNN-BP predicted physico-chemical properties groundwater in the coastal plain of Vinh Linh and Gio Linh, which is located in the north middle of Vietnam. For phases of training and testing carried out in the models, the observed data consisting of CO_2 and Ca was used as inputs, while $CaCO_3$ was used as output. The stimulated results pointed out that the three models have a high suitable presentation for forecasting water quality components. The best performance was related to the MARS. The results of DTR and FFNN-BP also showed that their accuracy is a suitable presentation for practical purposes. Furthermore, the conducting of a comparison of three models showed that the outcomes of MARS and DTR models were slightly more reliable in comparing with FFNN-BP. The qualitative description of the GWQI found that the whole region of Vinh Linh and Gio Linh districts gained 'Excellent' water quality with 100% cases. At the same time, the study results have shown that the water quality in these two districts is usable for humans, livestock, and agriculture.

In addition, this study demostrated that machine learning models play a key role in the decision-making progress for carring out an effects of climate changes, urbanization, and industrialization on quality of groundwater. Another possible future work is to enhance quality of groundwater analysis equipment may finding other elements in the groundwater, and samples should be collected evenly throughout the seasons of the year.

ACKNOWLEDGEMENTS

We would like to thank Quang Tri province of Departments of Natural Resources and Environment, and Science and Technology, and Vinh Linh and Gio Linh residence. We also thanks the project of Hue University 'Assessing the corrosion capacity of groundwater for foundation concrete structures in the northern coastal plains of Quang Tri province', the code DHH2013-01-151 has supported the collection of field data in Quang Tri province.

AUTHOR CONTRIBUTIONS

Conceptualization, Discussion, Writing, Material, Methods, Review, and Editing: Nguyen Hong Giang, Tran Dinh Hieu; Data collection, Writing: Hoang Ngo Tu Do; Methods: Thinh-Tien Nguyen. All authors have read and agreed to the published version of the manuscript.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Abba, S. I., Pham, Q. B., Usman, A. G., Linh, N. T. T., Aliyu, D. S., Nguyen, Q. & Bach, Q. V. 2020 Emerging evolutionary algorithm integrated with kernel principal component analysis for modeling the performance of a water treatment plant. *Journal of Water Process Engineering* **33**, 101081.

Ahmadi, M. H., Mohseni-Gharyehsafa, B., Farzaneh-Gord, M., Jilte, R. D., Kumar, R. & Chau, K. W. 2019 Applicability of connectionist methods to predict dynamic viscosity of silver/water nanofluid by using ANN-MLP, MARS and MPR algorithms. *Engineering Applications of Computational Fluid Mechanics* **13**(1), 220–228.

Alfarrah, N. & Walraevens, K. 2018 Groundwater overexploitation and seawater intrusion in coastal areas of arid and semi-arid regions. *Water* **10**(2), 143.

Al Iqbal, M. R., Rahman, S., Nabil, S. I. & Chowdhury, I. U. A. 2012 Knowledge based decision tree construction with feature importance domain knowledge. In: *2012 7th International Conference on Electrical and Computer Engineering* (Pran Kanai Saha, ed.). IEEE, Dhaka, Bangladesh, pp. 659–662.

Amiri-Ardakani, Y. & Najafzadeh, M. 2021 Pipe break rate assessment while considering physical and operational factors: a methodology based on global positioning system and data driven techniques. *Water Resources Management* **35**(11), 3703–3720.

Antoniadis, A., Lambert-Lacroix, S. & Poggi, J. M. 2020 Random forests for global sensitivity analysis: a selective review. *Reliability Engineering & System Safety* **206**, 107312.

Ashiquzzaman, A. & Tushar, A. K. 2017 Handwritten Arabic numeral recognition using deep learning neural networks. In: *2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR)* (Md. Atiqur Rahman Ahad, ed.). IEEE, Dhaka, Bangladesh, pp. 1–4.

Attoh-Okine, N. O., Cooger, K. & Mensah, S. 2009 Multivariate adaptive regression (MARS) and hinged hyperplanes (HHP) for doweled pavement performance modeling. *Construction and Building Materials* **23**(9), 3020–3023.

Ayoko, G. A., Singh, K., Balerea, S. & Kokot, S. 2007 Exploratory multivariate modeling and prediction of the physico-chemical properties of surface water and groundwater. *Journal of Hydrology* **336**(1–2), 115–124.

Bengio, Y., Delalleau, O. & Simard, C. 2010 Decision trees do not generalize to new variations. *Computational Intelligence* **26**(4), 449–467.

