Environmental Science and Pollution Research
https://doi.org/10.1007/s11356-023-25886-w

RESEARCH ARTICLE

A novel hybrid random forest linear model approach for forecasting groundwater fluoride contamination

Mouigni Baraka Nafouanti[1] · Junxia Li[1,2] · Edwin E. Nyakilla[3] · Grant Charles Mwakipunda[3] · Alvin Mulashani[4]

Received: 14 November 2022 / Accepted: 7 February 2023

Abstract

Groundwater quality in the Datong basin is threatened by high fluoride contamination. Laboratory analysis is a standard method for estimating groundwater quality parameters, which is expensive and time-consuming. Therefore, this paper proposes a hybrid random forest linear model (HRFLM) as a novel approach for estimating groundwater fluoride contamination. Light gradient boosting (LightGBM), random forest (RF), and extreme gradient boosting (Xgboost) were also employed in comparison with HRFLM for predicting fluoride contamination in groundwater. 202 groundwater samples were collected to draw up the performance capability of several models in forecasting subsurface water fluoride contamination. The performance of the models was assessed utilizing the receiver operating characteristic (ROC) area under the curve (AUC) and the confusion matrix (CM). The CM results reveal that with nine predictor variables, the hybrid HRFLM achieved an accuracy of 95%, outperforming the Xgboost, LightGBM, and RF models, which attained 88%, 88%, and 85%, respectively. Likewise, the AUC results of the hybrid HRFLM show high performance with an AUC of 0.98 compared to Xgboost, LightGBM, and RF, which achieved an AUC of 0.95, 0.90, and 0.88, respectively. The study demonstrates that the HRFLM can be applied as an advanced approach for groundwater fluoride contamination prediction in the Datong basin and could be adopted in various areas facing a similar challenge.

Keywords Hybrid random forest linear model · Fluoride · Light gradient boosting machine · Groundwater · Extreme gradient boosting

Introduction

Groundwater is the most important water source in the world (Brindha and Elango, 2011). It plays a principal role in economic growth in several countries like Denmark, Austria, and China (Khosravi et al., 2020; Manap et al., 2013). In the Datong basin, groundwater serves in agriculture and industrial activities and constitutes the primary drinking water (He et al., 2021; Nafouanti et al., 2021b). However, the quality of this water resource is vulnerable and continuously deteriorating due to the introduction of fluoride as a direct consequence of increasing numerous natural and anthropogenic activities (Li, 2001; Orban et al., 2010; Xie et al., 2011). Hence, monitoring the subsurface water quality is essential for the protection and sustainability of the groundwater source.

The groundwater fluoride in the Datong basin is derived from the release of fluorite and biotite under the arid and semiarid climate (Guo et al., 2007; Li et al., 2020; Mamatha and Rao, 2010). Furthermore, human activities directly

Responsible Editor: Marcus Schulz

✉ Mouigni Baraka Nafouanti
mouignibarakanafouanti@gmail.com

[1] State Key Laboratory of Biogeology and Environmental Geology, China University of Geosciences, Wuhan 430074, China

[2] China Laboratory of Basin Hydrology and Wetland Eco-restoration, China University of Geosciences, Wuhan 430074, China

[3] Department of Petroleum Engineering, Faculty of Earth Resources, China University of Geosciences, Wuhan 430074, China

[4] Department of Geosciences and Mining Technology, College of Engineering and Technology, Mbeya University of Science and Technology, Box 131, Mbeya, Tanzania

Published online: 17 February 2023

contribute to the groundwater fluoride concentration, including fertilizers (e.g., phosphatic fertilizer, nitrogen phosphorus potassium), irrigation, sewage, and sludge (Ramanaiah et al., 2006). Assigned by the World Health Organization, an optimum range of 0.5 to 1.5 mg/L of fluoride maintains the protection of teeth and bone growth in the human body (Al-Mohair et al., 2015; Su et al., 2013a). However, a high quantity of fluoride in groundwater is a source of several complications, including bone deformation and dental change, which have been reported in Africa, Pakistan, India, and China (Rafique et al., 2008; Tripathy et al., 2006). In this case, modeling can help identify groundwater fluoride quantity, and an accurate groundwater estimation is also essential.

The advancement of artificial intelligence (AI) has brought vast technology to study and estimate groundwater contamination (Chang et al., 2020; Feng et al., 2020; Gupta et al., 2021; Vesselinov et al., 2018; Wang and Wang, 2020). Machine learning (ML) approaches have been considered a crucial concept in hydrology research following their successful deployment in anticipating groundwater recently, as they can resolve complex problems (Huang et al., 2020c; Jain et al., 1996). An artificial neural network (ANN) is the frequently employed ML algorithm to estimate groundwater contamination, and it has been applied in Canada to predict groundwater levels (Adamowski and Chan, 2011). Likewise, the extreme learning machine (ELM), a feed-forward neural network, was employed to forecast groundwater fluoride in the Maku area (Barzegar et al., 2017). However, these algorithms suffer from overfitting and are susceptible to an intensive operation requiring much computational time. Currently, ensemble learning models such as extreme gradient boosting (Xgboost), light gradient boosting machine (LightGBM), and random forest (RF) have been employed to forecast groundwater contamination (Gupta and Natarajan, 2021; Rahmati et al., 2019; Singh et al., 2014; Taherdangkoo et al., 2021). In comparison with single models, these methods use more base learners to achieve accurate results. In addition, they can also reduce variance, minimize bias, and then decrease the overfitting problems in the model (Huang et al., 2009, 2020b). For instance, Xgboost was employed to predict the subsurface water levels in Malaysia and found that it performs better than other models (Ibrahem Ahmed Osman et al., 2021). LightGBM was utilized to forecast groundwater level anomalies in the aquifers of South Africa, and the model achieved satisfactory performance with less error (Gaffoor et al., 2022). RF was also used to forecast groundwater quality assessment in Miandoab and nitrate in Africa, and the model gave better results due to its capability to avoid overfitting (Norouzi and Moghaddam, 2020; Ouedraogo et al., 2019). However, these approaches are challenging to interpret and require ample space and extensive training time (Huang et al., 2020a). Therefore, an alternative hybrid model of ensemble learning (RF) and logistic regression (LR) is necessary as an adopted method for groundwater contamination.

Many researchers have recently employed hybrid models for estimating groundwater contamination generated by combining several ML techniques (Gupta et al., 2022; Khosravi et al., 2018; Kombo et al., 2020; Ransom et al., 2017; Talukdar et al., 2022). For instance, ELM was combined with adaptive neuro-fuzzy analysis for groundwater estimation (Afkhamifar and Sarraf, 2021; Azizpour et al., 2022). A hybrid RF K-nearest neighbor (KNN-RF) was applied to estimate groundwater levels (Cao and Yu, 2014; Mehta et al., 2018). However, those hybrid models were employed with no evident technique to confirm their findings in predicting groundwater contamination. To enhance further analysis and development for groundwater estimation, we suggested a novel hybrid random forest linear model (HRFLM), which is more flexible and can correlate more specific answers with three ensemble learnings as evident methods for predicting groundwater fluoride.

In the present research, we aim to employ a hybrid HRFLM model to estimate groundwater fluoride contamination in the Datong basin. Secondly, HRFLM, Xgboost, LightGBM, and RF were compared to identify the potential algorithm for predicting groundwater fluoride contamination. This study disclosed that the novel union of RF with the linear model (LR) to form the HRFLM model could improve groundwater contamination assessment and can be applied to various fields and study areas. Also, the study will prove the performance of classification analysis in estimating water parameters.

Study area

The Datong basin is around 6000 km^2 in the Shanxi Province and is part of the Cenozoic faulted basins. It appertains to the semiarid climate in East Asia characterized by seasonal regions (Su et al., 2015; Wang and Shpeyzer, 2000). The average precipitation is about 225 to 400 mm generally from July to August. The air temperature is 6.5 °C per year, and the evapotranspiration is below 2000mm (Xing et al., 2013). Also, mountains and slopes bordered the area from the northwest to the southwest. The Sanggan and Huang Shui rivers constitute the principal river across the site from the south to the north (Fig. 1), and they serve the area in the irrigation process for agricultural activities (Su et al., 2013b).

