Vietnam Journal of Earth Sciences, 44(3), 410-429, https://doi.org/10.15625/2615-9783/17240

Vietnam Academy of Science and Technology
Vietnam Journal of Earth Sciences
http://www.vjs.ac.vn/index.php/jse

Development and application of hybrid artificial intelligence models for groundwater potential mapping and assessment

Duong Hai Ha[1], Huong Thi Thanh Ngo[2*], Tran Van Phong[3], Nguyen Duc Dam[2], Mohammadtaghi Avand[4], Huu Duy Nguyen[5], Mahdis Amiri[6], Hiep Van Le[2], Indra Prakash[7], Binh Thai Pham[2]

[1]*Institute for Water and Environment, Hanoi 100000, Vietnam*
[2]*University of Transport Technology, Hanoi 100000, Vietnam*
[3]*Institute of Geological Sciences, VAST, Hanoi, Vietnam*
[4]*Department of Watershed Management Engineering, College of Natural Resources, Tarbiat Modares University, Tehran 14115-111, Iran*
[5]*Faculty of Geography, VNU University of Science, Vietnam National University, Hanoi, Vietnam*
[6]*Gorgan University of Agricultural Sciences & Natural Resources, Department of Watershed & Arid Zone Management, Gorgan 4918943464, Iran*
[7]*Dy. Director General (R), Geological Survey of India, Gandhinagar (Guj.) 382002, India*

Received 17 March 2022; Received in revised form 19 May 2022; Accepted 18 June 2022

ABSTRACT

Groundwater potential assessment is essential for optimum utilization and recharge of groundwater resources for the proper development and management of an area. The main aim of this study is to develop an accurate groundwater potential map of the Dak Nong province (Vietnam) using hybrid artificial intelligence models, which are a combination of Random Forest (RF) and its Ensemble Framework (AdaBoost - ABRF, Bagging - BRF and LogitBoost - LBRF). In this study, twelve conditioning factors, namely topography (aspect, elevation, Topographic Wetness Index - TWI, slope, and curvature), hydrology (infiltration and river density, rainfall, Sediment Transport Index - STI, Stream Power Index - SPI), land use, and soil were used to develop the models. Well, yield data was also utilized to develop and validate potential groundwater zones.

One Rule (R) feature selection method was utilized to prioritize the importance of groundwater potential affecting parameters. The results indicated that the Average Merit (AM) of the rainfall factor was the highest (68.039), and river density was the lowest (53,969). Performance evaluation of ML models was done using standard statistical indicators, including Area Under the Receiver Operating Characteristic (ROC) curve (AUC). The results showed that all the four models performed well in the training (AUC≥0.967) and testing (AUC≥0.734) phases, but the performance of the ABRF (AUC=0.992) model is the best in the training phase, whereas LBRF is the best in the testing phase (AUC=0.776). The present model study would be helpful in the proper groundwater potential assessment and management of groundwater resources for sustainable development.

Keywords: Groundwater potential mapping, Random Forest, artificial intelligence, hybrid models, Vietnam.

1. Introduction

Groundwater is a reliable source of fresh water, especially in semi-arid and arid regions (Jha et al., 2009). Recently, due to the increase in population and climate change effect, surface water resources cannot meet the requirement of fresh water for irrigation,

*Corresponding author, Email: huongntt@utt.edu.vn

drinking, and industrial use, thus increasing pressure on the utilization of groundwater resources (Jha et al., 2007). This has necessitated the systematic mapping of groundwater resources by using modern technology for proper groundwater potential assessment of an area for its optimum utilization and management in conjunction with surface water for sustainable development.

Traditional methods of groundwater assessment based on field mapping and exploration require a great deal of time and cost (Bonham-Carter, 2014; Chen et al., 2017). Therefore, newer statistical methods were utilized with GIS: Geographic Information System for developing potential groundwater maps (Arkoprovo et al., 2012; Lee et al., 2012). GIS and Remote Sensing technologies were utilized to map potential groundwater areas in the Musi watershed using thematic geo-environmental maps (Ganapuram et al., 2009). In Mul et al. (2007), the potential groundwater area of the South Pare Mountains Tanzania was identified using a geology map and spring chemical analysis data. Saraf et al. (2004) used a topographic map based on Digital Elevation Model (DEM) to analyze the groundwater potential in selected areas of West Bengal and Madhya Pradesh, India. They showed the relationship of groundwater potential with the intersection of drainage.

In recent decades, new artificial intelligence methods including Machine Learning (ML) have been developed and utilized to explore the potential of groundwater, the most popular of which are: Support Vector Machine (SVM) (Lee et al., 2018), Artificial Neural Network (ANN) (Lee et al., 2018), Classification and Regression Tree (CRT) (Naghibi et al., 2016), K-Nearest Neighbor (KNN) (Naghibi and Dashtpagerdi, 2017), Maximum Entropy (ME) (Rahmati et al., 2016), Functional Tree (FT) (Chen et al., 2018), Fisher's Linear Discriminant Function Analysis (FLDA) (Chen et al., 2019b), and Boosted Regression Tree (BRT) (Mousavi et al., 2017). Nowadays, hybrid ML approaches are being used more widely for the assessment and mapping of groundwater potential as the performance of these models is better in many cases in comparison to single ML models. Miraki et al. (Miraki et al., 2019) utilized a novel hybrid method, namely RF based on Random subspace to construct the map of groundwater potential of Qorveh-Dehgolan plain, Iran. The results showed that the hybrid method had a more precise predictive capacity for the groundwater potential than single ML models (Logistic Regression - LR, RF, and Naïve Bayes). Chen et al. (Chen et al., 2019a) applied a hybrid model based on FLDA with Bagging (BFLDA) and Rotation forest (RFLDA) to assess the potential of groundwater in the Ningtiaota area (China). The results indicated that the BFLDA model was better than the other FRLDA and FLDA models. Hossein et al. (Rizeei et al., 2019) applied a novel hybrid method based on MABLR - Multi-Adaptive Boosting Logistic Regression to build the map of the groundwater potential of the Gyeongsangbuk-do basin (South Korea) and compared the results with other models: LR, Multiple-Layer Perception (MPL), and SVM methods. The results indicated that the MABLR model was efficient in mapping the potential of groundwater.

