Water Practice & Technology

IWA PUBLISHING

Water Practice & Technology Vol 17 No 1, 336 doi: 10.2166/wpt.2021.120

Prediction of groundwater quality indices using machine learning algorithms

Hemant Raheja *, Arun Goel and Mahesh Pal

Department of Civil Engineering, NIT Kurukshetra, Haryana 136119, India
*Corresponding author. E-mail: raheja.hemant788@gmail.com

HR, 0000-0001-7751-1050; AG, 0000-0003-1150-0871; MP, 0000-0003-1805-2952

ABSTRACT

The present paper deals with performance evaluation of application of three machine learning algorithms such as Deep neural network (DNN), Gradient boosting machine (GBM) and Extreme gradient boosting (XGBoost) to evaluate the ground water indices over a study area of Haryana state (India). To investigate the applicability of these models, two water quality indices, namely Entropy Water Quality Index (EWQI) and Water Quality Index (WQI) are employed in the present study. Analysis of results demonstrated that DNN has exhibited comparatively lower error values and it performed better in the prediction of both indices, i.e. EWQI and WQI. The values of Correlation Coefficient (CC = 0.989), Root Mean Square Error (RMSE = 0.037), Nash–Sutcliffe efficiency (NSE = 0.995), Index of agreement (d = 0.999) for EWQI and CC = 0.975, RMSE = 0.055, NSE = 0.991, d = 0.998 for WQI have been obtained. From variable importance of input parameters, the Electrical conductivity (EC) was observed to be most significant and 'pH' was least significant parameter in predictions of EWQI and WQI using these three models. It is envisaged that the results of study can be used to righteously predict EWQI and WQI of groundwater to decide its potability.

Key words: deep neural network (DNN), entropy water quality index (EWQI), extreme gradient boosting (XGBoost), gradient boosting machine (GBM), water quality index (WQI)

HIGHLIGHTS

- Evaluating groundwater quality using WQI and EWQI method.
- Machine learning techniques are used for predicting groundwater quality.
- Prediction performance of DNN, XGBoost, and GBM models is compared.
- Present model optimizes number of parameters to be determined for evaluating water quality.
- DNN based groundwater quality prediction performs much superior than other two models.

1. INTRODUCTION

Water, being one of the most essential resources for mankind, is used for different purposes like drinking, irrigation, recreation, navigation, domestic, fisheries, industrial and many more. About 75% of the surface area of the earth is covered by water with a total quantity of about 1,386 M km^3 (Kaushik *et al.* 2004). Out of which, 97.5% of the earth's water is in the oceans, which is not appropriate for human utilization or consumption without proper treatments. Only 2.5% is available as a freshwater. Out of this, about 24.4 M km^3 is locked in polar ice caps and only 10.6 M km^3 is available as fresh water in reservoirs, rivers, lakes, streams, and groundwater. Groundwater is one of the basic sources, which is used for several purposes throughout the world such as for irrigation, drinking and industrial use. The groundwater quality is dependent on environmental conditions and geological features. Groundwater contamination is a major problem, which poses serious threats to human health and environmental quality worldwide (Mohamed *et al.* 2015). It may be due to various human activities such as industrial, agricultural and other related activities, which lead to leaching of organic matter, pesticides, and nitrates deep into the aquifer (Su *et al.* 2019). Further groundwater quality is usually determined by the concentration of physical, chemical, and biological parameters (Kumari & Rai 2020). Panghal *et al.* (2021) reported that approximately 82% of the area of Tosham block, Haryana, had poor and very poor water quality for drinking purposes and 18% of area was unsuitable for drinking. It was also observed by Kumari & Rai (2020) that 45.31% of the area of southern Haryana had poor and very poor water quality for drinking purposes in the month of May,

2014. Hence, it is necessary to know the chemical composition of groundwater in order to determine its suitability for drinking purposes.

Many researchers have evaluated the potability of groundwater based on major parameters; that is, pH, Electrical Conductivity (EC), Total Hardness (TH), Calcium (Ca^{2+}), Magnesium (Mg^{2+}), Sodium (Na^{+}), Potassium (K^{+}), Sulphate (SO_4^{2-}), Chloride (Cl^{-}), Bicarbonate (HCO_3^{-}), Nitrate (NO_3^{2-}), and Fluoride (F^{-}) as mentioned by (Vasanthavigar *et al.* 2010; Zahedi 2017; Kumari & Rai 2020; Maghrebi *et al.* 2021). However, measurement of all these parameters in groundwater often poses difficulty as it is a quite expensive and tedious task (Kumar *et al.* 2016). Hence, reducing the effective cost and the subjectivity for evaluating water quality is a great challenge. Keeping in view the typical issues related to easy determination of water quality, several water quality indices have been derived in the past based on various water quality parameters and are available in the literature. Therefore, the researchers have mentioned index-based methods for evaluating water quality for drinking purposes; that is, water quality index (WQI) by Adimalla *et al.* (2020b); Tyagi *et al.* (2020); Wang *et al.* (2020), Entropy water quality index method (EWQI) by Kumar & Augustine (2021), Canadian WQI by Dao *et al.* (2020), US National Sanitation Foundation WQI by Lumb *et al.* (2011) and so on. However, the EWQI is considered an effective method to provide exact and comprehensive information about the overall quality of water for drinking purposes and has been used in recent studies by Adimalla *et al.* (2020b) and Wang *et al.* (2020).

Various machine learning approaches have been used extensively to predict groundwater quality for drinking purposes due to their improved performance in comparison to statistical methods over last three decades (Haghiabi *et al.* 2018; Aldhyani *et al.* 2020; Lu & Ma 2020; Nayan *et al.* 2021). Due to easy availability of water quality data from different sources in India, machine learning approaches are being successfully applied. With the advancement of new machine learning tools like Deep Neural Network and XGBoost, these new approaches are being used to analyze and, further, to predict the water quality in comparison to already reported work using other conventional machine learning models. Granata *et al.* (2017) purposed two models, namely Support Vector Regression (SVR) and Regression Trees (RT), for predicting wastewater quality. Najafzadeh *et al.* (2019) suggested a model to forecast the water quality index of Karoun River in Iran by pairing gene expression programming (GEP), evolutionary polynomial regression (EPR), and model tree (MT). Another study conducted by Najafzadeh & Ghaemi (2019) predicted biochemical oxygen demand (BOD) and chemical oxygen demand (COD) of Karoun River in Iran using machine learning models. Najafzadeh *et al.* (2021) have used four different machine learning models such as Polynomial Regression (EPR), M5 Model Tree (MT), Gene-Expression Programming (GEP), and Multivariate Adaptive Regression Spline (MARS). These four models helped in predicting the water quality index of Karoun River in Iran. Najafzadeh & Niazmardi (2021) have attempted to predict the water quality of the Karoun River with eleven water quality indicators using a Support Vector Regression (SVR) model using different kernels.

Traditional groundwater quality modelling approaches use either time series analysis or statistical methods. These approaches work by assuming some relationship between the dependent and independent variables as well. Mostly they require data to satisfy some statistical characteristics (e.g. normal distribution etc.), hence, are found to achieve inferior results during prediction. On the other hand, Artificial Intelligence (AI) based models do not require any assumptions about data as well as the relationship between dependent and independent variables. Exhaustive literature review suggests that AI-based predictive models are also found to perform better than the statistical methods. Meyers *et al.* (2017) have proposed three different models such as ANN, SVM, and RF for evaluating water quality in the United Kingdom (UK). Nayan *et al.* (2021) have used the GBM model to predict the water quality of a Bangladesh river for irrigation purposes for the time period of 2013–2019. El Bilali *et al.* (2020) determined the groundwater quality for drinking purposes by using ANN models. Di *et al.* (2019) applied machine learning models to predict water quality in the Yangtze River in China. These studies suggest that ANN, SVM, and GBM work well in predicting water quality and hence can effectively be used for water quality modelling with the available datasets. Recently, new machine learning models like Extreme gradient boosting (XGBoost) and Deep Neural Network (DNN) have been applied in water engineering applications (Najah Ahmed *et al.* 2019; El Bilali *et al.* 2021; Ibrahem Ahmed Osman *et al.* 2021). DNN has emerged as a powerful tool for various applications in civil engineering as studied by (Wu *et al.* 2018; Dick *et al.* 2019; Kumar & Abraham 2019; Pal 2019). It seems from literature review that limited numbers of studies are available on the use of DNN, XGBoost and GBM algorithms in water resource engineering, especially predicting water quality for drinking purposes using two indices, EWQI and WQI, over Haryana

state in India. This study is aimed at comparing the performance of these three machine learning models on the basis of performance parameters such as Correlation Coefficient (CC), Root Mean Square Error (RMSE), Nash–Sutcliffe efficiency (NSE) and Index of agreement (d) on 392 data sets from Haryana state for 12 hydrochemical parameters. Two indices, EWQI and WQI, were manually calculated for the data set and compared with the predicted values obtained by applying DNN, XGBoost and GBM algorithms. Further most significant parameters affecting the groundwater quality were also determined from the same data set by using three algorithms. At the end uncertainty and reliability analyses were also carried out. The outcomes of the article are expected to provide scientific information about quality of the groundwater that will be further useful in management and sustainable development of the groundwater resource in Haryana state.