The bedrock outcrops are located in the north, west, and east. Archean gneiss and basalt are the outcrops for the north. The west comprises Cambrian–Ordovician limestone, shale, and Carboniferous–Permian–Jurassic sandstone. Shale, Cambrian–Ordovician limestone, and Carboniferous–Permian–Jurassic sandstone are also found in the research zone's western part. Granite and Archean

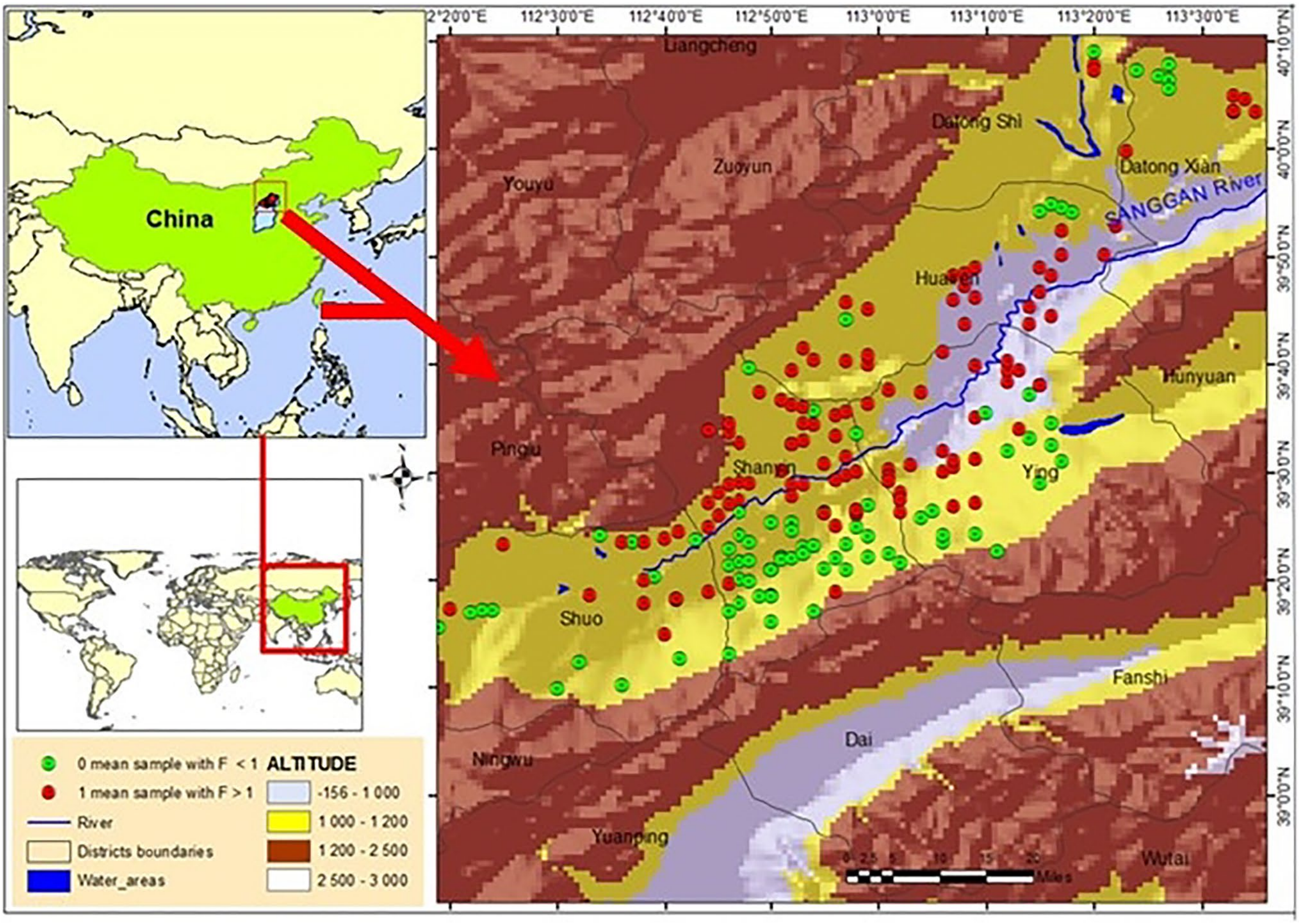


Fig. 1 The Dantong basin (research zone)

gneiss are sparsely located in the basin's northeast part. Aluvial-pluvial sand and gravel are the primary sediments in the area. Sandy loam soils, alluvial–lacustrine, and alluvial–pluvial sands are in the central part of the basin. Moreover, in the central part, silty clay and silts opulent in organic matter are identified (Guo and Wang, 2005).

Additionally, aquifers in the area are in the center below the plain flat alluvial-lacustrine, including upper, middle, and lower aquifers. Gravel and sands form the upper aquifer and are generally found at 5 to 60 m below the land, from a distance of 2–10 m. Furthermore, the middle aquifer is made of sand and sandy gravel from 60 to 160 m below the land surface (Jiang et al., 2018; Xie et al., 2009). Lastly, the lower aquifer contains fine sand and silt detected at the lowest point, more significant than 160 m below the land. Groundwater revitalization is through infiltration from mountains in front of the basin's bedrock, meteoric water vertically, an outflow from non-perennial rivers laterally, and irrigation return flow. Evaporation and abstraction are the foremost reasons for the subsurface water discharged in the study area (Guo and Wang, 2005).

Methodology

Sampling and laboratory analysis

In this research, 202 subsurface water samples were gathered from numerous wells in the research area formerly discussed by Nafouanti et al. (2021a, b). Samples were sifted via 0.45-mm membrane sieves and gathered in 500-mL pre-cleaned polyethylene flasks. They were washed with deionized water to avoid contamination while collecting and pumped for 5 to 10 min to flush the pipe-floating solids to find fresh groundwater. The physical elements, like total dissolved solids (TDS), were measured using HACH Instruments portable Hana meters (Sension+ MM150). The Ion Chromatograph DX-120 (Thermo Scientific, USA) was utilized in the analytical procedures to assess the significant anions (Cl^-, NO_3^-, HCO_3^-, SO_4^{2-}, F^-). The Inductively Coupled Plasma Atomic Emission Spectroscopy, ICP-AES (IRIS Intrepid II XSP), was employed to measure the concentration of major cations (K^+, Na^+, Ca^{2+}, and Mg^{2+}). Similarly, Inductively

Coupled Plasma Mass Spectrometry (ICP-MS) (Agilent, USA) was employed to analyze the trace elements (Fig. s2). After every ten (10) samples, replicates and standards were inserted to preserve quality control and assurance. The groundwater physicochemical parameters are summarized in the descriptive statistical analysis presented in Table 1.

Machine learning methods

Dataset preprocessing

Preprocessing is the first phase of dealing with machine learning data before building the model. The dataset is composed of 10 predictor variables: Cl^-, TDS, K^+, Na^+, Ca^{2+}, $Mg^{2+}{}_{,}$ HCO_3^-, SO_4^{2-}, NO_3^-, Zn, and the output variables. Furthermore, 80% of the dataset was employed in training and 20% for testing. Likewise, we have altered the dataset into classes assigning zero (0) when the fluoride concentration is inferior to 1 mg/L and 1 when the element fluoride exceeds 1 mg/L according to the Chinese recommendation (Pi et al., 2015). Likewise, all models' data were scaled from 0 to 1 to improve the model's speed and performance. The technique of min–max normalization was also used for the dataset scaling, as predictor variables in the different ranges can lead to inaccurate models. This technique has been previously utilized (Alkindi et al., 2022; Elbeltagi et al., 2022) for building a water resources model, and it is defined as follows:

$$X_2 = \frac{X - X_{\min}}{X_{\max -}X_{\min}} \tag{1}$$

Here, X is the predictor variable and X_2 is the normalized predictor variable. $X_{\min}$ and $X_{\max}$ correspondingly represent the input variable's minimum and maximum values. The flowchart of the modeling procedures has been demonstrated in Fig. 2.

Pearson correlation coefficient (Pr)

Pearson's correlation analysis was utilized to predict the linear association between the inputs and the output variables (Bolandi et al., 2017; Nyakilla et al., 2021) to study the significant linear association of each parameter (Yan and Au, 2019). Therefore, water parameters such as Cl^-, TDS, K^+, Na^+, Ca^{2+}, $Mg^{2+}{}_{,}$ HCO_3^-, SO_4^{2-}, NO_3^-, and Zn can significantly impact fluoride release due to natural and anthropogenic activities. It is donated as shown in Eq. 1. The result of the Pearson correlation can be found in the supplementary material section (Fig. s1).

$$R_{a,b} = \frac{\mathrm{cov}(a,b)}{\sigma_a \sigma_b} \tag{2}$$

where R is the linear correlation between two data a, and b, σ_a is the standard deviation of a, cov indicate the covariance, and σ_a is the standard deviation of b.

Xgboost model

This study used the Xgboost to estimate groundwater fluoride contamination. Xgboost is a machine learning approach of gradient boosting machine proposed in 2016 by (Chen and Guestrin 2016) (Fig. 3). The ensemble algorithm uses weak learners' capacity to achieve robust performance, and it is very speedy and efficient in avoiding overfitting due the introduction of a new tree model with a loss function. Column subsampling and shrinkage approaches are utilized to decrease model variance and bias. The shrinkage process simplifies the reduction of bias when employing an individual tree, which enhances the process. Similarly, the randomization and the boosting iteratively average base learners techniques simplify the variance reduction by subsampling into the algorithm (Hu et al., 2021).