In general, hybrid/ensemble ML models have shown better prediction of potential groundwater zones compared to single ML models. However, no known ML method can solve all groundwater problems, especially the assessment of groundwater potential in different regions (Bui et al., 2020). Therefore, an attempt has been made to apply advanced hybrid ML methods such as Random Forest (RF) and its Ensemble Framework (AdaBoost-ABRF, Bagging-BRF, and LogitBoost-LBRF) for the development of potential groundwater maps of the DakNong province, Vietnam, as model development is a continuous process for

improving predictive performance. This work is the first time LogitBoost ensemble with RF to develop a hybrid method for mapping the potential of groundwater. Weka and ArcGIS software was utilized for data processing and modeling.

2. Study area

The study area is the DakNong province, located between 11°45‘to 12°50’N latitude and 107°13’to 108°10‘E longitude (Fig. 1), covering two sub-regions of the southeastern and central part of Vietnam. The region's topography is relatively flat, with alternating highlands (plateau) divided by high mountains and low valleys running along the Serepok and Krong rivers. The average altitude of this region is approximately 650 m above sea level, and the highest is 1982 m. The area's climate is of tropical equatorial monsoon type with two main seasons: rainy season (April to November) and dry season (December to March next year). The mean temperature is 22-23°C, and the highest temperature is 35°C. The total annual precipitation in this area is about 2513 mm. Groundwater occurs in the DakNong province in three types of geological formations: Pliocene-Pleistocene Basalt Complex, Quaternary formation (alluvium), and Jurassic sedimentary rocks (Ha et al., 2021; Nguyen et al., 2020d). Following is a brief description of these three types of aquifer:

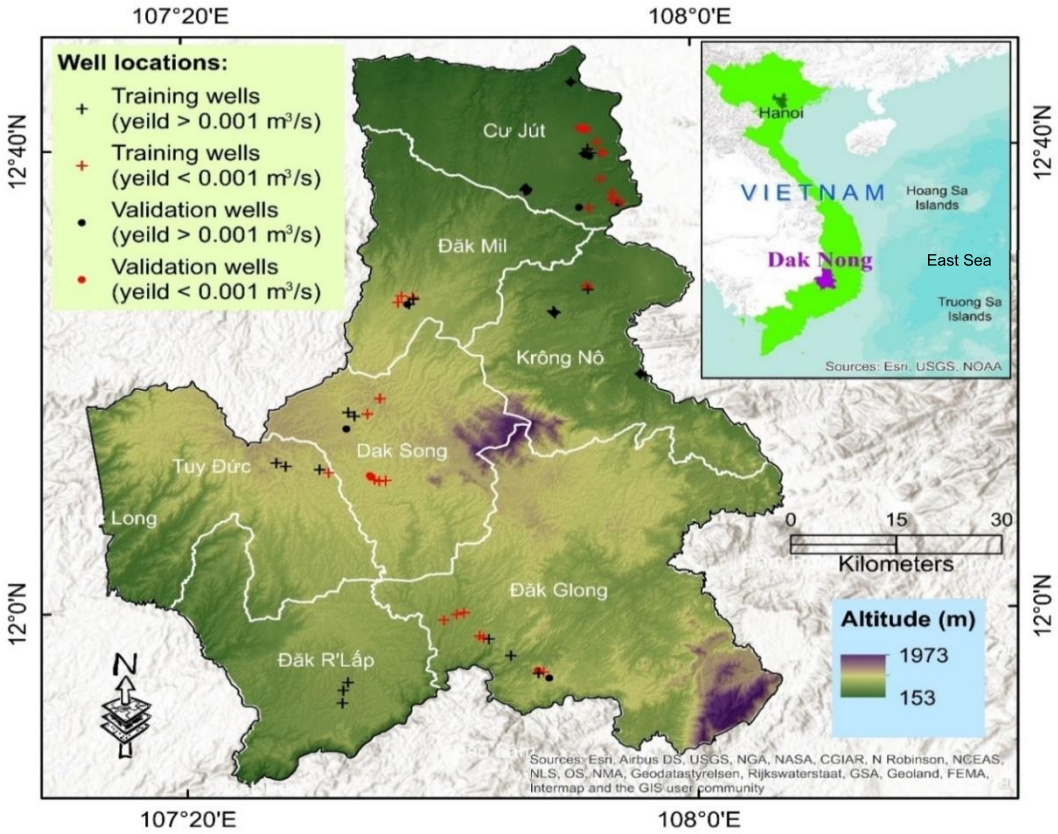


Figure 1. The study area with well locations

(i) The Quaternary alluvium aquifers (sand, grit, pebbles, gravel, and clay) occupy an area of about 27.16 square kilometers along major rivers and large waterways unconfined type. The average thickness of the aquifer is 5 to 7 m and a maximum of 20 m.