2. MATERIALS AND METHODOLOGY

2.1. Machine learning models

2.1.1. Gradient boosting machine (GBM)

Boosting algorithms were originally introduced by the machine learning community as stated by Freund (1995). The Gradient boosting machine learning brings together weak learners in a different way to develop a strong learner. As each weak learner is added, a new model is fitted to provide a more accurate estimate of the response variable. The new weak learners are maximally correlated with the negative gradient of the loss function, associated with the whole ensemble. The main aim of the GBM is to create a stronger prediction model by merging a group of relatively weak prediction models. The model forecasts values for the structure $\hat{y} = F(x)$ by limiting the mean squared error as given in Equation (1).

$$= \frac{1}{n}\sum_{i}(\hat{y}_i - y_i)^2 \tag{1}$$

where, $y_i =$ actual value, $\hat{y}_i =$ predicted value, i = equities over some test data of size, $n =$ number of samples in y.

2.1.2. Extreme gradient boosting (XGBoost)

The XGBoost algorithm has recently been used for classification/regression problems (Chen & Guestrin 2016) and is based on the gradient boosting algorithm (Friedman 2001). The XGBoost algorithm has been implemented in various engineering applications, such as shear strength (Zhang *et al.* 2021), and streamflow (Hadi *et al.* 2019; Yu *et al.* 2020). Two major changes in the design of the XGBoost over the gradient boosted decision tree are (1) using a regularization term in its objective function allowing it to be less prone to overfitting in comparison to the gradient-boosted decision tree and (2) using Taylor expansion on the objective function compared to the gradient-boosted decision tree, which uses first derivative in optimization, thus allowing XGBoost to define the loss function more accurately. The objective function of XGBoost that need to be minimized is defined by Equations (2) and (3) and given below:

$$\text{obj}(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(D_t) \tag{2}$$

where

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3}$$

where, $\hat{y}_i^{(t)}$ is the predicted value of target variable y_i at the t^{th} round, L is loss function, T is number of leaves in the tree, ω_j is score of j^{th} leaf and lambda (λ) is the regularization parameter (Chen & Guestrin 2016).

Further, D_t denotes an independent tree with γ as the penalty coefficient and $1/2\lambda \sum_{j=1}^{T} \omega_j^2$ is L_2 norm of leaf score. After t iterations, the function of the model is the $(t-1)^{\text{th}}$ iteration prediction function plus a new decision

tree. Thus, the Equation (2) is updated as Equation (4) and mentioned below:

$$\mathrm{Obj}^{t} = \sum_{i=1}^{n} L(y_i , \hat{y}_i^{(t-1)} + D_t(x_i)) + \Omega(D_t) \tag{4}$$

The importance of each feature is calculated based on their prediction performance. Each feature is replaced by the noise and prediction takes place for the particular instance. The higher the value of prediction indicates greater score of each feature. Thus, it assesses all features by the individual scores which become important during training and the feature with higher score plays the major role in making the key decision with boosted trees.

2.1.3. Deep neural network (DNN)

The processing node is basic component of a back-propagation neural network (BPNN). Each processing node of BPNN performs two functions and behaves as a biological neuron. Initially it sums up the input values and this sum of input is then passed through the activation function to create the output. The activation function *f* can be used as any differential function. In BPNN, all processing nodes are arranged into layers, which are fully interconnected to the next layer. There is no connection of nodes of the same layer. A BPNN, typically, consists of an input layer that serves as a distribution structure of data that is inputted into the network and is not used for any form of processing. After this layer, one or more processing layers are known as hidden layers. The output layer is the final processing layer. A neural network having two or more hidden layers with a huge number of nodes and using mathematical modeling is mostly known as a DNN (Pal 2019).

All the interconnections between each node have an associated weight. When a value is passed from the input layer, down these interconnections, these values are multiplied by the associated weight and summed to derive the net input n_j to the unit as shown below in Equation (5):

$$= \sum_{i} W_{ji} O_i \tag{5}$$

where, W_{ji} denotes the weight of the interconnection to unit *j* from unit *i* (called input) and O_i denotes the output of unit *i*. The net input obtained by the above equation is then transformed by the activation function to produce an output (O_i) for unit j.

Activation functions introduce non-linearity in the neural network to learn more complex features present in the data. Traditional activation functions: sigmoid and hyperbolic tangent with a BPNN are found to be affected by saturation and sensitivity to changes around their mid-point (Goodfellow *et al.* 2016). The rectified linear activation function (RELU); (Nair & Hinton 2010) is a piecewise linear function and considered to be a milestone in the design of a deep neural network. RELU activation function outputs the input value itself if it is positive, otherwise the output would be zero and easy to train. The use of RELU is found to achieve better performance than other activation functions with DNN. The RELU function is defined in Equation (6) as below:

$$f(n_j) = \max(0,\ n_j) \tag{6}$$

Initializing BPNN weights is an important factor that affects the functioning of the neural network. The initial weights should be selected before the start of network training and should be in a reasonable range. Random weight initialization is normally used with a standard BPNN to find an optimal set of weights using a stochastic gradient descent approach. Xavier weight initialization was proposed as the weight initialization technique for DNN because of the poor performance of random weight initialization with a standard gradient descent-based optimization approach. This approach assigns the weights from a Gaussian distribution with zero mean and some finite variance, thus allowing the variance of the outputs of a layer to be equal to the variance of its inputs.

During the training phase of BPNN, network weights are continuously updated and adjusted using the learning rate as one of the user-defined parameters. The value of the learning rate is generally selected randomly, based on the experiences and the earlier reported works. Gradient descent algorithms were generally used to update the network weights with the help of a learning rate in a BPNN. The introduction of adaptive learning rate methods allowed adjusting the learning rate adaptively throughout the training process of a DNN. In order to update the

weights of a DNN, an adaptive moment estimation based optimization algorithm was proposed (Kingma & Ba 2015) and found to perform well. Use of Adam requires several user-defined parameters during training but studies suggest that the default values as suggested by (Kingma & Ba 2015) work well with most of the datasets.

Similar to BPNN, optimal values of several user-defined parameters need to be obtained during training of a deep neural network. These parameters include the activation function, optimization algorithm, number and type of hidden layers, number of nodes in the hidden layer, hidden and dropout layers, updaters (i.e., learning rate optimization algorithm), weight initialization method, batch size (i.e., number of training samples used in one iteration) and the number of epochs. One epoch is defined as when an entire training dataset has passed once through the neural network both in forward and backward directions.

2.2. Study area and data

Haryana state is a landlocked state in the north-west part of India as shown in Figure 1. The latitude of Haryana state extends between 27°39′ and 30°35′N and longitudes between 74°27′ and 77°36′E. The total geographical area of Haryana state is 44,212 Km^2. The normal average annual rainfall is about 617 mm, which is received primarily through the southwestern monsoon. The climate of the state is arid to semi-arid with a very hot summer temperature of approximately 45 °C.