Bhatt, A. H., Karanjekar, R. V., Altouqi, S., Sattler, M. L., Hossain, M. S. & Chen, V. P. 2017 Estimating landfill leachate BOD and COD based on rainfall, ambient temperature, and waste composition: exploration of a MARS statistical approach. *Environmental Technology & Innovation* **8**, 1–16.

Bonansea, M., Rodriguez, M. C., Pinotti, L. & Ferrero, S. 2015 Using multi-temporal Landsat imagery and linear mixed models for assessing water quality parameters in Río Tercero reservoir (Argentina). *Remote Sensing of Environment* **158**, 28–41.

Brys, G., Hubert, M. & Struyf, A. 2004 A robust measure of skewness. *Journal of Computational and Graphical Statistics* **13**(4), 996–1017.

Chandanapalli, S. B., Reddy, E. S. & Lakshmi, D. R. 2018 DFTDT: distributed functional tangent decision tree for aqua status prediction in wireless sensor networks. *International Journal of Machine Learning and Cybernetics* **9**(9), 1419–1434.

Deo, R. C., Samui, P. & Kim, D. 2016 Estimation of monthly evaporative loss using relevance vector machine, extreme learning machine and multivariate adaptive regression spline models. *Stochastic Environmental Research and Risk Assessment* **30**(6), 1769–1784.

Devianto, D., Permathasari, P., Yollanda, M. & Ahmad, A. W. 2020 The model of artificial neural network and nonparametric MARS regression for Indonesian composite index. In *IOP Conference Series: Materials Science and Engineering*, Vol. 846, No. 1, p. 012007, IOP Publishing.

Emamgholizadeh, S., Kashi, H., Marofpoor, I. & Zalaghi, E. 2014 Prediction of water quality parameters of Karoon River (Iran) by artificial intelligence-based models. *International Journal of Environmental Science and Technology* **11**(3), 645–656.

Esmaeilbeiki, F., Nikpour, M. R., Singh, V. K., Kisi, O., Sihag, P. & Sanikhani, H. 2020 Exploring the application of soft computing techniques for spatial evaluation of groundwater quality variables. *Journal of Cleaner Production* **276**, 124206.

Friedman, J., Hastie, T. & Tibshirani, R. 2010 Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1), 1.

Gakii, C. & Jepkoech, J. 2019 *A Classification Model for Water Quality Analysis Using Decision Tree*.

Genuer, R., Poggi, J. M., Tuleau-Malot, C. & Villa-Vialaneix, N. 2017 Random forests for big data. *Big Data Research* **9**, 28–46.

Ghorbani, M. A., Deo, R. C., Yaseen, Z. M., Kashani, M. H. & Mohammadi, B. 2018 Pan evaporation prediction using a hybrid multilayer perceptron-firefly algorithm (MLP-FFA) model: case study in North Iran. *Theoretical and Applied Climatology* **133**(3), 1119–1131.

Giang, N. H., Hieu, T. D. & Do, H. N. T. 2021 *Predictive Modelling Physico-Chemical Properties Groundwater in Coastal Plain Area of Vinhlinh and Giolinh Districts of Quangtri Province, Vietnam*.

Gutiérrez, Á. G., Schnabel, S. & Contador, J. F. L. 2009 Using and comparing two nonparametric methods (CART and MARS) to model the potential distribution of gullies. *Ecological Modelling* **220**(24), 3630–3637.

Haghiabi, A. H. 2016 Prediction of longitudinal dispersion coefficient using multivariate adaptive regression splines. *Journal of Earth System Science* **125**(5), 985–995.

Haghiabi, A. H., Nasrolahi, A. H. & Parsaie, A. 2018 Water quality prediction using machine learning methods. *Water Quality Research Journal* **53**(1), 3–13.

He, Q., Dong, Z., Zhuang, F., Shang, T. & Shi, Z. 2012 Parallel decision tree with application to water quality data analysis. In: *International Symposium on Neural Networks* (Jun Wang, Gary G. Yen, Marios M. Polycarpou, eds.). Springer, Berlin, Heidelberg, pp. 628–637.

İlhan, N., Yetiş, A. D., Yeşilnacar, M. İ. & Atasoy, A. D. S. 2021 Predictive modelling and seasonal analysis of water quality indicators: three different basins of Şanlıurfa, Turkey. *Environment, Development and Sustainability* **24**(3), 1–35.