Water table depth ranges from 0.0 to 10.7 m, with an average depth of 2 to 4 m which varies with rainfall fluctuation. Water in the aquifer dries out during the dry season in many places. Water bearing capacity of the aquifer is limited, which is moderate to weak, and thus it can be used only for small residential areas.

(ii) The aquifer of the Pliocene-Pleistocene basalt complex includes basalt rocks that occupy around 3936.53 km^2 area, Its thickness ranges between 27 and 502 meters, and the mean thickness is approximately 100 m. Water in the Basalt Complex flows along with flow contacts, cracks, vesicles, and interconnected cavities in weathered basalt. Groundwater in this type of aquifer occurs mainly under confined conditions at moderate depth. Groundwater quality is good.

(iii)The Jurassic formation (sandstone, siltstone, limestone, and shale) aquifer covers an area of approximately 2116.78 km^2. Aquifer thickness varies from 17.5 to 79.6 m, with a mean thickness of 40 m. the water flows through cracks and flow contacts. The permeability of this aquifer and water quality are poor.

3. Materials and methods

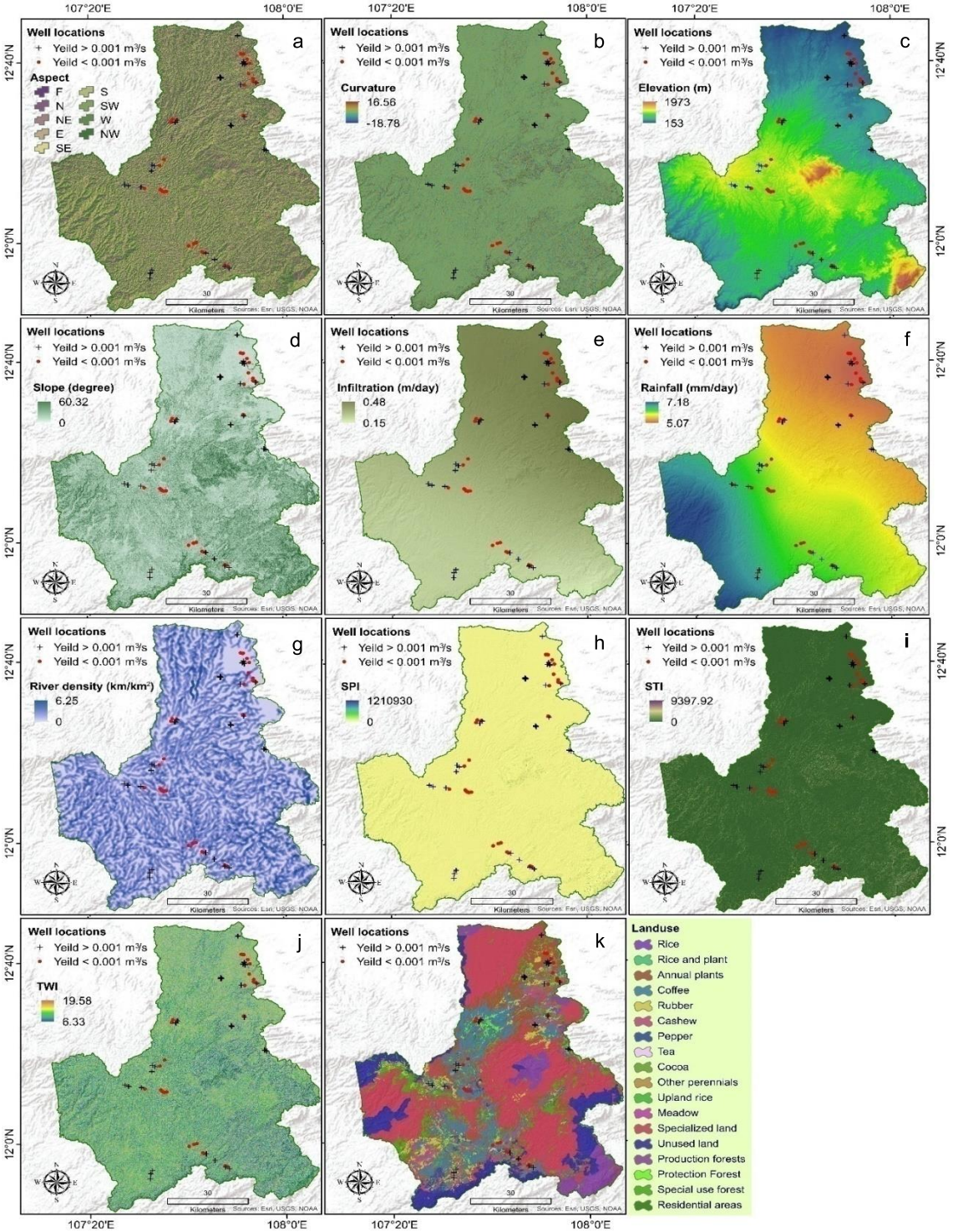
3.1. Building groundwater database

The groundwater inventory map of this research region was prepared from the spatial data of 72 wells (location, yield, aquifer characteristics, etc.) in conjunction with different thematic maps developed from the Digital Elevation Model (DEM), land use, geology, soil, and rainfall data obtained from various agencies and by conducting field surveys. Groundwater well data from 72 wells were used for the training and testing of the ML models in a 70:30 ratio (Chen et al., 2018; Oh et al., 2011). In this study, we have used the excellent yield of 0.001 m^3/s as a threshold value for the classification of high groundwater and lowly groundwater locations based on the expert analysis of the local groundwater (Ha et al., 2021; Nguyen et al., 2020d). Based on the local geo-environmental condition and data availability, twelve groundwater potential affecting factors (aspect, elevation, Sediment Transport Index (STI), Stream Power Index (SPI), Topographic Wetness Index (TWI), slope, curvature, soil, infiltration, land use, river density, and rainfall) were considered in the model study along with well yield data for groundwater modeling (Ha et al., 2021; Nguyen et al., 2020d; Senanayake et al., 2016; Souissi et al., 2018).