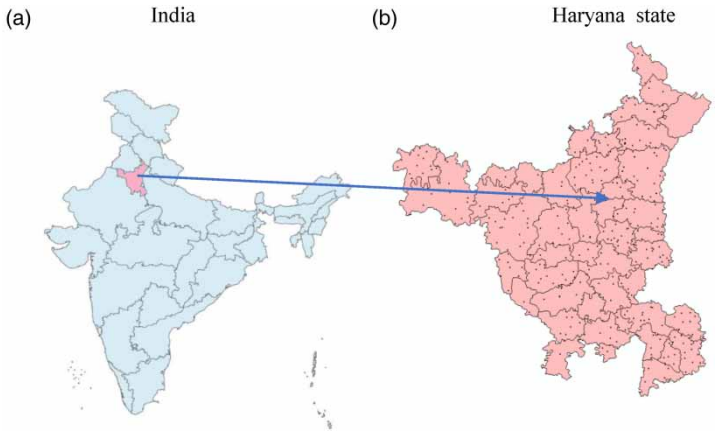

Figure 1 | Location map of (a) India and (b) Haryana showing the study area with groundwater samples position.

The dataset used in this study was downloaded from India Water Resources Information System (IWRIS; https://indiawris.gov.in). A dataset collected from national groundwater monitoring stations distributed Figure 1(b) over 20 districts of Haryana state during the year 2016, which was used to predict the groundwater quality for drinking purposes given in Table 1. A total of 392 samples were collected and analyzed for 12 important hydrochemical parameters so as to access the groundwater quality. This dataset was randomly divided in such a way that 294 (75%) samples are used for the training, whereas the remaining 98 (25%) were used for testing the models. To calculate the groundwater quality, two different water quality indices, Entropy water quality index (EWQI) and Water quality index (WQI) were calculated using 12 parameters. In order to reduce the effect of large variation in input and output values, normalization of different input and output parameters was also performed.

2.3. Entropy water quality index (EWQI)

EWQI has been frequently used to estimate the water quality for domestic/drinking, agricultural purpose etc. in the past and have been applied to determined water quality by (Shannon 1948). The steps involved in the calculation EWQI are as under follow.

In first instance, normalization is carried out on initial matrix 'X'. After this, the standard matrix as (7) & (8) is stated as $Y = (y_{ij})$ * (m × n) where:

$$X = \begin{bmatrix} x_{11} & x_{1n} \\ x_{m1} & x_{nm} \end{bmatrix} \tag{7}$$

$$Y = \begin{bmatrix} y_{11} & y_{1n} \\ y_{m1} & y_{nm} \end{bmatrix} \tag{8}$$

Table 1 | Summary of water quality observations of the Haryana state for year 2016

Name of parameters & unit	Minimum value	Mean value	Maximum value	Standard deviation value	Coefficient of variation (%)
pH	7.22	8.35	9.15	0.54	6.51
EC ($\mu s\ cm^{-1}$)	238.00	2,026.45	14,640.00	1,951.28	96.29
TH (mg/L)	39.00	457.06	6,505.00	587.06	128.4
Ca^{2+} (mg/L)	8.00	66.63	601.00	92.95	139.51
Mg^{2+} (mg/L)	0.17	75.73	1,216.00	103.33	136.44
Na^{+} (mg/L)	4.50	279.20	2,400.00	319.07	114.28
K^{+} (mg/L)	25	36.81	951.00	96.16	261.23
HCO_3^-(mg/L)	37.00	293.09	1,239.00	175.75	59.96
Cl^{-} (mg/L)	6.00	337.29	4,930.00	548.22	162.54
SO_4^{2-} (mg/L)	15	291.92	4,028.00	462.31	158.37
NO_3^{2-} (mg/L)	0.00	67.78	998.00	141.05	208.11
F^{-} (mg/L)	0.01	1.23	18.00	1.95	158.63

In Equations (7) and (8) m and n are the number of groundwater samples and parameter for a sample respectively.

Then the value of y_{ij} is obtained by Equation (9):

$$y_{ij} = \frac{x_{ij} - (x_{ij})_{\min}}{(x_{ij})_{\max} - (x_{ij})_{\min}} \tag{9}$$

where, x_{ij} is the j^{th} evaluation index of the i^{th} groundwater sample.

After calculating the standardized value, the ratio of index value of the j index using i sample is calculated by Equation (10):

$$P_{ij} = \frac{y_{ij}}{\sum_{i=1}^{m} y_{ij}} \tag{10}$$

Next step is to calculate entropy weight w_j and information entropy e_j by Equations (11) and (12) as:

$$e_j = -\frac{1}{\ln m} \sum_{i=1}^{m} y_{ij} \ln y_{ij} \tag{11}$$

$$w_j = \frac{1 - e_j}{\sum_{j=1}^{n} (1 - e_j)} \tag{12}$$

To calculate $EWQI$, a quality rating scale q_j for each parameter is assigned where q_j is obtained from the following Equation (13)

$$q_j = \frac{c_j}{s_j} \times 100 \tag{13}$$

where, c_j represents the concentration of the parameter (mg/l), and s_j denotes the water standards of groundwater for each drinking parameter according to Bureau of Indian standards (BIS). The EWQI can then be calculated by the following Equation (14):

$$EWQI = \sum_{j=1}^{n} w_j\, q_j \tag{14}$$

The results of EWQI obtained from above equation are listed in Table 2 (Aouiti *et al.* 2021). Out of 392 groundwater samples, only 14.80% of groundwater samples are unsuitable for drinking, while 32.40% of samples have good quality and only 17.60% samples exhibited excellent quality of groundwater for drinking purposes.

Table 2 | EWQI scale, water type, and % of samples in the study area (Wang *et al.* 2017)

Sr. No.	EWQI scale	Quality of groundwater	Number of samples	% of samples
1	<50	Excellent	69	17.60
2	50–100	Good	127	32.40
3	100–200	Poor	101	25.76
4	200–300	Very poor	37	9.44
5	>300	Not suitable for drinking	58	14.80

2.4. Water quality index (WQI)

The WQI method has been commonly used to access the drinking water quality in various geological topographies (Aly *et al.* 2015; Chung *et al.* 2015; Adimalla & Qian 2019; Adimalla *et al.* 2020a; Gaikwad *et al.* 2020). The following steps are includes for calculation of WQI according to Horton (1965):

1. In the first instance, 'assigning weight': twelve parameters are weighted according to its relative importance. The weight 5 is assigned to the most significant parameters and 1 is for the least significant. The relative weight (W_i) is obtained from Equation (15) and has been shown in Table 3.

$$W_i = \frac{wi}{\sum_{i=0}^{n} wi} \tag{15}$$

where, wi is the weight of each parameter and n is the number of parameters.

Table 3 | Relative weight of hydrochemical parameters

Parameters	Weight (wi)	Relative weight (W_i)
pH	4	0.091
EC	4	0.091
TH	4	0.091
Ca^{2+}	3	0.068
Mg^{2+}	3	0.068
Na^{+}	4	0.091
K^{+}	1	0.023
HCO_3^-	1	0.023
Cl^-	5	0.11
SO_4^{2-}	5	0.11
NO_3^{2-}	5	0.11
F^-	5	0.11

2. In second instance, 'quality rating (q_i) scale calculation' firstly, the water sample concentration is multiplied by 100 and then the results are divided by its limits given by the BIS (2015) in Equation (16):

$$q_i = \frac{C_i}{S_i} \times 100 \tag{16}$$

where, C_i is the concentration of chemicals in the water sample in (mg/l), and S_i is the drinking water standard for each chemical parameter BIS (2015).

3. In the third instance, 'calculation of water quality index (WQI)', firstly SI_i (water quality index of the i^{th} parameter) value is calculated by Equation (17) and WQI equals the sum of all the values of SI_i of each parameter as by Equation (18):

$$SI_i = W_i \times q_i \tag{17}$$

$$\text{WQI} = \sum_{i=1}^{n} SI_i \tag{18}$$

Wang *et al.* (2017) described the WQI into five categories as listed in Table 4. It can be noted from Table 4 that 16.58% of samples exhibited excellent water quality, 29.34% good water quality and 13.26% of water samples fell under unsuitable for drinking purposes.

Table 4 | WQI scale, water type and % of water samples in the study area (Aouiti *et al.* 2021)

Sr. No.	WQI scale	Quality of groundwater	Number of samples	% of samples
1	<50	Excellent	65	16.58
2	50–100	Good	115	29.34
3	100–200	Poor	119	30.36
4	200–300	Very Poor	41	10.46
5	>300	Not suitable for drinking	52	13.26

The values of EWQI and WQI obtained for predicting the groundwater quality are shown in Tables 2 and 4. The perusal of Tables 2 and 4 shows that 50% of the samples fall under 0–100 for EWQI, while 45.92% of the samples fall under 0–100 in WQI. The comparison of EWQI and WQI shows that most of the samples of EWQI are more effective than WQI. As compared with past studies undertaken by Kumari & Rai (2020) and Panghal *et al.* (2021), the results obtained in this study are superior thereby, suggesting the importance of successful application of the three machine learning models.