The procedure of Xgboost works as follows: For example, a data (DT) has ***n*** features with ***m*** number of instances, hence:

DT= [(x_i,y_i): i = 1··· ..***m***,x_i ƐR^m,y_i ƐR], $\acute{Y}_i$ will be the forecasted dependent variable of an ensemble tree model produced by the below calculations:

$$\hat{O}_i = \Phi\left(x_i\right) = \sum_{k=1}^{k} f_k \left(x_i\right) f_k \quad \epsilon\ \mathcal{F} \tag{3}$$

The variable k signifies the number of trees and F_k (k-th-tree). It is necessary to resolve the preceding equation by reducing the loss and regularization objective to discover the best functions.

Table 1 Statistical descriptive of physicochemical water parameters in the research zone (unit mg/l)

Features	Cl^-	TDS	F^-	Mg^{2+}	Na^+	K^+	HCO_3^-	Ca^{2+}	SO_4^{2-}	NO_3^+	Zn
Maximum	3086	9118	22.3	773	2346	327	1785	189.4	2689	572	0.22
Minimum	5.76	302	0.01	4.3	6.3	0.01	172	3.2	0.01	0.01	0.01
Mean	178	1223	16	56.8	219	7	479.4	44.3	196	41	0.07
Standard deviation	352	1293	1.9	77.9	319	25	296	29.1	341	79	0.02

Fig. 2 Flowchart for the modeling procedures

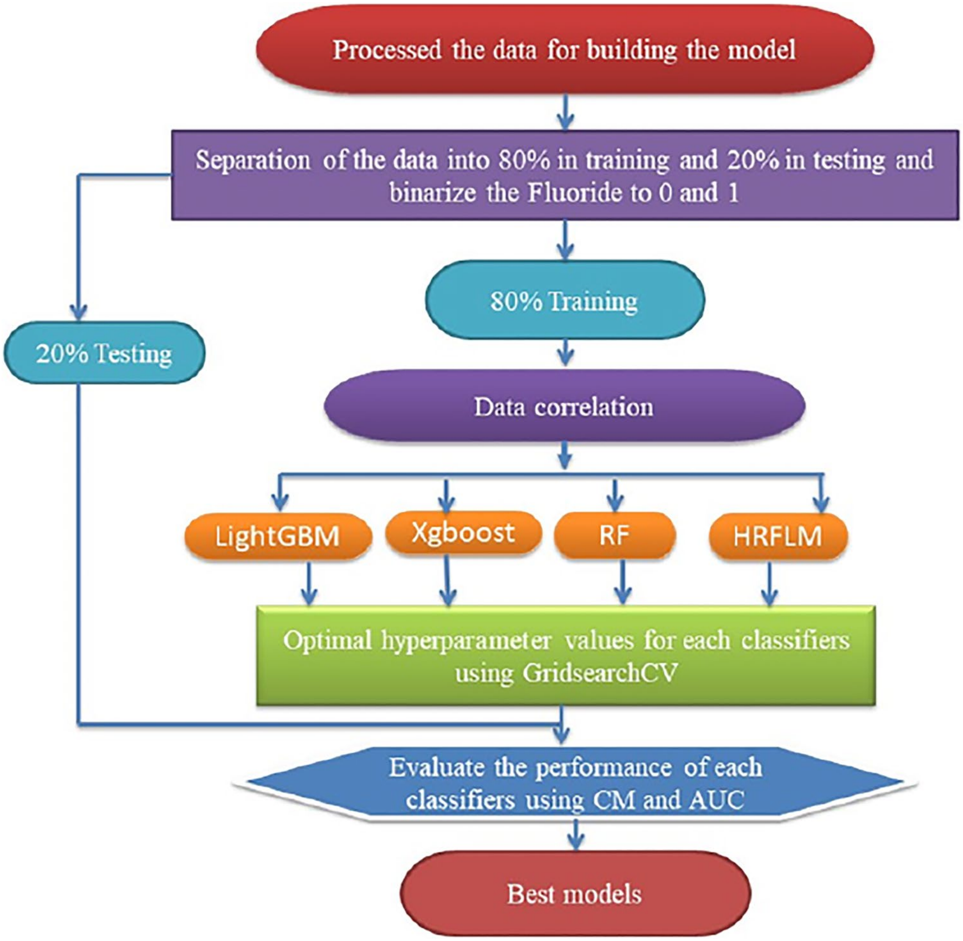
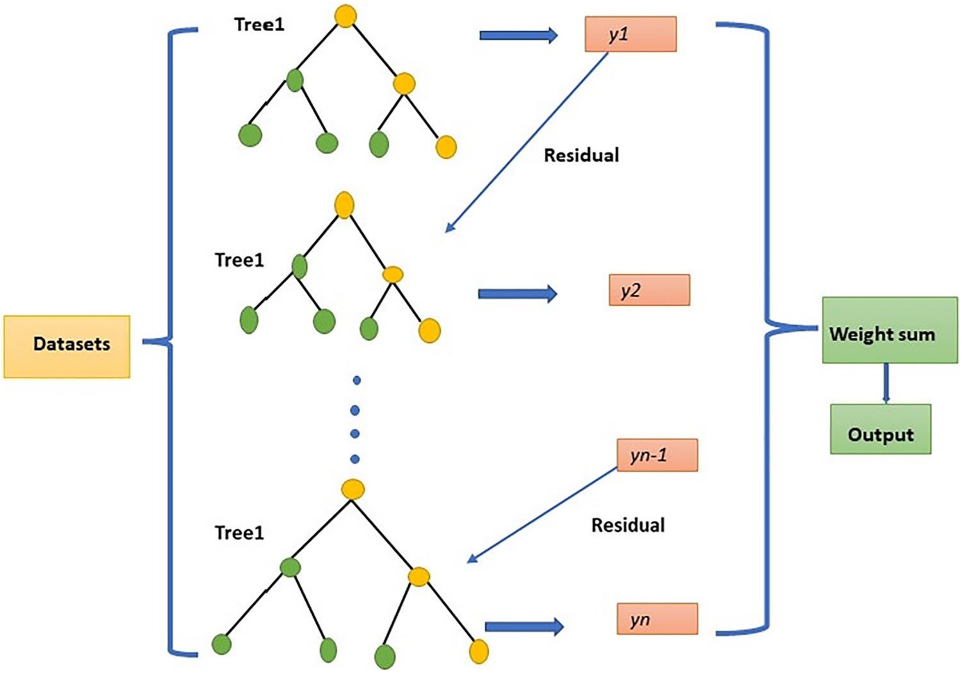

Processed the data for building the model
Separation of the data into 80% in training and 20% in testing and binarize the Fluoride to 0 and 1
80% Training
20% Testing
Data correlation
LightGBM
Xgboost
RF
HRFLM
Optimal hyperparameter values for each classifiers using GridsearchCV
Evaluate the performance of each classifiers using CM and AUC
Best models

Fig. 3 The Xgboost model procedures



$$\wp(\Phi) = \sum_{i} l\left(y_{\mathrm{i},}\dot{\mathrm{A}}._{i}\right) + \sum_{k} \Omega f_{(k)} \tag{4}$$

The variable *l* represents the loss function, the dissimilarity concerning the predicted output $\acute{Y}$, and the actual output y_{i}. Ω is a measure showing the complication of the model, and it helps the model avoid overfitting, and it is computed using:

$$\Omega\left(f_{\mathrm{k}}\right) = yT + \frac{1}{2}y\|w\|^{2} \tag{5}$$

The variable *T* symbolizes the number of leaves, and *w* symbolizes the weight of each leaf. The boosting function is utilized in decision trees when training the model to reduce the objective function. It works by adding a novel function *f* as the algorithm conserves training. Consequently, a novel function is added in the *t-th* iteration in the following procedures (Ibrahem Ahmed Osman et al., 2021):

$$\ell^{(t)} = \sum_{i=1}^{n} l\Big(y_{\mathrm{i},}\dot{\mathrm{A}}._{i}^{(t-1)} + f_{t(x_{\mathrm{i}})}\Omega\left(f_{\mathrm{t}}\right) \tag{6}$$

$$\ell_{\mathrm{split}} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_{\mathrm{L}}} g_{\mathrm{i}}\right)^{2}}{\sum_{i\in I_{\mathrm{L}}} h_{\mathrm{i}} + \lambda} + \frac{\left(\sum_{i\in I_{\mathrm{R}}} g_{\mathrm{i}}\right)^{2}}{\sum_{i\in I_{\mathrm{R}}} h_{\mathrm{i}} + \lambda} - \frac{\left(\sum_{i\in I} g_{\mathrm{i}}\right)^{2}}{\sum_{i\in I} h_{i} + \lambda}\right] - \gamma \tag{7}$$

$$g_{i} = \partial_{\dot{\mathrm{A}}.^{(t-1)}} 1\left(y_{i},\ \dot{\mathrm{A}}^{(t-1)}\right) \tag{8}$$

$$h_{1} = \partial^{2}_{\dot{\mathrm{A}}^{(\mathrm{t}-1)}} 1\left(y_{i},\ \dot{\mathrm{A}}.^{(t-1)}\right) \tag{9}$$

LightGBM

This research employed the LightGBM to predict fluoride contamination in groundwater. LightGBM is an algorithm with a great-performance gradient boosting framework (GBDT) built on a decision tree model. The LightGBM includes a Gradient-based One-Side Sampling (GOSS), Exclusive Feature Bundling (EFB), and histogram leaf-wise tree growth technique. The GOSS is a sampling technique that conserves all instances with high gradients and makes random sampling on the cases with a slight gradient. A constant multiplier for the dataset cases with a slight gradient is employed to compensate for the dataset distribution during the sampling process.