United States Geological Survey Aster DEM at the resolution of 30 m (https://earthexplorer.usgs.gov) was utilized for the extraction of topographical features (i.e., slope, aspect, curvature, and elevation) and hydrological features (STI, TWI, and SPI) and development of thematic maps. Soil and land use maps were compiled from the DakNong Ministry of Natural Resources and Environment and modified from Google Earth images. A rainfall map was prepared from the data of the Meteorology Department of Vietnam. Following is a brief description of the characteristics of these factors:

3.1.2. Topographic factors

The elevation is an essential factor for controlling the rainfall, soil formation, weathering, vegetation, and depth of percolation infiltration. Elevation in the area varies from 153 to 1973 m (Fig. 2). Slope (degree) controls the runoff thus infiltration in the aquifers (Chen et al., 2018). The study region varies from 0 to 60.32 degrees (Fig. 2). The aspect map indicates the direction of the slope face. It is related to groundwater potential because it determines the solar radiation amount falling on the surface of the Earth and impacts the precipitation. The aspect map was prepared from DEM and divided into categories (Fig. 2). Profile curvature and plan curvature maps were built from DEM as curvature influences runoff and infiltration (Shirzadi et al., 2017). A concave surface is more appropriate for storing surface water, thus helping infiltration/ recharge (Nguyen et al., 2020d). The curvature map of the region varies from -18.78 to 15.56 (Fig. 2).


107°20'E
108°0'E
12°40'N
12°0'N
Well locations
Yeild > 0.001 m³/s
Yeild < 0.001 m³/s
a
Aspect
F
S
N
SW
NE
W
E
NW
SE
b
Curvature
16.56
-18.78
c
Elevation (m)
1973
153
d
Slope (degree)
60.32
0
e
Infiltration (m/day)
0.48
0.15
f
Rainfall (mm/day)
7.18
5.07
g
River density (km/km²)
6.25
0
h
SPI
1210930
0
i
STI
9397.92
0
j
TWI
19.58
6.33
k
Kilometers
Sources: Esri, USGS, NOAA
Landuse
Rice
Rice and plant
Annual plants
Coffee
Rubber
Cashew
Pepper
Tea
Cocoa
Other perennials
Upland rice
Meadow
Specialized land
Unused land
Production forests
Protection Forest
Special use forest
Residential areas

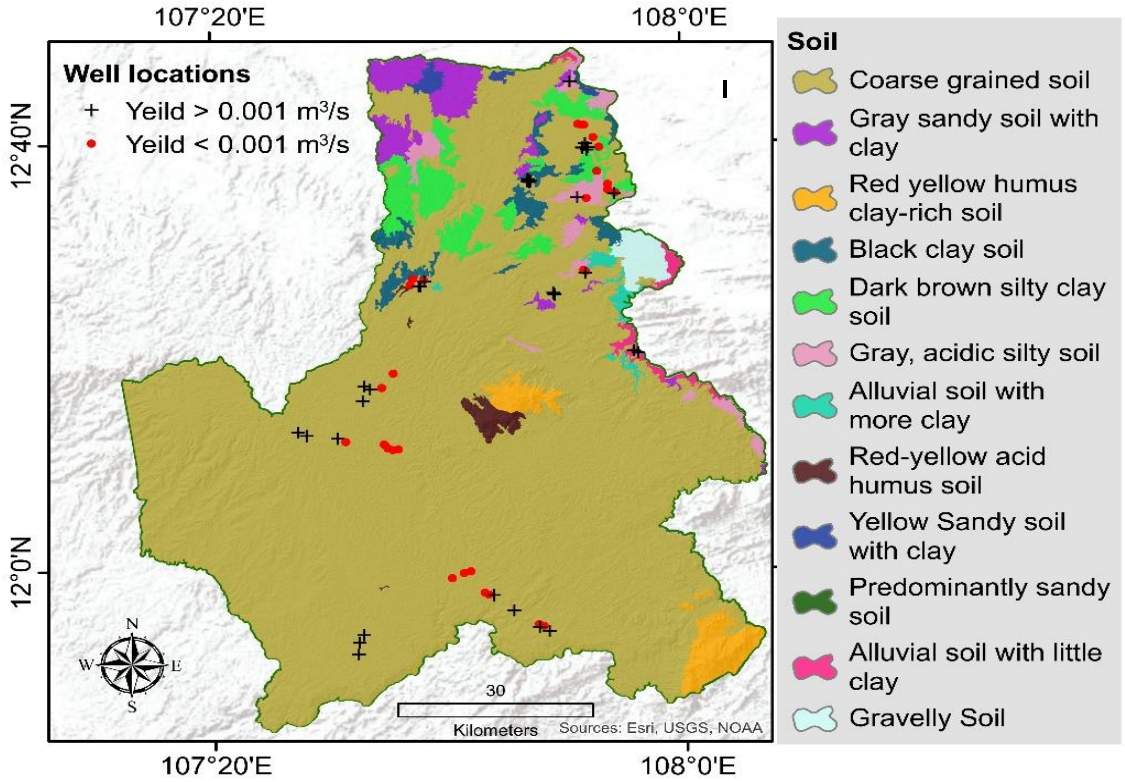


Figure 2. Maps of groundwater conditioning factors: (a) Aspect, (b) Curvature, (c) Elevation, (d) Slope, (e) Infiltration, (f) Rainfall, (g) River density, (h) SPI, (i) STI, (j) TWI, (k) Land use, (l) Soil