3. PARAMETER OPTIMIZATION OF DIFFERENT MACHINE LEARNING ALGORITHMS

The use of algorithms XGBoost, DNN and GBM needs setting of several user-defined parameters. For all three algorithms used in the present study, H2o Auto-ML software (H2O.ai. 2020) was applied. To obtain the optimal value of various hyper parameters of different algorithms, Auto-ML uses 5 -fold cross validation and grid search method. During training phase, the DNN produces a ranked list of different input variables using Geodon's method (Gedeon 1997) whereas GBM and XGBoost use a gain value to rank various input attributes as discussed in (Xu *et al.* 2014; Chen & Guestrin 2016). The optimal values of various parameters obtained in predicting EQWI and WQI are provided in Table 5.

Table 5 | Optimal value of user-defined parameters DNN, XGBoost and GBM

Name of algorithm	User-defined parameters for EWQI	User-defined parameters for WQI
DNN	One Hidden layer with 200 nodes, Epochs = 4211 Activation Rectifier with Dropout = 0.2, initial weight distribution = Uniform Adaptive, Batch size = 1, distribution = gaussian	One Hidden layer with 50 nodes, Epochs = 8380 Activation Rectifier with Dropout = 0.2, initial weight distribution = Uniform Adaptive, Batch size = 1, distribution = gaussian
XGBoost	Distribution = gaussian, booster = gbtree, ntree = 54, maxbins = 256, max depth = 5, eta = 0.3 and lambda (λ) = 0.01	Distribution = gaussian, booster = gbtree, ntree = 43, maxbins = 256, max depth = 20, eta = 0.3 and lambda (λ) = 0.1
GBM	Distribution = gaussian, ntree = 72, nbins = 20, max depth = 9	Distribution = gaussian, ntree = 75, nbins = 20, max depth = 10

3.1. Performance indices

The performance of these three algorithms (DNN, XGBoost and GBM) used in this study is predicted after the desired training and testing phases are completed. The calculated output values are estimated using the following measures of goodness-of-fit: Correlation Coefficient (CC), Root Mean Square Error (RMSE), Nash–Sutcliffe efficiency (NSE) and Index of agreement (d). The CC range is varied from −1 to 1. If the CC value is approaching 1 and the RMSE value is approaching 0 then the accuracy of the model is high. If the CC value is approaching 0 and the RMSE value is approaching 1 then the accuracy of model is low. Nash & Sutcliffe (1970) initially proposed Nash–Sutcliffe efficiency (NSE) and commonly used statistic index. The range of NSE is varied from 0 to +1 and is also called the efficiency index (E_f). Willmott (1981) first used the Index of agreement (d), which varied between 0 and 1. The values of CC, RMSE, NSE, and d are computed by using the Equations (19)–(22) as given below:

$$\mathrm{CC} = \frac{\sum_{i=1}^{N}(X_0 i - \bar{X}_0)(X_p i - \bar{X}_p)}{\sqrt{\sum_{i=1}^{N}(X_0 i - \bar{X}_0)^2 \sum_{i=1}^{N}(X_p i - \bar{X}_p)^2}} \tag{19}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum (X_p i - X_0 i)^2}{N}} \tag{20}$$

$$\mathrm{NSE} = 1 - \left[\frac{\sum_{i=1}^{N}(X_p - X_o)^2}{\sum_{i=1}^{N}(X_o - \bar{X}_o)^2}\right] \tag{21}$$

$$\mathrm{d} = 1 - \left[\frac{\sum_{i=1}^{N}(X_p - X_o)^2}{\sum_{i=1}^{N}(|X_p - X_o| + |X_o - \bar{X}_o|)^2}\right] \tag{22}$$

where X_0, X_p and $\bar{X}_o$ are the observed and predicted and mean of observed values respectively.

3.2. Uncertainty and reliability analysis

Saberi-Movahed *et al.* (2020) carried out the uncertainty (U_{95}) and reliability analysis, for evaluating the overall consistency of the model used in predictions for longitudinal dispersion coefficients in water pipelines. The main purpose of uncertainty analysis is to adjust the expected range within which the true value of the experimental result lies and is obtained from Equation (23) and the reliability analysis is determined by using Equations (24) and (25), which are given below:

$$U_{95} = \left(\frac{1.96}{N}\right)\sqrt{\sum_{i=1}^{N}(X_0 i - \bar{X}_0)^2 + \sum_{i=1}^{N}(X_0 i - X_p i)^2} \tag{23}$$

$$\text{Reliability} = \left(\frac{100\%}{N}\right)\sum_{i=0}^{N} K_i \tag{24}$$

$$\mathrm{RAE} = \left|\frac{X_o - X_p}{X_o}\right| \tag{25}$$

To obtain the K_i value, first relative average error is determined from Equation (25) and next, if the value of RAE $\leq \Delta$ then $K_i = 1$, otherwise $K_i = 0$.

Where, Δ is the threshold value of the water quality parameter and as per Chinese standards the optimum value of Δ is 0.2, or 20%. K_i is determined as the number of times the value of RAE$\leq\Delta$. The same equations have been applied in determining the water quality of a data set for estimating EWQI and WQI.

4. ANALYSIS OF RESULTS AND DISCUSSION

Table 6 provides results obtained in terms of CC, RMSE, NSE and d values with both training and test datasets for predicting the groundwater quality using EQWI and WQI. A comparison of these values suggests the slightly improved performance of EWQI by optimized DNN (CC = 0.989, RMSE = 0.037, NSE = 0.995 and d = 0.999) in comparison to XGBoost (CC = 0.976, RMSE = 0.056, NSE = 0.994 and d = 0.999) and GBM (CC = 0.959, RMSE = 0.074, NSE = 0.992 and d = 0.998). However, the performance of WQI using DNN (CC = 0.975, RMSE = 0.055, NSE = 0.991 and d = 0.998) in comparison to XGBoost (CC = 0.944, RMSE = 0.082, NSE = 0.986 and d = 0.998) and GBM (CC = 0.968, RMSE = 0.062, NSE = 0.980 and d = 0.998). The overall results also suggest slightly improved performance by GBM in comparison to XGBoost while using WQI as a groundwater quality index whereas XGBoost achieves better results with EWQI.

Table 6 | Modeling performance for EWQI and WQI with a training and test dataset

		CC		RMSE		NSE		d	
Index	**Modeling approach**	**Training**	**Testing**	**Training**	**Testing**	**Training**	**Testing**	**Training**	**Testing**
EWQI	DNN	**0.998**	**0.989**	**0.012**	**0.037**	**0.989**	**0.995**	**0.997**	**0.999**
	XGBoost	0.998	0.976	0.016	0.056	0.935	0.994	0.982	0.999
	GBM	0.998	0.959	0.015	0.074	0.922	0.992	0.978	0.998
WQI	DNN	**0.999**	**0.975**	**0.018**	**0.055**	**0.965**	**0.991**	**0.990**	**0.998**
	XGBoost	0.996	0.944	0.048	0.082	0.943	0.986	0.982	0.998
	GBM	0.996	0.968	0.020	0.062	0.949	0.980	0.986	0.998

Table 7 shows the performance results obtained in terms of uncertainty (U_{95}) and reliability analysis for predicting the groundwater quality using EWQI and WQI in both training and testing stages. From this table, it is evident that the performance of EWQI using DNN achieved the lowest value of U_{95} (21.60) when compared with XGBoost (U_{95} = 22.31) and GBM (U_{95} = 22.17) in the training stage. While in the testing stage, the value of U_{95} of EWQI optimized by DNN is more accurate (U_{95} = 19.78) in comparison to XGBoost (U_{95} = 20.91) and GBM (U_{95} = 20.22). Additionally, prediction of EWQI using DNN is more reliable (98.59) in comparison to XGBoost (96.46) and GBM (97.17) in the training stage. While in the testing stage, the performance of EWQI using DNN has a higher level of reliability (98.16) when compared with XGBoost (97.08) and GBM (97.16). To conclude, in terms of uncertainty (U_{95}) and reliability, DNN performs better than the other two models for training as well as testing. The same trend has been observed for estimation of WQI in terms of uncertainty and reliability under the training and testing category. However, when two indices, EWQI and WQI, are compared, the prediction of EWQI is superior to WQI in the three algorithms used in this article.