Furthermore, the EFB technique increases computational proficiency by allocating the variables into smaller bundles. In the LightGBM, the histogram of a leaf node could be computed by the difference between the parent node and the sibling node, which can reinforce the model speed in training and decrease memory consumption. In the histogram, it can be observed that consecutive floating-point eigenvalues are divided into small bins, which are employed in the construction of the histograms (Fig. 4). Firstly, statistical calculations are processed in the histogram which is the summation of gradients with the number of samples in the respective bin. This process will decrease the cost of calculation and storage in the model (Weng et al., 2019). In addition, the leaf-wise tree growth technique is necessary for optimizing and controlling model complexity. The leaves on the identical layer are extensively treated with various information gains (Kodaz et al., 2009). Thus, it indicates the reduction in entropy produced by dividing the nodes established on attributes, and it is determined as follows:

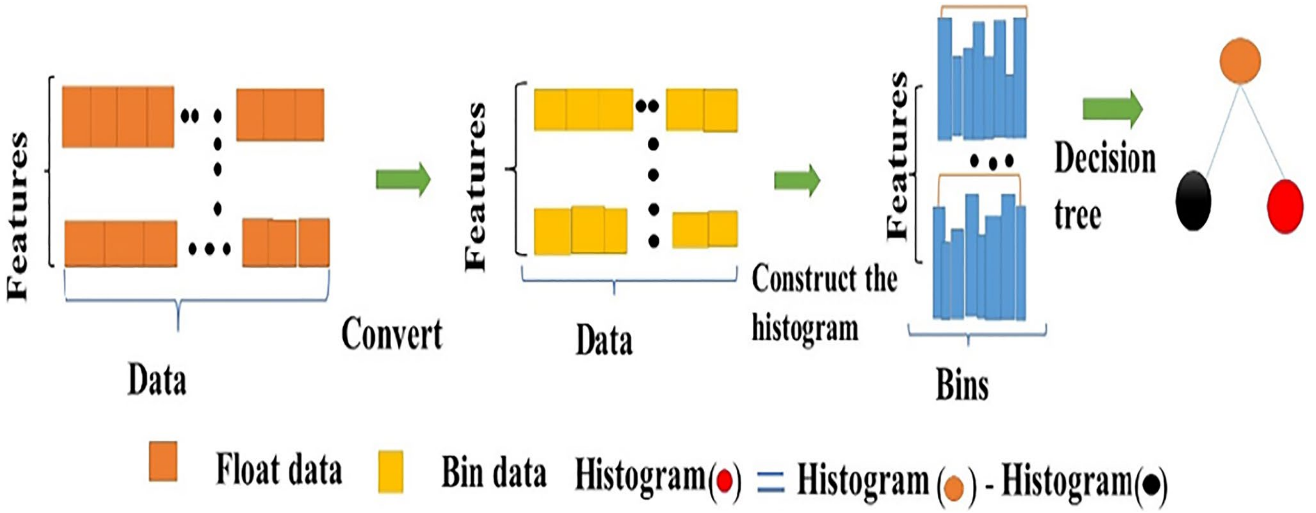


Fig. 4 Histogram-based decision tree technique

$$IG(B,V) = E_\mathrm{n}(\mathrm{B}) - \sum_{\mathrm{v\epsilon Values(W)}} \frac{\left|B_\vartheta\right|}{B} En\left(B_\vartheta\right) \tag{10}$$

$$E_\mathrm{n}(B) = \sum_{d=1}^{D} {}_{-}P_{d\log_2 P_\mathrm{d}} \tag{11}$$

where E_n(B) is the information entropy, p_d is the ratio of B relating to category d, and D is the category number. The attribute's value of V is shown as ϑ and B_ϑ is the subset of B.

RF

In this work, the RF was employed to forecast fluoride contamination in groundwater. RF forms a robust model by generating a thousand random trees to form a forest. In the RF, the model's principal parameters, including the number of trees and the predictors, are selected at each node (Tesoriero et al., 2017). The uncertainty of forecasting on the RF tree is calculated via its standard deviation equation below:

$$\rho = \sqrt{\frac{\sum_d^D (fb(x) - f)^2}{D - 1}} \tag{12}$$

where x is the unseen sample calculated by averaging the prediction $\sum_d^D fd(x)$ from every single tree, d and D are the repeated bagging from d to D, and d can be considered equal to 1.

In the RF, the random is accessible in two processes in the development of the tree. Primarily, a random is chosen with the substitute of the whole dataset rows from one-third of the dataset and "out-of-bag" (OBB), representing the data not arbitrarily chosen in the decision of the tree. Second is the limited number of randomly chosen parameters accessible in every node, and the output in the RF is expressed as:

$$y = \frac{1}{n}\sum_i^n = 1p_\mathrm{i} \tag{13}$$

n represents the number of trees, and p_i signifies each tree's prediction. In the model, the dimension of the tree can be controlled by setting the required samples at the trees' maximum depth and leaf node. The entropy is essential in the RF for determining the variable split at each node. It determines the homogeneity of the subset dataset, and when entropy is equivalent, the class label is identically split (Nguyen, 2020). However, zero entropy signifies that the sample is entirely homogenous and is expressed as below:

$$Entropy = -T\log_2(T) - p\log_2(p) \tag{14}$$

where T and p represent the probability of a randomly designated variable in a class n

Hybrid random forest linear model (HRFLM proposed method)

The HRFLM is an ML suggested by Senthilkumar Mohan (Mohan et al., 2019) to increase the prediction accuracy in classification analysis. A hybrid approach combines two or more algorithms to solve the same issue and is widely employed to predict various datasets (Hazarika et al., 2022; Khosravi et al., 2021). The hybrid was formed using a random forest classifier and linear model. Thus, three random forests and a linear model called the logistic regression (LR) algorithms were built. The LR is a generalized linear model employed as a linear classifier in classification analysis, and it is expressed as follows:

$$p = \frac{1}{1 + e^{-(b_o + b_1 x)}} \tag{15}$$

where b_o and b_1 are the estimated parameters

Therefore, a combination of RF and LR is a new method proposed to improve the prediction of groundwater fluoride contamination. The proposed method was implemented using the sklearn library in python, including matplotlib, pandas, and various compulsory libraries. HRFLM is a computational approach mining three association rules such as apriori, predictive, and Tertius (Nahar et al., 2013). Three RF classifiers and one linear model (LR) were described to build the hybrid. The log loss function has been employed to update the optimal weight in the combination of the two models. This technique will minimize the classification error and measure the degree to which the forecasting differs from the actual label, and it is determined as follows:

$$Logloss_i = -\left[y_\mathrm{i}\ln p_\mathrm{i} + \left(1 - y_\mathrm{i}\right)\ln\left(1 - p_\mathrm{i}\right)\right] \tag{16}$$

where i represents the specified observation or record, y means the actual value, and p is the prediction probability. Also, *ln* demonstrates the natural logarithm of a number which is the base of a mathematical constant. Thus, the model was combined with a controlling weight average, trained, and tested to evaluate the performance accuracy of the HRFLM.

Metrics for the various model evaluation

The ROC (AUC) and the confusion matrix (CM) were employed to assess the performance of various models. The CM demonstrates the ability of the model to classify the actual values compared to the predictive values. The accuracy, specificity, sensitivity, and error rate were computed to evaluate the algorithms' estimation. The different metrics equations are defined as follows (Fan et al., 2022; Hazarika and Gupta, 2022):

$$Sensitivity = \frac{TP}{TP + FN} \tag{17}$$

$$Specificity = \frac{TP}{TN + FP} \tag{18}$$

$$Error\ rate = \frac{FP + FN}{TP + TN + FN + FP} \tag{19}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{20}$$

The sensitivity is the same as true positive (TP) and represents the proportion of samples predicted with high fluoride. Specificity is the same as true negative (TN) and represents the proportion predicted with no fluoride. The accuracy signifies the percentages that are appropriately classified. In the evaluation, the prediction error is known as the false-positive (FP), which means the model predicts an element as fluoride while the prediction is false. The false-negative (FN) represents the reverse error in binary classification analysis, meaning an element is a fluoride, still, the model distinguishes it mistakenly as non-fluoride, and FN and FP represent the error rate.