3.1.3 Hydrology factors

TWI is utilized to assess the effect of topography on water infiltration in the ground. It helps in quantifying the flow control and accumulation of the surface runoff (hydrological processes) (Elmahdy and Mohamed, 2014; Mokarram et al., 2015). In this study region, TWI ranges from 6.33 to 19.58 (Fig. 2).

SPI and STI help understand erosive processes resulting from the surface runoff and are proxies for the mid-scale topography (slope, valley bottom, or ridge) and the landscape flow capacity (Bourque and Bayat, 2015). In general, areas with higher STI and SPI values have a higher potential for groundwater potential as they indicate a higher water table. SPI and STI values vary from 0 to 1210930 and from 0 to 9397.92, respectively (Fig. 2). River density is a measure of the draining of watershed by stream channels. It is defined as the total drainage (river and streams) length per total drainage basin area. The river density map was classified from 0 to 6.25 km/km² (Fig. 2).

3.1.4. Other geo-environmental factors

Land use patterns affect the infiltration and consumption of water resources due to anthropological activities (Mafi-Gholami et al., 2019). The land use map of the region was divided into 17 classes based on types of cultivation, vegetation, land and built up, vegetation (rice, rice, and plant, annual plants, coffee, rubber, cashew, pepper, tea, cocoa, other perennials cultivation, upland rice), meadow, specialized land, unused land, production forests, protection

forest, particular use forest, and residential areas (Fig. 2).

The soil characteristics are essential in assessing the potential of groundwater in an area (Naghibi et al., 2017). In this study, a soil map was extracted from the data of the Soil Survey Department and classified into 12 types of soil: coarse grained soil, gray sandy soil with clay, red yellow humus clay-rich soil, black clay soil, dark brown silty clay soil, gray, acidic silty soil, alluvial soil with more clay, red-yellow acid humus soil, yellow sandy soil with clay, predominantly sandy soil, alluvial soil with little clay, and gravelly soil (Fig. 2). Infiltration depends on the characteristics of groundmass, including permeability and porosity of the underlying strata influencing the movement of surface water downward (Moghaddam et al., 2015). The ground permeability in this area varies from 0.15 to 0.48 m/day (Fig. 2). Rainfall directly helps recharge the groundwater through infiltration surface water in the groundmass. Recharge depends on the average annual rainfall of an area (Zaidi et al., 2015). The study region's daily rainfall ranges from 5.07 to 7.18 mm/day (Fig. 2).

3.2. Methods used

3.2.1. Random Forest (RF)

In an RF algorithm for the formation of each tree, a different set of available patterns is determined, taking into account the replacement of each selected pattern (Liaw and Wiener, 2002). The size of this sampled sample will equal the total number of available patterns. The random forest was described in 2001 by Breiman as a way of developing new decision-making trees, combining several unique algorithms' predictions using standard rules (Breiman, 2001). The general principles of group training techniques have relied on the assumption that their accuracy of them is higher than other training algorithms because the combination of various prediction models is more accurately than one single model, and groups increase the strength of individual and specific sets of classes, and they reduce class weaknesses at the same time (Avand et al., 2019; Shahabi et al., 2019b).

3.2.2. AdaBoost ensemble

AdaBoost is an ML algorithm with supervision. AdaBoost can combine a large number of learning algorithms with improving performance. The basic classifier used for the AdaBoost algorithm is only better than the random classifier, thus increasing the algorithm's performance with more repetitions. Even classifiers with an error higher than the random classifier improve overall performance by taking a negative coefficient. This method is sensitive to noisy data, has separate sections, and is less sensitive to over-adaptation issues than other learning algorithms (Pham et al., 2020). The basic idea of this algorithm is that each training sample is assigned a specific weight. First, the weight of all samples is the same, but in each iteration, the poorly trained structure provides classification, and the weight of the samples incorrectly classified by that class is increased. Thus, the focus of the algorithm is on hard-to-class samples. The final classification is made by a majority vote on the classifiers, in which the less erroneous classes are given more weight (An and Kim, 2010).

3.2.3. Bagging Ensemble

To improve the accuracy of basic methods such as decision trees, it is recommended to use integrated methods. It has been proven

that a combined method's correctness is often better than any of the components (Breiman, 1996; Opitz and Maclin, 1999). One of the proposed methods in the field of data mining is the Bagging method. In this method, several categories of data, for example, several decision trees, are created from the data, and all announce their opinions. The final decision is based on a majority vote (Maclin and Opitz, 1998). More specifically, if the D data set, which includes groundwater information, is considered, the Bearing method works so that it forms the base number of N categories, each of which is denoted by Ci (1). In repeating i, the algorithm selects the Di training set with the help of sampling by replacing the D set (Nhu et al., 2020). Because the sampling method is alternative, some groundwater location is repeated several times in sample Di, while in others, they may not be present in this sample. Di is then generated as a training set for the Ci model i used and showed its vote for the test data. This is done for all i. Then, the test data class is determined on the base of the maximum number of votes cast (presence or absence of groundwater). Of course, this is done once for each test data to determine the target class for all of them (Avand et al., 2020a; Barzegar et al., 2019).