Table 7 | Performance results for the uncertainty and reliability analysis of DNN, XGBoost and GBM algorithms

	EWQI		WQI	
Algorithm	**Uncertainty (U_{95})**	**Reliability (%)**	**Uncertainty (U_{95})**	**Reliability (%)**
Training				
DNN	**21.60**	**98.59**	**21.71**	**97.82**
XGBoost	22.31	96.46	21.88	96.82
GBM	22.17	97.17	21.74	97.23
Testing				
DNN	**19.78**	**98.16**	**20.60**	**97.08**
XGBoost	20.91	97.08	20.62	96.16
GBM	20.22	97.16	20.61	96.33

Further, Figures 2 and 3 provide plots between predicted and actual values of EWQI and WQI using DNN, XGBoost, and GBM respectively. It can be observed from Figure 2 that generally, most of the predicted values

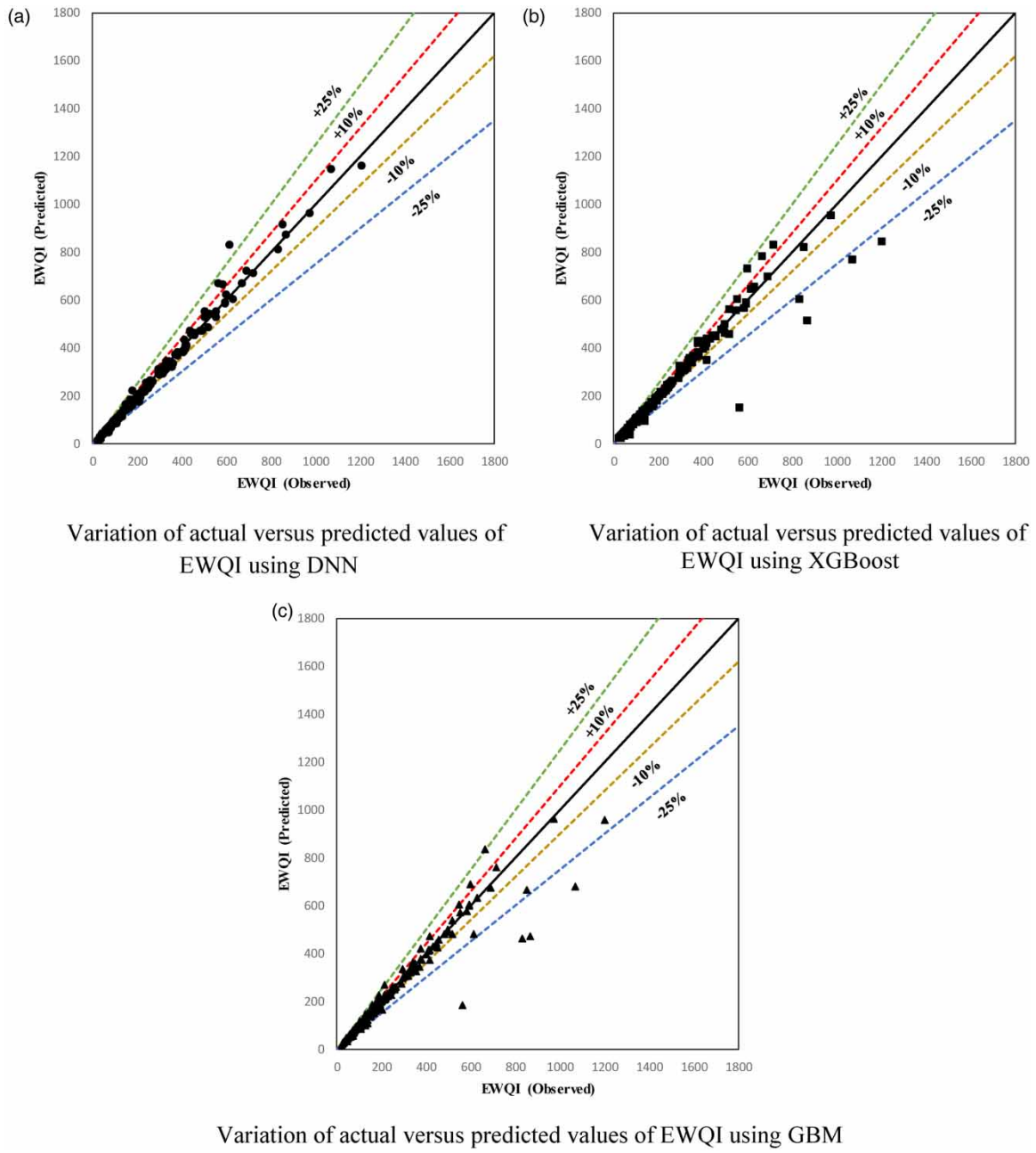

Figure 2 | Variation of actual versus predicted values of EWQI using (a) DNN (b) XGBoost and (c) GBM algorithms.

for EWQI as the groundwater quality index fall near the line of perfect agreement (i.e. a line at 45°). Four more lines in the range of ±25% and ±10% error between the actual and predicted values of EWQI and WQI are also plotted in the graphs. Figure 2(a) shows that most of the value predicted by DNN of EWQI were well ±10% error line from the line of perfect agreement. It can be inferred from Figure 2(b) that some predicted values by XGBoost of EWQI are also lying away from the ±25% error line. Figure 2(c) shows that some predicted values by GBM are lying away from ±10% error line and ±25% error line. Figure 3(a) demonstrates that most of the value predicted for WQI by DNN were well ±10% error line from the line of perfect agreement. Comparison of Figures 2(a)–2(c) and 3(a)–3(c) also suggests that the predicted EWQI and WQI values by DNN are in good agreement with the actual EWQI and WQI respectively, suggesting its improved performance in comparison to both XGBoost and GBM in the present study.

In this study the relative importance of each hydrochemical parameter has been compared with each algorithm, so that the variability in the dependent variable can be explained by the independent variables.