Results and discussions

Models' performance results using the confusion matrix

The CM results were employed to evaluate the HRFLM, Xgboost, LightGBM, and RF performance in forecasting groundwater fluoride contamination and detailed in the supplementary material section (Table s1, Table s2, Table s3, and Table s4), respectively. Furthermore, from the scikit-learn GridSearchCV Python library, a set of hyperparameters have chosen for the HRFLM model by searching the equilibrium between good accuracy and regularization (Table 2). A depth of 10 was chosen for the model as a large depth could lead to overfitting (Huang et al., 2019). Also, the parameter "C" equals 10 because selecting a small value yields a better regularization (Uscanga-junco et al., 2021). The performance of the HRFLM model in the training and testing stage was achieved with satisfactory results in the prediction of groundwater fluoride contamination. The training phase's accuracy, sensitivity, specificity, and error rate were 98%, 98.2%, 98%, and 2%, respectively. In the testing stage, the model provided an accuracy of 95%, a sensitivity of 95.2%, a specificity of 95%, and a minimum error of 5% (Table 3). The model has shown a high sensitivity and specificity, signifying that the HRFLM model has shown a prediction ability between the different classes. The high sensitivity of the model confirms the presence of a large quantity of fluoride in groundwater in the research zone. Likewise, the training time of the model was 0.18 ms and 0.02 ms in the testing, demonstrating a less time-consuming model. Therefore, the model is acceptable due to its high accuracy, sensitivity, specificity, less error rate, and computational time. Our findings reveal a considerable ability of the HRFLM model to estimate fluoride contamination in groundwater. Based on the high performance of HRFLM, it can be applied to evaluate many contaminants in the subsurface water and can be adapted to different domains.

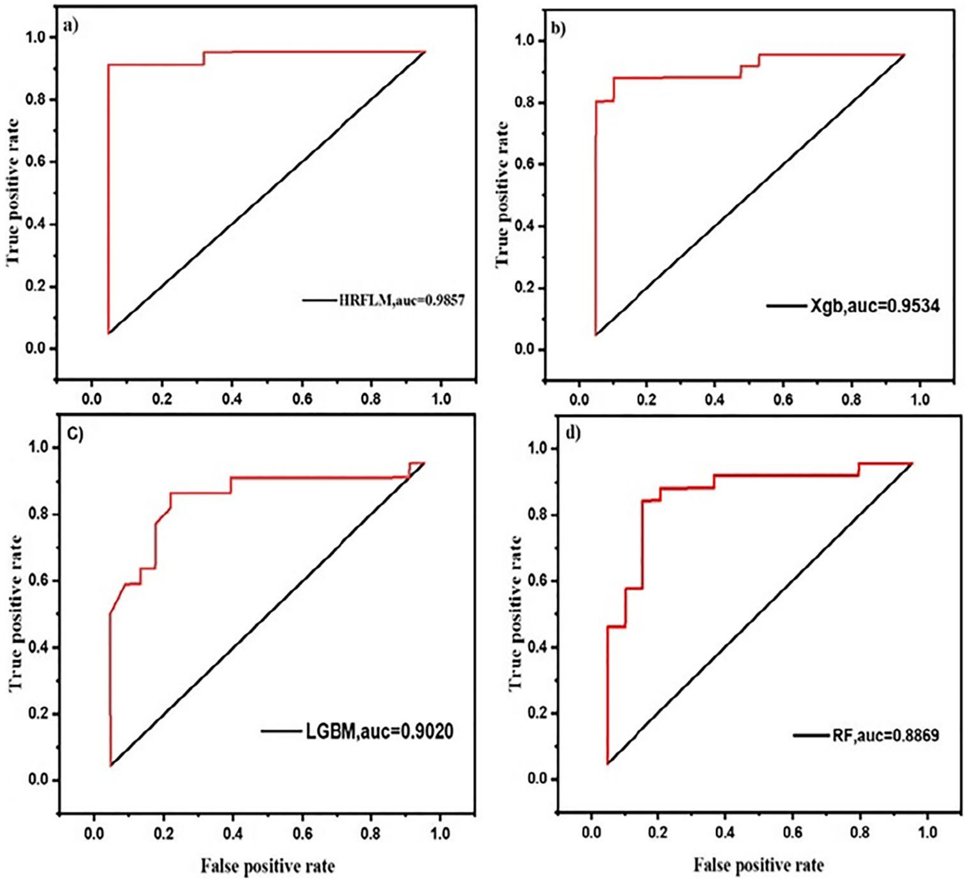
Table 2 Hyperparameter optimization of the HRFLM model using GridsearchCV

HRFLM hyperparameters	Meaning	Optimal value
n_estimators	Number of trees	50
min_samples_split	Minimum number of samples for nodes split	2
min_samples_leaf	Minimum number of samples for leaf node	1
max_depth	Maximum depth of trees	10
random_state	an integer value implying the selection of a random	42
max_iter	Maximum number of iterations	500
C	Regularization parameter	10

Table 3 Statistical evaluation using the confusion matrix metric

Training					Testing			
	HRFLM	Xgboost	LightGBM	RF	HRFLM	Xgboost	LightGBM	RF
Accuracy	98	96	96	94	95	88	88	85
Sensitivity	98.2	97	97	96	95.2	92	92	90
Specificity	98	95	95	94	95	82	82	80
Error rate	2	4	4	6	5	13	13	15
Time (ms)	0.18	0.27	0.28	0.33	0.02	0.08	0.08	0.010

Fig. 5 The area under the curve of the HRFLM (a), Xgboost (b), LightGBM (c), and RF(d)



Furthermore, the results of LightGBM in the training phase yield an accuracy of 96 %, a sensitivity of 97%, a specificity of 94%, and an error rate of 4%, whereas, in the testing, the results were 88%, 92%, 82%, and 13%, respectively (Table 3). The model used 0.27 ms in the training and 0.08 ms for testing during the prediction. Therefore, the prediction results demonstrated in the LightGBM suggest good performance modeling. The sensitivity result of the model confirms the HRFLM model's findings, showing a large number of fluorides in the research area. Besides, for the Xgboost, the accuracy, sensitivity, specificity, and error rate in training were 96%, 97%, 94%, and 4%, respectively. In the testing, the model accuracy, sensitivity, specificity, and error rate were 88%, 92%, 82%, and 13%, correspondingly. The model's training and testing times were 0.28 ms and 0.08 ms, respectively. Therefore, the model achieved good results and confirmed the high amount of fluoride in the subsurface water of the research area, confirming the HRFLM model findings. The GridSearchCV technique was employed to select the LightGBM and the Xgboost hyperparameters in Tables s5 and Table s6 in the supplementary materials. The good performance of the LightGBM and the Xgboost in predicting subsurface water fluoride contamination was enhanced by the best hyperparameters selected using the GridSearchCV technique. The Xgboost was previously employed in predicting water quality (Li et al., 2022; Liang et al., 2020). In our research, Xgboost has shown an ability to predict groundwater fluoride contamination.

Additionally, for the RF, the accuracy, sensitivity, specificity, and error rate for training were 94%, 96%, 94%, and 6%, and those for testing were 85%, 90%, 80%, and 15%, respectively (Table 3). Likewise, as demonstrated in Table 3, the model utilized 0.33 ms time in training and 0.010 ms in the testing. The RF demonstrates high sensitivity, meaning that the model identifies more samples with high fluoride groundwater. This result confirmed the result yielded by the HRFLM model, showing a high amount of fluoride in groundwater in the research area. Similarly, a set of hyperparameters have been selected to enhance the model performance by employing the GriserachCV technique (Table s7). The RF was applied previously in forecasting subsurface water fluoride contamination (Naghibi et al., 2016; Wu et al., 2020). Our result reveals the RF's performance ability in predicting groundwater fluoride contamination.