3.2.4. LogitBoost Ensemble

Boosting is presented to combine several algorithms and improve predictive performance. The structure of this model is a logical and generalized structure of the famous multiple logit model, which can estimate any model with random desirability, and the three significant shortcomings of the multiple logit model considering random disagreement, unlimited succession pattern, and dependence on unseen factors in time have passed (Avand et al., 2020a). It notes that this model, unlike the standard Logit and Probit models, is not limited to a specific distribution and can find heterogeneity in factors and the source of this heterogeneity. In the case of discrete selection modeling, the desirability function defined for the n decision maker is defined as Equation 1 to select i option from the available selection set (Jou et al., 2011; Li et al., 2010).

$$U_{ni} = V_{ni} + \varepsilon_{ni} \quad (1)$$

V_{ni} is defined as the specific desirability (observable) of option i for person n and is defined as the indefinite and random (invisible) part of the desirability of option i for person n. In the ensemble logit model, the invisible part of the utility function ( ) includes two parts. The first part includes a custom distribution, and the second one, like the standard Logit model, consists of the distribution of a limit value with independent and identical distribution, so it imposes fewer assumptions on the data (Hess, 2005; Train, 2009). The general form of the combined logit model is as follows:

$$P_{ni} = \int L_{ni}(\beta) f(\beta) d\beta \quad (2)$$

where, P_{ni} is defined as the probability of the choice of option i by the individual n, and L_{ni} is the probability of the choice of the option i by the person n in the Logit model.

3.2.5. Validation methods

Various criteria are used to evaluate the ML models (Avand et al., 2019); (Janizadeh et al., 2019). One of the most important criteria used for the validation is the confusion matrix (Table 1). This matrix shows how the categorization algorithm works according to the input data set by the different categories of the categorization problem (Visa et al., 2011).

Table 1. Confusion matrix used

		Predicted	
Actual		Positive	Negative
	Positive	TP	FN
	Negative	FP	TN

Where TP (True Positive) indicates the number of records whose real category is positive and the classification algorithm correctly identifies their category, TN (True Negative) is the number of records whose real category is negative. The classification algorithm has correctly identified their category, FN (False Negative) indicates the number of records whose actual category is positive and their category categorization algorithm misdiagnosis has given, FP (False Positive) indicates the number of records whose real category is negative and the category categorization algorithm has misdiagnosed them positively. Other statistical criteria used to evaluate model results include: Accuracy (ACC), Specificity (SPF), Negative Predictive Value (NPV), Sensitivity (SST), and Positive Predictive Value (PPV), whose equations are as follows:

$$CC = (TP + TN)/(TP + FN + FP + TN) \quad (3)$$

$$SST = TP/TP + FN \quad (4)$$

$$SPF = TN/FP + TN \quad (5)$$

$$PPV = TP/TP + FP \quad (6)$$

$$NPV = TN/FN + TN \quad (7)$$

Root Mean Square Error (RMSE) which measures the error rate between two data sets, was also utilized to evaluate and compare the proposed models (Hadzima-Nyarko and Trinh, 2022; Kumar, 2022; Tran et al., 2022). This parameter usually compares the predicted values and the measured values (Le et al., 2020). RMSE number represents the mean of the available errors, and when our goal is to evaluate the accuracy of the total data, this number can be used as an important indicator (Chai and Draxler, 2014), and its equation is as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2} \quad (8)$$

where $\frac{1}{N}\sum$ performs the averaging operation and $(x_i - \hat{x}_i)^2$ calculates the square error of each data.

Kappa is used to evaluate the percentage of agreement between two people when assessing a certain phenomenon after eliminating the role of luck factors. If both agree on the idea in all cases, k = + 1. If the observed agreement rate between the two > the expected agree rate then k > 0. If the observed agreed rate is less than or equal to the chance by chance then k < 0 (Chen et al., 2019b; Nguyen et al., 2020c). It can be calculated by following equation:

$$K = \frac{P_p - P_{exp}}{1 - P_{exp}} \quad (9)$$

where P_p is defined as the rate of instances predicted correctly for groundwater or non-groundwater. P_{exp} is defined as expected agreements.

ROC curve: One of the most important criteria for determining the efficiency of a category is the use of the level criterion below the ROC curve. The AUC is the area below the ROC, and the higher the value of the AUC, the more optimal performance of the model (Shahabi et al., 2019a). ROC curve is a way to evaluate ML models' performance quantitatively (McClish, 1989). The value of AUC for a category that randomly determines the sample category is 0.5. Also the maximum value of this criterion is equal to one, and it occurs for a situation where the category is ideal and can detect all positive samples without any false alarms (Nguyen et al., 2020a; Nhu et al., 2020).

3.2.6. OneR feature selection

OneR is an abbreviation of "One Rule" that generates a rule for each predictor in the dataset and then selects the rule with the least total error as its "one rule" (Morariu et al., 2005). Using machine learning methods increases the quality of selected features and also increases the quality of learning. In this method, each feature is evaluated separately using 1R classifier. In this classification, the rules are expressed based on the value of the

property and the subject (Avand et al., 2020a). If only those features are used to classify the data, it will cause an error. Then, the feature that has the lowest error value will be selected. So, this method sorts the properties (each feature independently) based on their error value and finally selects the most important ones (Nguyen et al., 2020a).