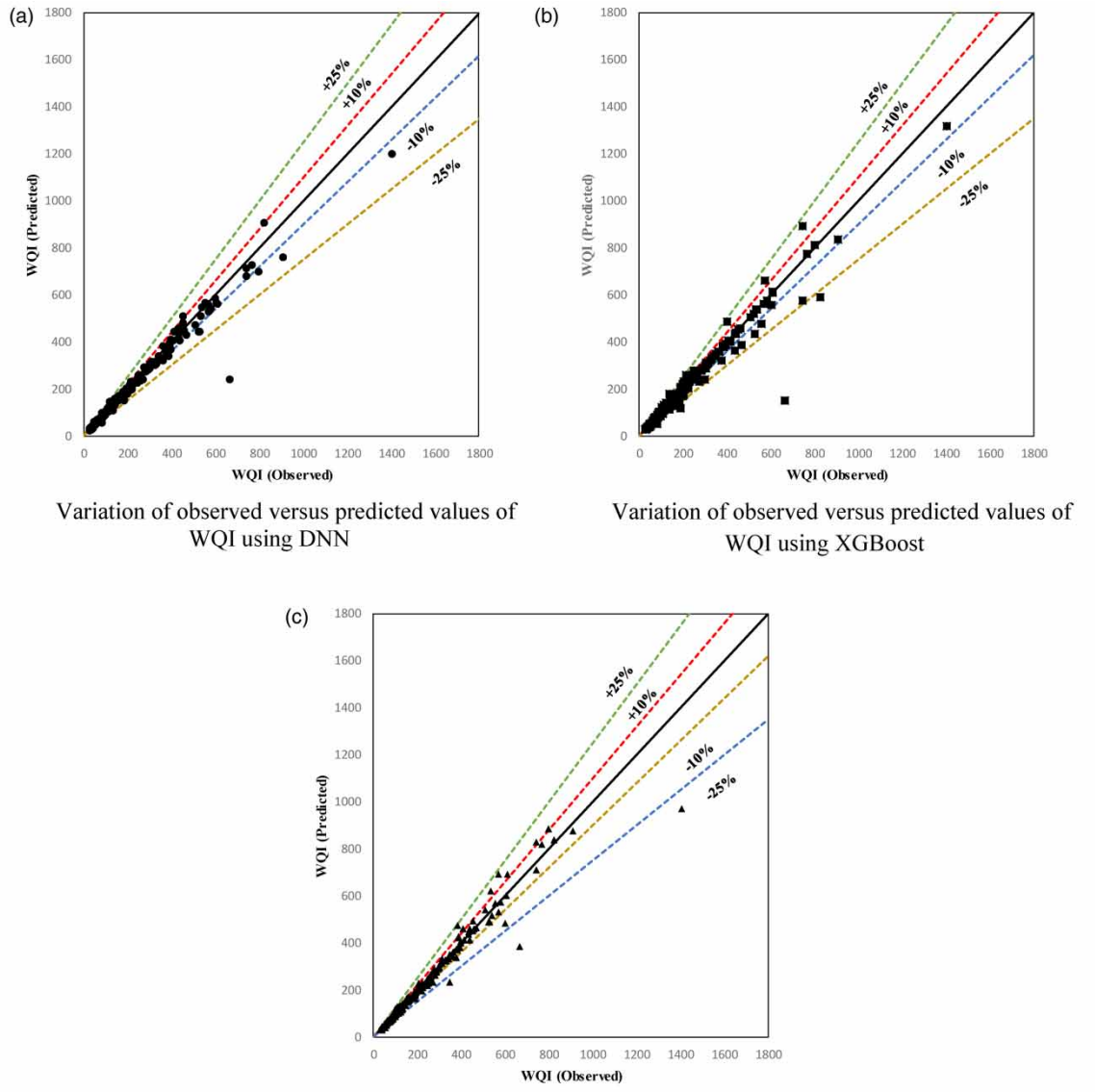

Figure 3 | Variation of observed versus predicted values of WQI using (a) DNN (b) XGBoost and (c) GBM algorithms.

Figure 4(a)–4(f) depicts the relative importance of different independent chemical variables calculated by using DNN, XGBoost and GBM algorithms. Figure 4(a)–4(c) represents the relative importance of variables for modeling EWQI and Figure 4(d)–4(f) represents the relative importance of variables for modeling WQI. From Figure 4(a)–4(f), it can be observed that the EC is the most significant variable having the highest relative importance to predict the EWQI and the WQI by using the three algorithms DNN, XGBoost and GBM. The pH has the lowest relative importance in prediction of EWQI (DNN = 0.2212, XGBoost = 0.0013 and GBM = 0.0008) and WQI (DNN = 0.1476, XGBOOST = 0.04781 and GBM = 0.0016). The most interesting fact is that Mg attains moderate relative importance for the prediction of EWQI (DNN = 0.4239 and GBM = 0.0033) and WQI (DNN = 0.3645 and GBM = 0.0314) but XGBoost shows a higher relative importance in WQI (0.5545) and lower relative importance in EWQI (0.0021) of this variable. Similarly, TH has moderate relative importance for the prediction of EWQI (DNN = 0.4145 and GBM = 0.0032) and WQI (DNN = 0.3534 and GBM = 0.0213) but XGBoost shows a higher relative importance in WQI (0.873) and lower relative importance in EWQI (0.0481) of this variable. However, the perusal of Figure 4(a)–4(f) shows that there is no definite trend of variable importance of hydrochemical parameters in estimating EWQI and WQI using the three algorithms; that is, DNN, XGBoost and GBM, on this data set.

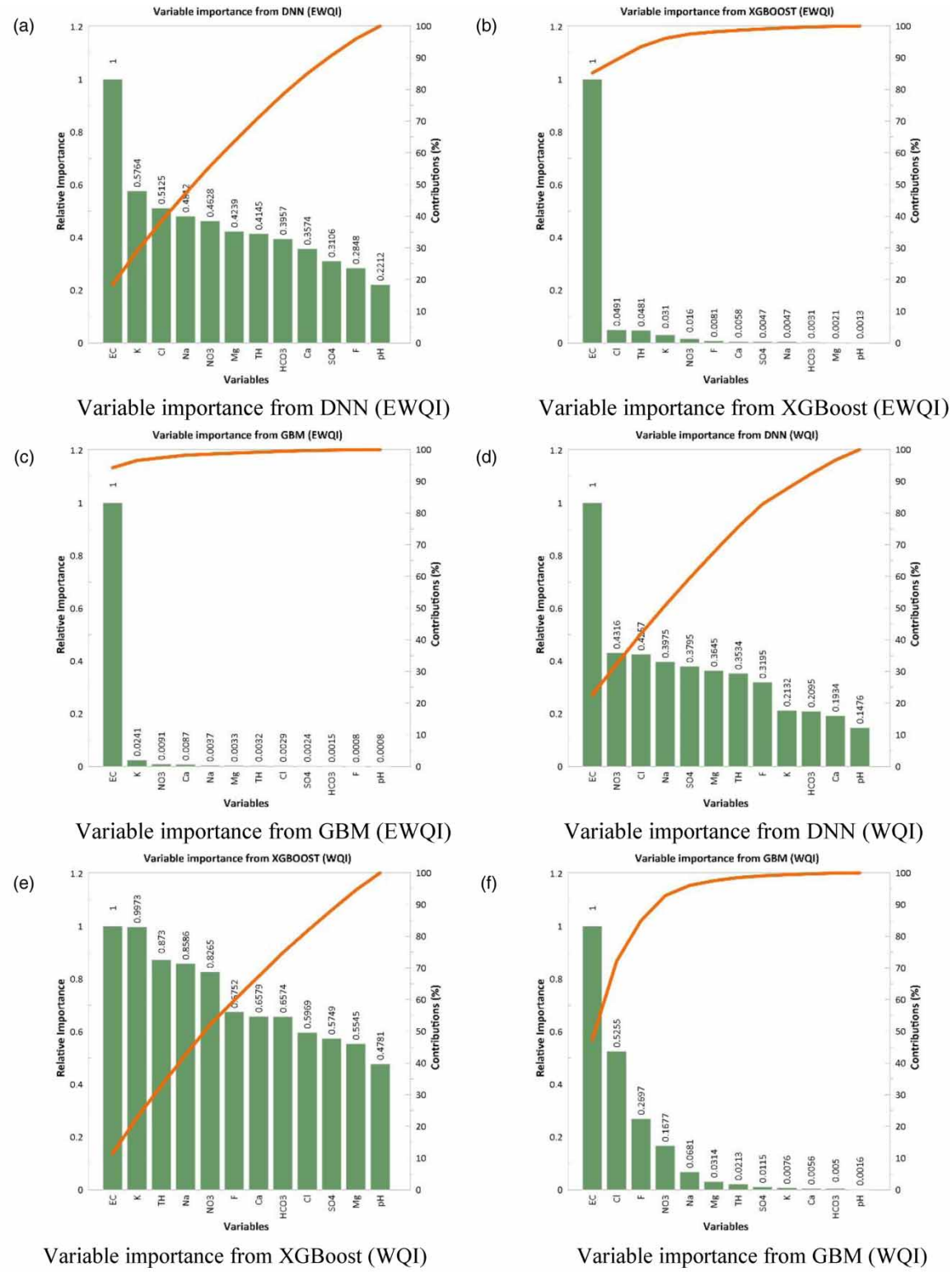

Figure 4 | Relative importance of input variables from DNN, XGBoost and GBM algorithms.