Evaluation results using the metrics ROC (AUC)

Our findings demonstrate that the performance of the HRFLM model using the AUC reached 0.98 in the prediction of fluoride contamination (Fig. 5a). The AUC measures

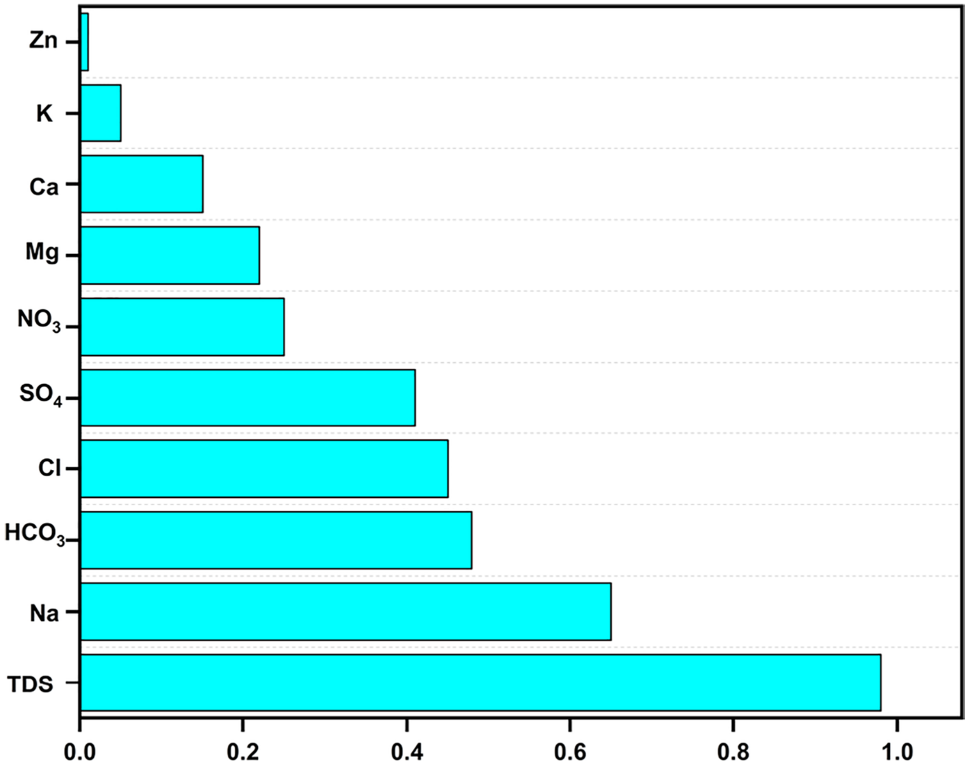


Fig. 6 The relative significance of the model's independent variables

the TP (sensitivity) and FP (false positive); when the AUC is closer to 1, it demonstrates that the model can classify the positive class. The results of the AUC of the HRFLM model indicate a high TP confirming the presence of a high quantity of fluoride in groundwater in the research zone, showing a useful application of the model. Likewise, the Xgboost model yielded an AUC of 0.95 in the prediction of groundwater fluoride contamination (Fig. 5b). A previous study has shown the prediction ability of the Xgboost model using the AUC metric with an AUC of 0.87 in the estimation of subsurface water (Arabameri et al., 2021). Regarding our findings, the model suggests a good performance in estimating fluoride in groundwater.

The LightGBM model's result using the AUC is demonstrated in Fig. 5c, and the model achieved an AUC of 0.90 in estimating groundwater fluoride contamination. The RF model achieved an AUC of 0.88, showing a good forecast as the value of AUC is closer to 1 (Fig. 5d). Therefore, the LightGBM and RF models achieved a good performance in forecasting groundwater fluoride contamination.

Model comparisons: HRFLM, LightGBM, Xgboost, and RF

The results presented in Table 3 demonstrate that the HRFLM model achieved high accuracy, sensitivity, specificity, and error rate in forecasting groundwater fluoride contamination compared to LightGBM, Xgboost, and RF models. Furthermore, the AUC of HRFLM outperforms the other models in predicting fluoride contamination in groundwater (Fig. 5a). Therefore, the HRFLM model has shown significant performance in estimating fluoride pollution in groundwater and reveals a useful application. The model has demonstrated less computational time during the prediction. Therefore, the great achievement of the HRFLM algorithm is attributed to the reason that hybrid models are soft computing due to various optimization techniques that improve the model's flexibility and accuracy (Ardabili, Mosavi, and Várkonyi-Kóczy 2020; Kondababu et al., 2021).

As shown in Table 3 and Fig. 5 (b and c), the performance results of the Xgboost and LightGBM are lower than the hybrid HRFLM in the prediction ability of groundwater fluoride. The LightGBM model can be prone to overfitting, which might be attributed to the lower performance of the model (Liu, 2022), and the Xgboost model is sensitive to outliers (Budholiya et al., 2020; Duan et al., 2021). The RF model has the lowest accuracy, sensitivity, specificity, and error rate (Table 3) in estimating subsurface water fluoride contamination. Similarly, the RF model has the lowest performance AUC in the prediction (Fig. 5d). The model's lower performance can be assigned to the model's requirement for much training time (Lopez et al., 2020). Our findings reveal that all models yielded satisfactory results in predicting fluoride contamination in groundwater. However, the adopted hybrid HRFLM gave the best performance results ($accuracy = 95\%$, $AUC = 0.98$), followed by Xgboost (*accuracy* = 88%, *AUC* = 0.95), LightGBM (*accuracy* = 88%, *AUC* = 0.90), and RF (*accuracy* = 85%, *AUC* = 0.88).

Sensitivity analysis of inputs parameters

Sensitivity analysis was executed to identify the importance of input variables on fluoride contamination in groundwater estimation (Eq. 16) (Nyakilla et al., 2022). Figure 6 demonstrates the comparative significance of every input element to the target variable. It is evident that TDS and Na^+ are the effective majority factors influencing fluoride in groundwater, contributing to 95% and 65% of model architecture, respectively. HCO_3^-, Cl^-, and SO_4^{2+} also have an important effect contributing about 50%, 48%, and 46 % to model development, and NO_3^-, Mg^{2+}, and Ca^{2+} contribute about 28%, 26%, and 18%, respectively.

$$WL = \frac{1}{T}\sum_{i=1}^{T}\left(\frac{\delta Output\%}{\delta Input\%}\right)_i x100 \tag{21}$$

The analysis reveals that each input contributes significantly to fluoride estimation except for Zn and K^+, which have 5% and 2% of fluoride estimation, respectively, where *δInput%* signifies the change in percent of input and *δOutput%* signifies the change in percent of output. This result indicates that variables alter from maximum to minimum values. The smaller *WL* value specifies that an independent variable has less impact on the fluoride released in groundwater, whereas the higher value of *WL* demonstrates that input variables affected the release of fluoride in groundwater. Therefore the parameters TDS, Na^+, HCO_3^-, Cl^-, SO_4^{2+}, NO_3^-, and Mg^{2+} have significant importance to the fluoride intrusion in the subsurface water, which has been confirmed by previous studies (Nafouanti et al., 2021a, b; Yang et al., 2021).

Conclusion

This study investigated the prediction ability of the proposed hybrid HRFLM to estimate subsurface water fluoride contamination in the Datong basin. Three ensemble learning, including LightGBM, Xgboost, and RF, were also employed as evident approaches for forecasting fluoride contamination in groundwater. Our findings revealed that the proposed hybrid HRFLM outperformed the Xgboost, LightGBM, and RF in forecasting fluoride pollution in groundwater.

By employing the GridsearchCV hyperparameters, the HRFLM model was achieved with high sensitivity of 95.2%, an accuracy of 95%, a specificity of 95%, and a lower error rate of 5%. The model has used less computational time in training (0.18 ms) and testing (0.02 ms).

Moreover, the AUC of HRFLM was compared with the AUC of ensemble learning models such as LightGBM, Xgboost, and RF, which also demonstrated promising results by achieving an AUC of 98% in estimating groundwater fluoride contamination. Thus, due to the model's reliable findings, the hybrid HRFLM is recommended for estimating groundwater fluoride contamination due to the reliable and flexible results of the algorithm, and the method can be applied to various fields and research areas. Notwithstanding the above insights, it is difficult to understand the process of groundwater contamination owing to the presence of several parameters. Therefore, future studies need to be conducted on developing various hybrid models that are more advanced to enhance the progress of groundwater prediction for better protection and sustainability.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11356-023-25886-w.

Acknowledgments The authors are grateful to all research members who contributed to the data collection and any work on this research and express their gratitude to the anonymous reviewers and the editor for the time they used for this study.

Authors' contributions Mouigni Baraka Nafouanti: Conceptualization, investigation, machine learning methodology analysis, and writing. Junxia Li: Conceptualization, laboratory analysis, review, and supervision. Edwin E. Nyakilla: Review, editing, and model checking. Grant Charles Mwakipunda: Review and editing. Alvin Mulashani: Review and editing

Funding (42020104005 and 42177067), (2021YFA0715900)

The research work was financially supported by the National Natural Science Foundation of China, the National Key Research and Development Program of China, the Ministry of Education of China (111 projects), and the Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan).

Data availability All the data and materials related to the manuscript, such as code, can be found at this link https://github.com/Nafouant/Nafouanti-Mouigni-Baraka and data will be available upon request.

Declarations

Ethics approval and consent to participate Not applicable

Consent for publication All the co-authors agreed to publish the manuscript.

Competing interests The authors declare no competing interests.

References

Adamowski J, Chan HF (2011) A wavelet neural network conjunction model for groundwater level forecasting. J. Hydrol. 407:28–40

Afkhamifar, S., Sarraf, A., 2021. Comparative study of groundwater level forecasts using hybrid neural network models, in Proceedings of the Institution of Civil Engineers-Water Management. Thomas Telford Ltd, pp. 267–277.