Jalal, D. & Ezzedine, T. 2020 Decision tree and support vector machine for anomaly detection in water distribution networks. In: *2020 International Wireless Communications and Mobile Computing (IWCMC)* (Azzam Mourad, Chadi Abou Rjeily, George Hadjichristofi, eds). IEEE, Beijing, China, pp. 1320–1323.

Jaloree, S., Rajput, A. & Gour, S. 2014 Decision tree approach to build a model for water quality. *Binary Journal of Data Mining & Networking* **4**, 25–28.

Kanwal, S., Gabriel, H., Mahmood, K., Ali, R., Haidar, A. & Tehseen, T. 2015 Lahore's groundwater depletion-A review of the aquifer susceptibility to degradation and its consequences. *University of Engineering and Technology Taxila. Technical Journal* **20**(1), 26.

Kardani, N., Zhou, A., Nazem, M. & Shen, S. L. 2020 Estimation of bearing capacity of piles in cohesionless soil using optimised machine learning approaches. *Geotechnical and Geological Engineering* **38**(2), 2271–2291.

Khaldi, R., El Afia, A. & Chiheb, R. 2019 Forecasting of BTC volatility: comparative study between parametric and nonparametric models. *Progress in Artificial Intelligence* **8**(4), 511–523.

Khan, M. J., Shah, B. A. & Nasir, B. 2020a Groundwater quality assessment for drinking purpose: a case study from Sindh Industrial Trading Estate, Karachi, Pakistan. *Modeling Earth Systems and Environment* **6**(1), 263–272.

Khan, Y. A., Shan, Q. S., Liu, Q. & Abbas, S. Z. 2020b A nonparametric copula-based decision tree for two random variables using MIC as a classification index. *Soft Computing* **25**(15), 1–16.

Khan, F. M., Gupta, R. & Sekhri, S. 2021 Superposition learning-based model for prediction of *E. coli* in groundwater using physico-chemical water quality parameters. *Groundwater for Sustainable Development* **13**, 100580.

Khoshhal, J. & Mokarram, M. 2012 Model for prediction of evapotranspiration using MLP neural network. *International Journal of Environmental Sciences* **3**(3), 1000–1009.

Kohler, M., Krzyzak, A. & Langer, S. 2019 Deep learning and MARS: a connection. *Stat* **1050**, 8.

Krutwagen, M. 2007 *Impact of Shrimp Pond Wastewater on the Estuaries and the Issue of Salinity Intrusion in the Quang Tri Province*. *Bachelor's Thesis*, University of Twente.

Kumar, C. P. 2012 Climate change and its impact on groundwater resources. *International Journal of Engineering and Science* **1**(5), 43–60.

LeBlanc, M. & Tibshirani, R. 1994 Adaptive principal surfaces. *Journal of the American Statistical Association* **89**(425), 53–64.

Liao, H. & Sun, W. 2010 Forecasting and evaluating water quality of Chao Lake based on an improved decision tree method. *Procedia Environmental Sciences* **2**, 970–979.

Loaiciga, H. A., Charbeneau, R. J., Everett, L. G., Fogg, G. E., Hobbs, B. F. & Rouhani, S. 1992 Review of ground-water quality monitoring network design. *Journal of Hydraulic Engineering* **118**(1), 11–37.

Loh, W. Y. 2011 Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1**(1), 14–23.

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

Mukate, S., Panaskar, D., Wagh, V., Muley, A., Jangam, C. & Pawar, R. 2018 Impact of anthropogenic inputs on water quality in Chincholi industrial area of Solapur, Maharashtra, India. *Groundwater for Sustainable Development* **7**, 359–371.

Najafzadeh, M. & Ghaemi, A. 2019 Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods. *Environmental Monitoring and Assessment* **191**(6), 1–21.

Najafzadeh, M. & Niazmardi, S. 2021 A novel Multiple-Kernel support vector regression algorithm for estimation of water quality parameters. *Natural Resources Research* **30**(5), 1–15.

Najafzadeh, M., Homaei, F. & Farhadi, H. 2021a Reliability assessment of water quality index based on guidelines of national sanitation foundation in natural streams: integration of remote sensing and data-driven models. *Artificial Intelligence Review* **54**(6), 1–33.

Najafzadeh, M., Homaei, F. & Mohamadi, S. 2021b Reliability evaluation of groundwater quality index using data-driven models. *Environmental Science and Pollution Research* **29**(6), 1–17.