5. CONCLUSIONS

In the present work, a study has been taken up to assess the performance of DNN, XGBoost, and GBM algorithms for predicting the groundwater quality based on two indices; that is, EWQI and WQI. The comparison of three techniques has been made on the basis of four goodness-of-fit indices, namely: CC, RMSE, NSE and d. The conclusions drawn from this work are as given below:

1. Three different machine learning models were used to predict EWQI; that is, DNN, XGBoost, and GBM. Out of the three algorithms it is observed that the DNN algorithm performs best in terms of the training as well as testing dataset. For the testing dataset, performance indices have the following values: (CC = 0.989, RMSE = 0.037, NSE = 0.995 and d = 0.999) and for training (CC = 0.998, RMSE = 0.012, NSE = 0.989 and d = 0.997).
2. Also, in terms of predictions of WQI, DNN performed best in terms of the training and testing dataset. For the testing dataset, performance indices have the following values: (CC = 0.975, RMSE = 0.055, NSE = 0.991 and d = 0.998) and for training (CC = 0.999, RMSE = 0.018, NSE = 0.965 and d = 0.990).
3. After comparison of all the three algorithms, it was found that DNN based modeling performs slightly better than XGBoost and GBM in estimating both EWQI and WQI in the training as well as testing data set.
4. After observing the values of various relative importance of variables as given by the three models, the variable 'EC' was given the most significance in predicting the dependent variable. However, the parameter 'pH' has been considered as the least significant parameter in prediction using the three machine learning models in the present study.
5. The uncertainty (U_{95}) and reliability analysis of the data set also revealed that the DNN performs better than the other two algorithms (XGBoost and GBM) for training as well as testing in prediction of EWQI and WQI.
6. The promising results observed in the study suggest that DNN can be used to predict various other bio-chemical and physio-chemical properties of groundwater consumed for drinking purposes.
7. The findings of this study can be extended further by examining the performance of the DNN model as compared to other machine learning models, considering different possible hydrochemical input parameters.
8. Furthermore, this study aims to be a mere suggestion to the research community about the application of machine learning models in prediction of ground water quality indices, as it is much needed for economically weaker regions, where not much equipment and resources are available.

ACKNOWLEDGEMENTS

First author (Hemant Raheja) is grateful to MHRD, GOI for financially supporting the present work for a Ph.D. scholarship grant (2K19/NITK/PHD/61900011) and to the site (https://indiawris.gov.in) for the data used in this study.

CONFLICT OF INTEREST

Authors claim there exists no conflict of interest in any form concerning the work done within the present manuscript.

DATA AVAILABILITY STATEMENT

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

REFERENCES

Adimalla, N. & Qian, H. 2019 Groundwater quality evaluation using water quality index (WQI) for drinking purposes and human health risk (HHR) assessment in an agricultural region of Nanganur, South India. *Ecotoxicology and Environmental Safety* **176**(126), 153–161. https://doi.org/10.1016/j.ecoenv.2019.03.066.

Adimalla, N., Dhakate, R., Kasarla, A. & Taloor, A. K. 2020a Appraisal of groundwater quality for drinking and irrigation purposes in Central Telangana, India. *Groundwater for Sustainable Development* **10**(126), 100334. https://doi.org/10.1016/j.gsd.2020.100334.

Adimalla, N., Qian, H. & Li, P. 2020b Entropy water quality index and probabilistic health risk assessment from geochemistry of groundwaters in hard rock terrain of Nanganur County, South India. *Chemie Der Erde* **80**(4), 125544. https://doi.org/10.1016/j.chemer.2019.125544.

Aldhyani, T. H. H., Al-Yaari, M., Alkahtani, H. & Maashi, M. 2020 Water quality prediction using artificial intelligence algorithms. *Applied Bionics and Biomechanics* **2020**, 6659314. https://doi.org/10.1155/2020/6659314.

Aly, A. A., Al-Omran, A. M. & Alharby, M. M. 2015 The water quality index and hydrochemical characterization of groundwater resources in Hafar Albatin, Saudi Arabia. *Arabian Journal of Geosciences* **8**(6), 4177–4190. https://doi.org/10.1007/s12517-014-1463-2.

Aouiti, S., Hamzaoui Azaza, F., El Melki, F., Hamdi, M., Celico, F. & Zammouri, M. 2021 Groundwater quality assessment for different uses using various water quality indices in semi-arid region of central Tunisia. *Environmental Science and Pollution Research* **28**(34), 46669–46691. https://doi.org/10.1007/s11356-020-11149-5.

BIS 2015 *Indian Standard Drinking Water–Specification (Second Revision)*. Bureau of Indian Standards (BIS), IS 10500, New Delhi, pp. 2–6.

Chen, T. & Guestrin, C. 2016 XGBoost A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **42**(8), 785–794. https://doi.org/10.1145/2939672.2939785.

Chung, S. Y., Venkatramanan, S., Kim, T. H., Kim, D. S. & Ramkumar, T. 2015 Influence of hydrogeochemical processes and assessment of suitability for groundwater uses in Busan City, Korea. *Environment, Development and Sustainability* **17**(3), 423–441. https://doi.org/10.1007/s10668-014-9552-7.

Dao, V., Urban, W. & Hazra, S. B. 2020 Introducing the modification of Canadian Water Quality Index. *Groundwater for Sustainable Development* **11**, 100457. https://doi.org/10.1016/j.gsd.2020.100457.

Di, Z., Chang, M. & Guo, P. 2019 Water quality evaluation of the Yangtze River in China using machine learning techniques and data monitoring on different time scales. *Water (Switzerland)* **11**(2), 339. https://doi.org/10.3390/w11020339.

Dick, K., Russell, L., Souley Dosso, Y., Kwamena, F. & Green, J. R. 2019 Deep learning for critical infrastructure resilience. *Journal of Infrastructure Systems* **25**(2), 05019003. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000477.

El Bilali, A., Abdeslam, T., Mazigh, N. & Moukhliss, M. 2020 Prediction of chemical water quality used for drinking purposes based on artificial neural networks. *Moroccan Journal of Chemistry* **3**, 665–672. https://doi.org/https://doi.org/10.48317/IMIST.PRSM/morjchem-v8i3.19786.

El Bilali, A., Taleb, A. & Brouziyne, Y. 2021 Groundwater quality forecasting using machine learning algorithms for irrigation purposes. *Agricultural Water Management* **245**(July), 106625. https://doi.org/10.1016/j.agwat.2020.106625.

Freund, Y. 1995 Boosting a weak learning algorithm by majority. In: *Information and Computation* **121**(2), 256–285. https://doi.org/10.1006/inco.1995.1136.

Friedman, J. H. 2001 Greedy function approximation: a gradient boosting machine. *Annals of Statistics* **29**(5), 1189–1232. https://doi.10.1214/aos/1013203451.

Gaikwad, S. K., Kadam, A. K., Ramgir, R. R., Kashikar, A. S., Wagh, V. M., Kandekar, A. M., Gaikwad, S. P., Madale, R. B., Pawar, N. J. & Kamble, K. D. 2020 Assessment of the groundwater geochemistry from a part of west coast of India using statistical methods and water quality index. *HydroResearch* **3**, 48–60. https://doi.org/10.1016/j.hydres.2020.04.001.

Gedeon, T. D. 1997 Data mining of inputs: analysing magnitude and functional measures. *International Journal of Neural Systems* **8**(2), 209–218. https://doi.org/10.1142/S0129065797000227.

Goodfellow, I., Bengio, Y. & Courville, A. 2016 *Deep Learning*. MIT Press, Cambridge, USA.

Granata, F., Papirio, S., Esposito, G., Gargano, R. & de Marinis, G. 2017 Machine learning algorithms for the forecasting of wastewater quality indicators. *Water (Switzerland)* **9**(2), 1–12. https://doi.org/10.3390/w9020105.

H2O.ai 2020 *H2O: Scalable Machine Learning Platform. Version 3.30.0.6*. Available from: https://github.com/h2oai/h2o-3 (3.30.0.6).

Hadi, S. J., Abba, S. I., Sammen, S. S. H., Salih, S. Q., Al-Ansari, N. & Mundher Yaseen, Z. 2019 Non-linear input variable selection approach integrated with non-tuned data intelligence model for streamflow pattern simulation. *IEEE Access* **7**, 141533–141548. https://doi.org/10.1109/ACCESS.2019.2943515.

Haghiabi, A. H., Nasrolahi, A. H. & Parsaie, A. 2018 Water quality prediction using machine learning methods. *Water Quality Research Journal* **53**(1), 3–13. https://doi.org/10.2166/wqrj.2018.025.

Horton, R. K. 1965 An index number system for rating water quality. *Journal of Water Pollution Control Federation* **37**(3), 300–306.

Ibrahem Ahmed Osman, A., Najah Ahmed, A., Chow, M. F., Feng Huang, Y. & El-Shafie, A. 2021 Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. *Ain Shams Engineering Journal* **12**, 1545–1556. https://doi.org/10.1016/j.asej.2020.11.011.

Kaushik, A., Kumar, K., Sharma, I. S. & Sharma, H. R. 2004 Groundwater quality assessment in different land-use areas of Faridabad and Rohtak cities of Haryana using deviation index. *Journal of Environmental Biology* **25**(2), 173–180.