Alkindi KM, Mukherjee K, Pandey M, Arora A, Janizadeh S, Pham QB, Anh DT, Ahmadi K (2022) Prediction of groundwater nitrate concentration in a semiarid region using hybrid Bayesian artificial intelligence approaches. Environ Sci. Pollut. Res. 29:20421–20436

Al-Mohair HK, Saleh JM, Suandi SA (2015) Hybrid human skin detection using neural network and K-means clustering technique. Appl Soft Comput 33:337–347

Arabameri A, Chandra Pal S, Costache R, Saha A, Rezaie F, Seyed Danesh A, Pradhan B, Lee S, Hoang N-D (2021) Prediction of gully erosion susceptibility mapping using novel ensemble machine learning algorithms. Geomatics. Nat. Hazards Risk 12:469–498

Ardabili S, Mosavi A, Várkonyi-Kóczy AR (2020) Advances in machine learning modeling reviewing hybrid and ensemble methods. Lect. Notes Networks Syst. 101:215–227. https://doi.org/10.1007/978-3-030-36841-8_21

Azizpour A, Izadbakhsh MA, Shabanlou S, Yosefvand F, Rajabi A (2022) Simulation of time-series groundwater parameters using a hybrid metaheuristic neuro-fuzzy model. Environ. Sci. Pollut. Res. 29:28414–28430

Barzegar R, Asghari Moghaddam A, Adamowski J, Fijani E (2017) Comparison of machine learning models for predicting fluoride contamination in groundwater. Stoch. Environ. Res. Risk Assess. 31:2705–2718. https://doi.org/10.1007/s00477-016-1338-z

Bolandi V, Kadkhodaie A, Farzi R (2017) Analyzing organic richness of source rocks from well log data by using SVM and ANN classifiers: a case study from the Kazhdumi formation, the Persian Gulf basin, offshore Iran. J. Pet. Sci. Eng. 151:224–234

Breiman, L., 2004. Consistency for a simple model of random forests.

Brindha K, Elango L (2011) Fluoride in groundwater: causes, implications and mitigation measures. Fluoride Prop. Appl. Environ. Manag.:113–136

Budholiya K, Shrivastava SK, Sharma V (2020) An optimized XGBoost based diagnostic system for effective prediction of heart disease. J. King Saud Univ. - Comput. Inf. Sci. https://doi.org/10.1016/j.jksuci.2020.10.013

Cao, L., Yu, P.S., 2014. A hybrid coupled k-nearest neighbor algorithm on imbalance data. https://doi.org/10.1109/IJCNN.2014.6889798

Chang Z, Du Z, Zhang F, Huang F, Chen J, Li W, Guo Z (2020) Landslide susceptibility prediction based on remote sensing images and GIS: comparisons of supervised and unsupervised machine learning models. Remote Sens. 12. https://doi.org/10.3390/rs12030502

Chen, T., Guestrin, C., 2016 Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. pp. 785–794.

Duan J, Asteris PG, Nguyen H, Bui X-N, Moayedi H (2021) A novel artificial intelligence technique to predict compressive strength of recycled aggregate concrete using ICA-XGBoost model. Eng. Comput. 37:3329–3346

Elbeltagi A, Pande CB, Kouadri S, Islam ARM (2022) Applications of various data-driven models for the prediction of groundwater quality index in the Akot basin, Maharashtra. India. Environ. Sci. Pollut. Res. 29:17591–17605

Fan, X., Wang, X., Zhang, X., ASCE Xiong (Bill) Yu, P.E.F., 2022. Machine learning based water pipe failure prediction: the effects of engineering, geology, climate, and socio-economic factors. Reliab. Eng. Syst. Saf. 219, 108185. https://doi.org/10.1016/j.ress.2021.108185

Feng DC, Liu ZT, Wang XD, Chen Y, Chang JQ, Wei DF, Jiang ZM (2020) Machine learning-based compressive strength prediction for concrete: an adaptive boosting approach. Constr. Build. Mater. 230:117000. https://doi.org/10.1016/j.conbuildmat.2019.117000

Gaffoor Z, Gritzman A, Pietersen K, Jovanovic N, Bagula A, Kanyerere T (2022) An autoregressive machine learning approach to forecast high-resolution groundwater-level anomalies in the Ramotswa/North West/Gauteng dolomite aquifers of Southern Africa. Hydrogeol. J.:1–26

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

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

Gupta D, Hazarika BB, Berlin M, Sharma UM, Mishra K (2021) Artificial intelligence for suspended sediment load prediction: a review. Environ. Earth Sci. 80:346. https://doi.org/10.1007/s12665-021-09625-3

Gupta D, Natarajan N (2021) Prediction of uniaxial compressive strength of rock samples using density weighted least squares twin support vector regression. Neural Comput Appl 33:15843–15850. https://doi.org/10.1007/s00521-021-06204-2

Gupta D, Natarajan N, Berlin M (2022) Short-term wind speed prediction using hybrid machine learning techniques. Environ. Sci. Pollut. Res. 29:50909–50927. https://doi.org/10.1007/s11356-021-15221-6

Hazarika BB, Gupta D (2022) MODWT—random vector functional link for river-suspended sediment load prediction. Arab. J. Geosci. 15:966. https://doi.org/10.1007/s12517-022-10150-1

Hazarika BB, Gupta D, Natarajan N (2022) Wavelet kernel least square twin support vector regression for wind speed prediction. Environ. Sci. Pollut. Res. 29:86320–86336. https://doi.org/10.1007/s11356-022-18655-8

He X, Li P, Wu J, Wei M, Ren X, Wang D (2021) Poor groundwater quality and high potential health risks in the Datong Basin, northern China: research from published data. Environ. Geochem. Health 43:791–812

Hu L, Wang C, Ye Z, Wang S (2021) Estimating gaseous pollutants from bus emissions: a hybrid model based on GRU and XGBoost. Sci. Total Environ. 783:146870

Huang F, Cao Z, Guo J, Jiang S-H, Li S, Guo Z (2020a) Comparisons of heuristic, general statistical, and machine learning models for landslide susceptibility prediction and mapping. CATENA 191:104580 https://doi.org/10.1016/j.catena.2020.104580

Huang F, Cao Z, Jiang S-H, Zhou C, Huang J, Guo Z (2020b) Landslide susceptibility prediction based on a semi-supervised multiple-layer perceptron model. Landslides 17:2919–2930. https://doi.org/10.1007/s10346-020-01473-9

Huang, F., Xie, G., Xiao, R., 2009. Research on ensemble learning, in 2009 International Conference on Artificial Intelligence and Computational Intelligence. IEEE, pp. 249–252.

Huang F, Zhang J, Zhou C, Wang Y, Huang J, Zhu L (2020c) A deep learning algorithm using a fully connected sparse autoencoder neural network for landslide susceptibility prediction. Landslides 17:217–229. https://doi.org/10.1007/s10346-019-01274-9

Huang G, Wu L, Ma X, Zhang W, Fan J, Yu X (2019) Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. J. Hydrol. 574:1029–1041. https://doi.org/10.1016/j.jhydrol.2019.04.085

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 Eng. J. 12, 1545–1556. https://doi.org/10.1016/j.asej.2020.11.011

Jain AK, Mao J, Mohiuddin KM (1996) Artificial neural networks: a tutorial. Computer (LongBeach. Calif) 29:31–44

Jiang S-H, Huang J, Huang F, Yang J, Yao C, Zhou C-B (2018) Modeling spatial variability of soil undrained shear strength by conditional random fields for slope reliability analysis. Appl. Math. Model. 63:374–389 https://doi.org/10.1016/j.apm.2018.06.030

Khosravi K, Barzegar R, Golkarian A, Busico G, Cuoco E, Mastrocicco M, Colombani N, Tedesco D, Ntona MM, Kazakis N (2021)

Predictive modeling of selected trace elements in groundwater using hybrid algorithms of iterative classifier optimizer. J. Contam. Hydrol. 242:103849

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

Khosravi K, Sartaj M, Tsai FT-C, Singh VP, Kazakis N, Melesse AM, Prakash I, Bui DT, Pham BT (2018) A comparison study of DRASTIC methods with various objective methods for groundwater vulnerability assessment. Sci. Total Environ. 642:1032–1049

Kodaz H, Özşen S, Arslan A, Güneş S (2009) Medical application of information gain based artificial immune recognition system (AIRS): diagnosis of thyroid disease. Expert Syst. Appl. 36:3086–3092

Kombo OH, Kumaran S, Sheikh YH, Bovim A, Jayavel K (2020) Long-term groundwater level prediction model based on hybrid KNN-RF technique. Hydrology 7:59

Kondababu A, Siddhartha V, Kumar BHKB, Penumutchi B (2021) A comparative study on machine learning-based heart disease prediction. Mater, Today Proc

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

Li L, Qiao J, Yu G, Wang L, Li H-Y, Liao C, Zhu Z (2022) Interpretable tree-based ensemble model for predicting beach water quality. Water Res. 211:118078 https://doi.org/10.1016/j.watres.2022.118078

Li, Y.M., 2001. Environmental chemistry study of waterborne poisoning in Shanyin, Shanxi province.