3.2.7. Research methodology

In this research, the methodological steps of potential groundwater maps are described briefly below (Fig. 3):

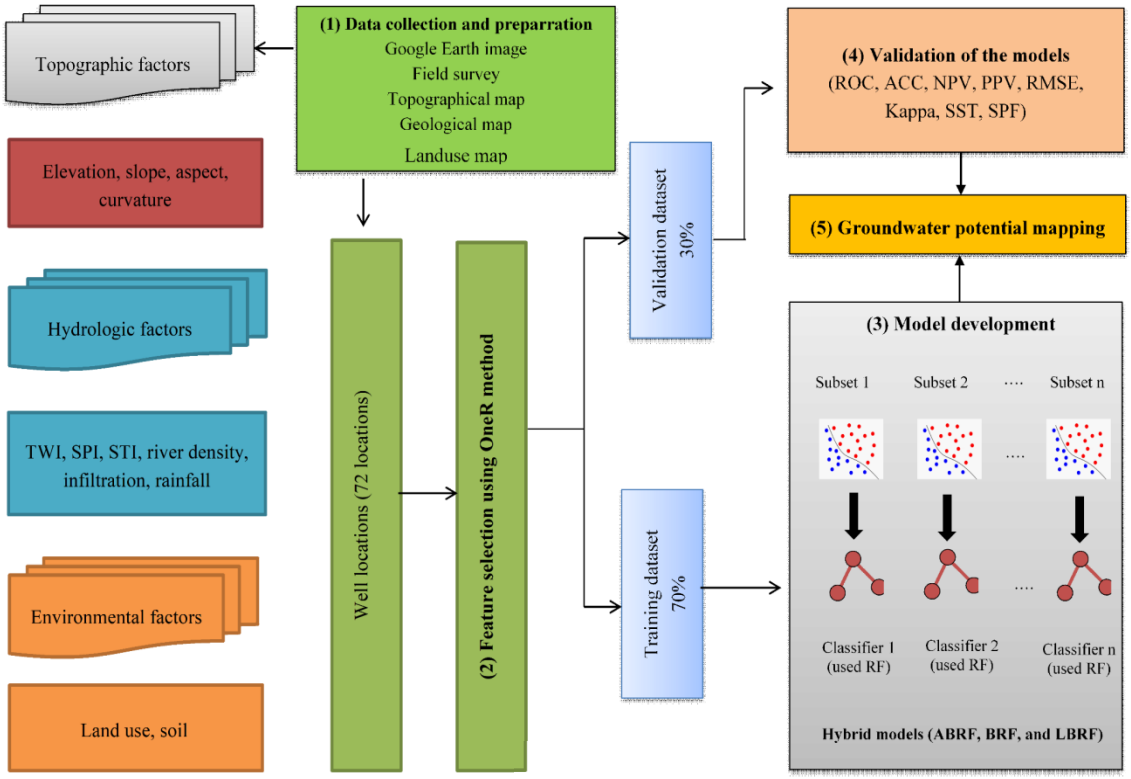


Figure 3. Methodological flow chart

Collection of geospatial data of groundwater wells, meteorology, and other thematic maps was done from various government and non-government organizations, websites, and field works. The geospatial database was divided into two parts at 70:30 for generating training and testing datasets for building and validating models, respectively. The yield of 0.001 cubic meters per second was used as a threshold value to differentiate potential and non-potential groundwater groups.

Feature selection: The oneR method was utilized to evaluate the input factors' importance and choose the suitable factors for modeling groundwater potential.

Development of the models: RF and its ensembles (BRF, ABRF, and LBRF) were developed using a training dataset for groundwater potential modeling and mapping. Out of these models, BRF was constructed by combing RF and Bagging, which Bagging utilized to optimize the training dataset for the classification of groundwater and non-potential

groundwater classes. ABRF was built by searching RF and AdaBoost ensemble, which AdaBoost used to optimize the training dataset for classifying groundwater and non-potential groundwater classes. LBRF was constructed by combing RF and LogitBoost ensemble, which LogitBoost was used to optimize the training dataset for classification of potential groundwater and non-potential groundwater classes.

Evaluation of the models: Statistical indices including SST, SPF, PPV, NPV, AUC, RMSE, and Kappa were calculated using the values obtained from the confusion matrix utilized to evaluate and compare the predictive capability of the proposed models on both training and testing datasets.

Construction of potential groundwater maps: Utilizing the trained models, the groundwater potential maps were constructed with various classes namely very low, low, moderate, high, and very high zones.

4. Results

4.1. Importance of input variables using One Rule method

Significant and insignificant factors in predicting potential groundwater zones using different models were validated based on the One Rule algorithm (Table 2). The results show that rainfall is the most critical factor (AM: 68.039), whereas river density is the least important factor (AM: 53.969) in building the models. Other factors (soil, SPI, infiltration, STI, land use, curvature, TWI, aspect, slope degree, and elevation) ranked second to eleventh, respectively, in the importance of variable factors for the generation groundwater potential zone maps.