Kingma, D. P. & Ba, J. 2015 Adam: a method for stochastic optimization. In: *3rd International Conference on Learning Representations*, May 7–9, 2015, San Diego. https://Arxiv.Org/Abs/1412.6980 (accessed June 15, 2019).

Kumar, V. S., Amarender, B., Dhakate, R., Sankaran, S. & Raj Kumar, K. 2016 Assessment of groundwater quality for drinking and irrigation use in shallow hard rock aquifer of Pudunagaram, Palakkad District Kerala. *Applied Water Science* **6**(2), 149–167. https://doi.org/10.1007/s13201-014-0214-6.

Kumar, S. S. & Abraham, D. M. 2019 A deep learning based automated structural defect detection system for sewer pipelines. In: *Computing in Civil Engineering 2019: Smart Cities, Sustainability, and Resilience – Selected Papers From the ASCE International Conference on Computing in Civil Engineering 2019*. https://doi.org/10.1061/9780784482445.029.

Kumar, P. J. S. & Augustine, C. M. 2021 Entropy-weighted water quality index (EWQI) modeling of groundwater quality and spatial mapping in Uppar Odai Sub-Basin, South India. *Modeling Earth Systems and Environment 0123456789*. https://doi.org/10.1007/s40808-021-01132-5.

Kumari, M. & Rai, S. C. 2020 Hydrogeochemical evaluation of groundwater quality for drinking and irrigation purposes using water quality index in Semi Arid Region of India. *Journal of the Geological Society of India* **95**(2), 159–168. https://doi.org/10.1007/s12594-020-1405-4.

Lu, H. & Ma, X. 2020 Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere* **249**, 126169. https://doi.org/10.1016/j.chemosphere.2020.126169.

Lumb, A., Sharma, T. C., Bibeault, J.-F. & Klawunn, P. 2011 A comparative study of USA and Canadian water quality index models. *Water Quality, Exposure and Health* **3**, 203–216. https://doi.org/10.1007/s12403-011-0056-5.

Maghrebi, M., Noori, R., Partani, S., Araghi, A., Barati, R., Farnoush, H. & Torabi Haghighi, A. 2021 Iran's groundwater hydrochemistry. *Earth and Space Science* **8**(8), 1–18. https://doi.org/10.1029/2021EA001793.

Meyers, G., Kapelan, Z. & Keedwell, E. 2017 Short-term forecasting of turbidity in trunk main networks. *Water Research* **124**, 67–76. https://doi.org/10.1016/j.watres.2017.07.035.

Mohamed, I., Othman, F., Ibrahim, A. I. N., Alaa-Eldin, M. E. & Yunus, R. M. 2015 Assessment of water quality parameters using multivariate analysis for Klang River basin, Malaysia. *Environmental Monitoring and Assessment* **187**, 4182. https://doi.org/10.1007/s10661-014-4182-y.

Nair, V. & Hinton, G. E. 2010 Rectified linear units improve Restricted Boltzmann machines. In: *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*.

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), 380. https://doi.org/10.1007/s10661-019-7446-8.

Najafzadeh, M. & Niazmardi, S. 2021 A novel multiple-kernel support vector regression algorithm for estimation of water quality parameters. *Natural Resources Research* **30**(5), 3761–3775. https://doi.org/10.1007/s11053-021-09895-5.

Najafzadeh, M., Ghaemi, A. & Emamgholizadeh, S. 2019 Prediction of water quality parameters using evolutionary computing-based formulations. *International Journal of Environmental Science and Technology* **16**(10), 6377–6396. https://doi.org/10.1007/s13762-018-2049-4.

Najafzadeh, M., Homaei, F. & Mohamadi, S. 2021 Reliability evaluation of groundwater quality index using data-driven models. *Environmental Science and Pollution Research*. https://doi.org/10.1007/s11356-021-16158-6.

Najah Ahmed, A., Binti Othman, F., Abdulmohsin Afan, H., Khaleel Ibrahim, R., Ming Fai, C., Shabbir Hossain, M., Ehteram, M. & Elshafie, A. 2019 Machine learning methods for better water quality prediction. *Journal of Hydrology* **578**(May), 124084. https://doi.org/10.1016/j.jhydrol.2019.124084.

Nash, J. E. & Sutcliffe, J. V. 1970 River flow forecasting through conceptual models part I – a discussion of principles. *Journal of Hydrology* **10**(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

Nayan, A.-A., Kibria, M. G., Rahman, M. O. & Saha, J. 2021 *River Water Quality Analysis and Prediction Using GBM*. *November*, 219–224. https://doi.org/10.1109/icaict51780.2020.9333492.

Pal, M. 2019 Deep neural network based pier scour modeling. *ISH Journal of Hydraulic Engineering* 1–6. https://doi.org/10.1080/09715010.2019.1679673.

Panghal, V., Sharma, P., Mona, S. & Bhateria, R. 2021 Determining groundwater quality using indices and multivariate statistical techniques: a study of Tosham block, Haryana, India. *Environmental Geochemistry and Health* **9**. https://doi.org/10.1007/s10653-021-01120-9.

Saberi-Movahed, F., Najafzadeh, M. & Mehrpooya, A. 2020 Receiving more accurate predictions for longitudinal dispersion coefficients in water pipelines: training group method of data handling using extreme learning machine conceptions. *Water Resources Management* **34**(2), 529–561. https://doi.org/10.1007/s11269-019-02463-w.

Shannon, C. E. 1948 A mathematical theory of communication. *Bell System Technical Journal* **27**(4), 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x.

Su, F., Wu, J. & He, S. 2019 Set pair analysis-Markov chain model for groundwater quality assessment and prediction: a case study of Xi'an city, China. *Human and Ecological Risk Assessment*. https://doi.org/10.1080/10807039.2019.1568860.

Tyagi, S., Sharma, B., Singh, P. & Dobhal, R. 2020 Water quality assessment in terms of water quality index. *American Journal of Water Resources* **1**(3), 34–38. https://doi.org/10.12691/ajwr-1-3-3.

Vasanthavigar, M., Srinivasamoorthy, K., Vijayaragavan, K., Rajiv Ganthi, R., Chidambaram, S., Anandhan, P., Manivannan, R. & Vasudevan, S. 2010 Application of water quality index for groundwater quality assessment: Thirumanimuttar sub-basin, Tamilnadu, India. *Environmental Monitoring and Assessment*. https://doi.org/10.1007/s10661-009-1302-1.

Wang, X., Zhang, F. & Ding, J. 2017 Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China. *Scientific Reports* **7**(1), 1–18. https://doi.org/10.1038/s41598-017-12853-y.

Wang, D., Wu, J., Wang, Y. & Ji, Y. 2020 Finding high-quality groundwater resources to reduce the hydatidosis incidence in the Shiqu County of Sichuan Province, China: analysis, assessment, and management. *Exposure and Health* **12**(2), 307–322. https://doi.org/10.1007/s12403-019-00314-y.

Willmott, C. J. 1981 On the validation of models. *Physical Geography* **2**(2), 184–194. https://doi.org/10.1080/02723646.1981.10642213.

Wu, Z., Wang, X., Chen, Y., Cai, Y. & Deng, J. 2018 Assessing river water quality using water quality index in Lake Taihu Basin, China. *Science of the Total Environment* **612**, 914–922. https://doi.org/10.1016/j.scitotenv.2017.08.293.

Xu, Z., Huang, G., Weinberger, K. Q. & Zheng, A. X. 2014 Gradient boosted feature selection categories and subject descriptors. In: *20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 522–531.

Yu, X., Wang, Y., Wu, L., Chen, G., Wang, L. & Qin, H. 2020 Comparison of support vector regression and extreme gradient boosting for decomposition-based data-driven 10-day streamflow forecasting. *Journal of Hydrology* **582**, 124293. https://doi.org/10.1016/j.jhydrol.2019.124293.

Zahedi, S. 2017 Modification of expected conflicts between Drinking Water Quality Index and Irrigation Water Quality Index in water quality ranking of shared extraction wells using Multi Criteria Decision Making techniques. *Ecological Indicators*. https://doi.org/10.1016/j.ecolind.2017.08.017.

Zhang, W., Wu, C., Zhong, H., Li, Y. & Wang, L. 2021 Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geoscience Frontiers* **12**(1), 469–477. https://doi.org/10.1016/j.gsf.2020.03.007.

First received 25 September 2021; accepted in revised form 16 November 2021. Available online 1 December 2021