Liang W, Luo S, Zhao G, Wu H (2020) Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms. Mathematics 8:1–17. https://doi.org/10.3390/MATH8050765

Liu Y (2022) Grocery Sales Forecasting. In: in: 2022 International Conference on Creative Industry and Knowledge Economy (CIKE 2022). Atlantis Press, pp 215–219

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

Mamatha P, Rao SM (2010) Geochemistry of fluoride-rich groundwater in Kolar and Tumkur districts of Karnataka. Environ. Earth Sci. 61:131–142

Manap MA, Sulaiman WNA, Ramli MF, Pradhan B, Surip N (2013) A knowledge-driven GIS modeling technique for groundwater potential mapping at the Upper Langat Basin. Malaysia. Arab. J. Geosci. 6:1621–1637

Mehta S, Shen X, Gou J, Niu D (2018) A new nearest centroid neighbor classifier based on K local means using harmonic mean distance. https://doi.org/10.3390/info9090234

Mohan S, Thirumalai C, Srivastava G (2019) Effective heart disease prediction using hybrid machine learning techniques. IEEE Access 7:81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707

Nafouanti MB, Li J, Mustapha NA, Uwamungu P, Al-Alimi D (2021a) Prediction on the fluoride contamination in groundwater at the Datong basin, Northern China: comparison of random forest, logistic regression, and artificial neural network. Appl. Geochemistry 132:105054. https://doi.org/10.1016/j.apgeochem.2021.105054

Nafouanti MB, Li J, Mustapha NA, Uwamungu P, Dalal A-A (2021b) Prediction on the fluoride contamination in groundwater at the Datong Basin, Northern China: comparison of random forest, logistic regression, and artificial neural network. Appl. Geochemistry 132:105054

Naghibi SA, Pourghasemi HR, Dixon B (2016) GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. Environ. Monit. Assess. 188:1–27

Nahar J, Imam T, Tickle KS, Chen YPP (2013) Association rule mining to detect factors which contribute to heart disease in males and females. Expert Syst. Appl. 40:1086–1093. https://doi.org/10.1016/j.eswa.2012.08.028

Nguyen XH (2020) Combining statistical machine learning models with ARIMA for water level forecasting: the case of the Red River. Adv. Water Resour. 142:103656

Norouzi H, Moghaddam AA (2020) Groundwater quality assessment using random forest method based on groundwater quality indices (case study: Miandoab plain aquifer, NW of Iran). Arab. J. Geosci. 13. https://doi.org/10.1007/s12517-020-05904-8

Nyakilla EE, Silingi SN, Shen C, Jun G, Mulashani AK, Chibura PE (2021) Evaluation of source rock potentiality and prediction of total organic carbon using well log data and integrated methods of multivariate analysis, machine learning, and geochemical analysis. Nat. Resour. Res. https://doi.org/10.1007/s11053-021-09988-1

Nyakilla EE, Silingi SN, Shen C, Jun G, Mulashani AK, Chibura PE (2022) Evaluation of source rock potentiality and prediction of total organic carbon using well log data and integrated methods of multivariate analysis, machine learning, and geochemical analysis. Nat. Resour. Res. 31:619–641. https://doi.org/10.1007/s11053-021-09988-1

Orban P, Brouyère S, Batlle-Aguilar J, Couturier J, Goderniaux P, Leroy M, Maloszewski P, Dassargues A (2010) Regional transport modeling for nitrate trend assessment and forecasting in a chalk aquifer. J. Contam. Hydrol. 118:79–93. https://doi.org/10.1016/j.jconhyd.2010.08.008

Ouedraogo I, Defourny P, Vanclooster M (2019) Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale. Hydrogeol. J. 27:1081–1098. https://doi.org/10.1007/s10040-018-1900-5

Pi K, Wang Y, Xie X, Su C, Ma T, Li J, Liu Y (2015) Hydrogeochemistry of co-occurring geogenic arsenic, fluoride, and iodine in groundwater at Datong Basin, northern China. J. Hazard. Mater. 300:652–661

Rafique T, Naseem S, Bhanger MI, Usmani TH (2008) Fluoride ion contamination in the groundwater of Mithi sub-district, the Thar Desert. Pakistan. Environ. Geol. 56:317–326

Rahmati O, Choubin B, Fathabadi A, Coulon F, Soltani E, Shahabi H, Mollaefar E, Tiefenbacher J, Cipullo S, Ahmad BB (2019) Predicting uncertainty of machine learning models for modelling nitrate pollution of groundwater using quantile regression and UNEEC methods. Sci. Total Environ. 688:855–866

Ramanaiah SV, Mohan SV, Rajkumar B, Sarma PN (2006) Monitoring of fluoride concentration in groundwater of Prakasham district in India: correlation with physico-chemical parameters. J. Environ. Sci. Eng. 48:129

Ransom, K.M., Nolan, B.T., A. Traum, J., Faunt, C.C., Bell, A.M., Gronberg, J.A.M., Wheeler, D.C., Z. Rosecrans, C., Jurgens, B., Schwarz, G.E., Belitz, K., M. Eberts, S., Kourakos, G., Harter, T., 2017. A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA. Sci. Total Environ. 601–602, 1160–1172. https://doi.org/10.1016/j.scitotenv.2017.05.192

Singh KP, Gupta S, Mohan D (2014) Evaluating influences of seasonal variations and anthropogenic activities on alluvial groundwater hydrochemistry using ensemble learning approaches. J. Hydrol. 511:254–266 https://doi.org/10.1016/j.jhydrol.2014.01.004

Su C, Wang Y, Pan Y (2013a) Hydrogeochemical and isotopic evidences of the groundwater regime in Datong Basin. Northern China. Environ. Earth Sci. 70:877–885. https://doi.org/10.1007/s12665-012-2176-z

Su C, Wang Y, Pan Y (2013b) Hydrogeochemical and isotopic evidences of the groundwater regime in Datong Basin. Northern China. Environ. earth Sci. 70:877–885

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

Taherdangkoo R, Liu Q, Xing Y, Yang H, Cao V, Sauter M, Butscher C (2021) Predicting methane solubility in water and seawater by machine learning algorithms: application to methane transport modeling. J. Contam. Hydrol. 242:103844

Talukdar S, Mallick J, Sarkar SK, Roy SK, Islam ARMT, Praveen B, Naikoo MW, Rahman A, Sobnam M (2022) Novel hybrid models to enhance the efficiency of groundwater potentiality model. Appl. Water Sci. 12:62. https://doi.org/10.1007/s13201-022-01571-0

Tesoriero AJ, Gronberg JA, Juckem PF, Miller MP, Austin BP (2017) Predicting redox-sensitive contaminant concentrations in groundwater using random forest classification. Water Resour. Res. 53:7316–7331

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

Uscanga-junco, O.A., Rosales-rivera, M., Díaz-gonz, L., 2021. Development and comparison of machine learning models for water multidimensional classification 598. https://doi.org/10.1016/j.jhydrol.2021.126234

Vesselinov VV, Alexandrov BS, O'Malley D (2018) Contaminant source identification using semi-supervised machine learning. J. Contam. Hydrol. 212:134–142. https://doi.org/10.1016/j.jconhyd.2017.11.002

Wang Y, Wang T (2020) Application of improved LightGBM model in blood glucose prediction. Appl. Sci. 10. https://doi.org/10.3390/app10093227

Wang YX, Shpeyzer G (2000) Hydrogeochemistry of mineral waters from rift systems on the East Asia continent: case studies in Shanxi and Baikal. China Environ. Sci. Press, Beijing (in Chinese with English Abstr

Weng T, Liu W, Xiao J (2019) Supply chain sales forecasting based on lightGBM and LSTM combination model. Ind. Manag. Data Syst. 120:265–279

Wu C, Fang C, Wu X, Zhu G (2020) Health-risk assessment of arsenic and groundwater quality classification using random forest in the Yanchi region of Northwest China. Expo. Heal. 12:761–774

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

Xie X, Wang Y, Ellis A, Su C, Li J, Li M (2011) The sources of geogenic arsenic in aquifers at Datong basin, northern China: constraints from isotopic and geochemical data. J. Geochemical Explor. 110:155–166 https://doi.org/10.1016/j.gexplo.2011.05.006

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

Yan N, Au OT-S (2019) n locating the important variables on which other variables depend. Open Univ. J, Asian Assoc

Yang J, Zeng L, He X, Su Y, Li Y, Tan H, Jiang B, Zhu H, Oh SK (2021) Improving the durability of heat-cured high-volume fly ash cement mortar by wet-grinding activation. Constr. Build. Mater. 289:123157. https://doi.org/10.1016/j.conbuildmat.2021.123157

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