Najah, A., El-Shafie, A., Karim, O. A. & El-Shafie, A. H. 2014 Performance of ANFIS versus MLP-NN dissolved oxygen prediction models in water quality monitoring. *Environmental Science and Pollution Research* **21**(3), 1658–1670.

National technical regulation on domestic water quality of Vietnam, QCVN 01: 2009/BYT. 2021. (accessed 24 December 2021).

Niroobakhsh, M., Musavi-Jahromi, S. H., Manshouri, M. & Sedghi, H. 2012 Prediction of water quality parameter in Jajrood River basin: application of multi layer perceptron (MLP) perceptron and radial basis function networks of artificial neural networks (ANNs). *African Journal of Agricultural Research* **7**(29), 4131–4139.

Omarova, A., Tussupova, K., Berndtsson, R., Kalishev, M. & Sharapatova, K. 2018 Protozoan parasites in drinking water: a system approach for improved water, sanitation and hygiene in developing countries. *International Journal of Environmental Research and Public Health* **15**(3), 495.

Organisation mondiale de la santé, Światowa Organizacja Zdrowia, World Health Organization, & World Health Organisation Staff 2004 *Guidelines for Drinking-Water Quality*, Vol. 1. World Health Organization.

Pekel, E. 2020 Estimation of soil moisture using decision tree regression. *Theoretical and Applied Climatology* **139**(3), 1111–1119.

Qin, B. & Xiao, F. 2018 A non-parametric method to determine basic probability assignment based on kernel density estimation. *IEEE Access* **6**, 73509–73519.

Raheli, B., Aalami, M. T., El-Shafie, A., Ghorbani, M. A. & Deo, R. C. 2017 Uncertainty assessment of the multilayer perceptron (MLP) neural network model with implementation of the novel hybrid MLP-FFA method for prediction of biochemical oxygen demand and dissolved oxygen: a case study of Langat River. *Environmental Earth Sciences* **76**(14), 1–16.

Ramchoun, H., Idrissi, M. A. J., Ghanou, Y. & Ettaouil, M. 2016 Multilayer perceptron: architecture optimization and training. *International Journal of Interactive Multimedia and Artificial Intelligence* **4**(1), 26–30.

Resh, V. H. 2008 Which group is best? Attributes of different biological assemblages used in freshwater biomonitoring programs. *Environmental Monitoring and Assessment* **138**(1), 131–138.

Saghebian, S. M., Sattari, M. T., Mirabbasi, R. & Pal, M. 2014 Ground water quality classification by decision tree method in Ardebil region, Iran. *Arabian Journal of Geosciences* **7**(11), 4767–4777.

Sahu, S. K., Dey, D. K. & Branco, M. D. 2003 A new class of multivariate skew distributions with applications to Bayesian regression models. *Canadian Journal of Statistics* **31**(2), 129–150.

Sekulic, S. & Kowalski, B. R. 1992 MARS: a tutorial. *Journal of Chemometrics* **6**(4), 199–216.

Sunardi, S., Ariyani, M., Withaningsih, S., Darma, A. P., Wikarta, K., Parikesit, P., Kamarudin, M. K. A. & Abdoellah, O. S. 2001 An alternative to neural nets: multivariate adaptive regression splines (MARS). *PC AI* **15**(1), 38–41.

Sunardi, S., Ariyani, M., Withaningsih, S., Darma, A. P., Wikarta, K., Parikesit, P., Kamarudin, M. K. A. & Abdoellah, O. S. 2021 Peri-urbanization and sustainability of a groundwater resource. *Environment, Development and Sustainability* **23**(6), 8394–8404.

Tam, V. T., Batelaan, O., Le, T. T. & Nhan, P. Q. 2014 Three-dimensional hydrostratigraphical modelling to support evaluation of recharge and saltwater intrusion in a coastal groundwater system in Vietnam. *Hydrogeology Journal* **22**(8), 1749–1762.

Taylor, K. E. 2001 Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres* **106**(D7), 7183–7192.

Touzani, S., Granderson, J. & Fernandes, S. 2018 Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings* **158**, 1533–1543.

Yang, J. H. & Yang, M. S. 2005 A control chart pattern recognition system using a statistical correlation coefficient method. *Computers & Industrial Engineering* **48**(2), 205–221.

Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N. & Khazaeni, Y. 2019 Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*. PMLR, pp. 7252–7261.

Zheng, X., Dan, C., Aragam, B., Ravikumar, P. & Xing, E. 2020 Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 3414–3425.

First received 12 July 2022; accepted in revised form 29 September 2022. Available online 6 October 2022