Table 2. Selection of attribute feature using One R attribute evaluation method

Rank	Average Merit	Error	attribute
1	68.039	4.411	Rainfall
2	65.659	3.582	Soil
3	62.878	3.73	SPI
4	61.298	5.201	Infiltration
5	59.89	6.391	STI
6	59.906	3.952	Landuse
7	58.737	6.002	Curvature
8	58.302	6.428	TWI
9	55.973	8.883	Aspect
10	55.78	6.595	Slope
11	55.141	4.707	Elevation
12	53.969	6.282	River density

4.2. Validation of model performance

The predictive capability of the models was evaluated using different statistical indexes for predicting potential groundwater zones. In the training step, LBRF and RF algorithms have almost the same values in all indexes. However, the accuracy of the RF model is higher based on RMSE (0.929) and K (0.227) indices. In comparison to ABRF and BRF models, the ABRF has the highest NPV (96.55) and SST (96.15) value, thus better model performance (Table 3). In the testing phase, the LBRF has the lowest RMSE (0.459) and the highest PPV (72.73), NPV (84.62), SST (80.00), SPF (78.57), ACC (79.17), and K (0.459) (Table 4). Therefore, the LBRF has the highest accuracy among other models used in the present study in the validation phase.

Table 3. Validation of the models using training dataset

No	Parameters	ABRF	BRF	LBRF	RF
1	TP	25	26	27	27
2	TN	28	25	27	27
3	FP	2	1	0	0
4	FN	1	4	2	2
5	PPV (%)	92.59	96.30	100.00	100.00
6	NPV (%)	96.55	86.21	93.10	93.10
7	SST (%)	96.15	86.67	93.10	93.10
8	SPF (%)	93.33	96.15	100.00	100.00
9	ACC (%)	94.64	91.07	96.43	96.43
10	K	0.8926	0.8219	0.9287	0.929
11	RMSE	0.1986	0.2806	0.2436	0.227

Table 4. Validation of the models using testing dataset

No	Parameters	ABRF	BRF	LBRF	RF
1	TP	7	7	8	6
2	TN	11	9	11	9
3	FP	4	4	3	5
4	FN	2	4	2	4
5	PPV (%)	63.64	63.64	72.73	54.55
6	NPV (%)	84.62	69.23	84.62	69.23
7	SST (%)	77.78	63.64	80.00	60.00
8	SPF (%)	73.33	69.23	78.57	64.29
9	ACC (%)	75.00	66.67	79.17	62.50
10	K	0.489	0.329	0.578	0.2394
11	RMSE	0.475	0.466	0.459	0.478

Analysis of the ROC curve in this study indicates that all four models: RF (0.987), ABRF (0.992), BRF (0.967), and LBRF (0.976) models have excellent predictive accuracy in the training phase, whereas, on validation datasets, the LBRF has the highest precision in predicting the potential of groundwater with a value of AUC_{LBRF}= 0.776, followed by the accuracy of the ABRF (AUC= 0.766), BRF (AUC= 0.745) and RF (AUC= 0.734) models (Fig. 4).

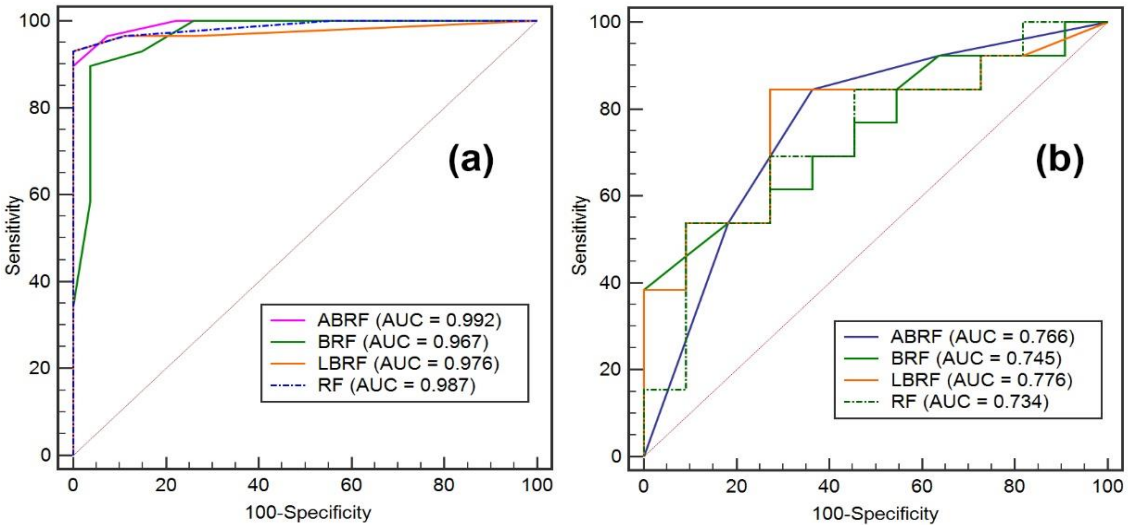


Figure 4. Plots of the ROC curves of the models a: training datasets and b: validation datasets

4.3. Groundwater potential mapping

Groundwater potential maps were constructed using one single (RF) and three ensemble models (ABRF, BRF, and LBRF). Developed maps were classified into various possible categories: shallow, low, moderate, high, and very high (Fig. 5) using a natural break method (El-Hoz et al., 2014). In Fig. 6, the highest percentage of wells with higher yield (yield> 0.001 m^3/s) falls in the high category of groundwater potential zones map developed from the ABRF model compared to LBRF, ABRF, BRF, and RF models, respectively. In contrast, the opposite is the case of wells having lesser yield (yield < 0.001 m^3/s). The maps were validated using the frequency ratio of wells with higher yield (yield > 0.001 m^3/s) falling in potentially very high groundwater classes. It can be stated that the accuracy of the groundwater zones in the maps developed by models is good. Therefore, the constructed maps could be used for water resource management in the study area.

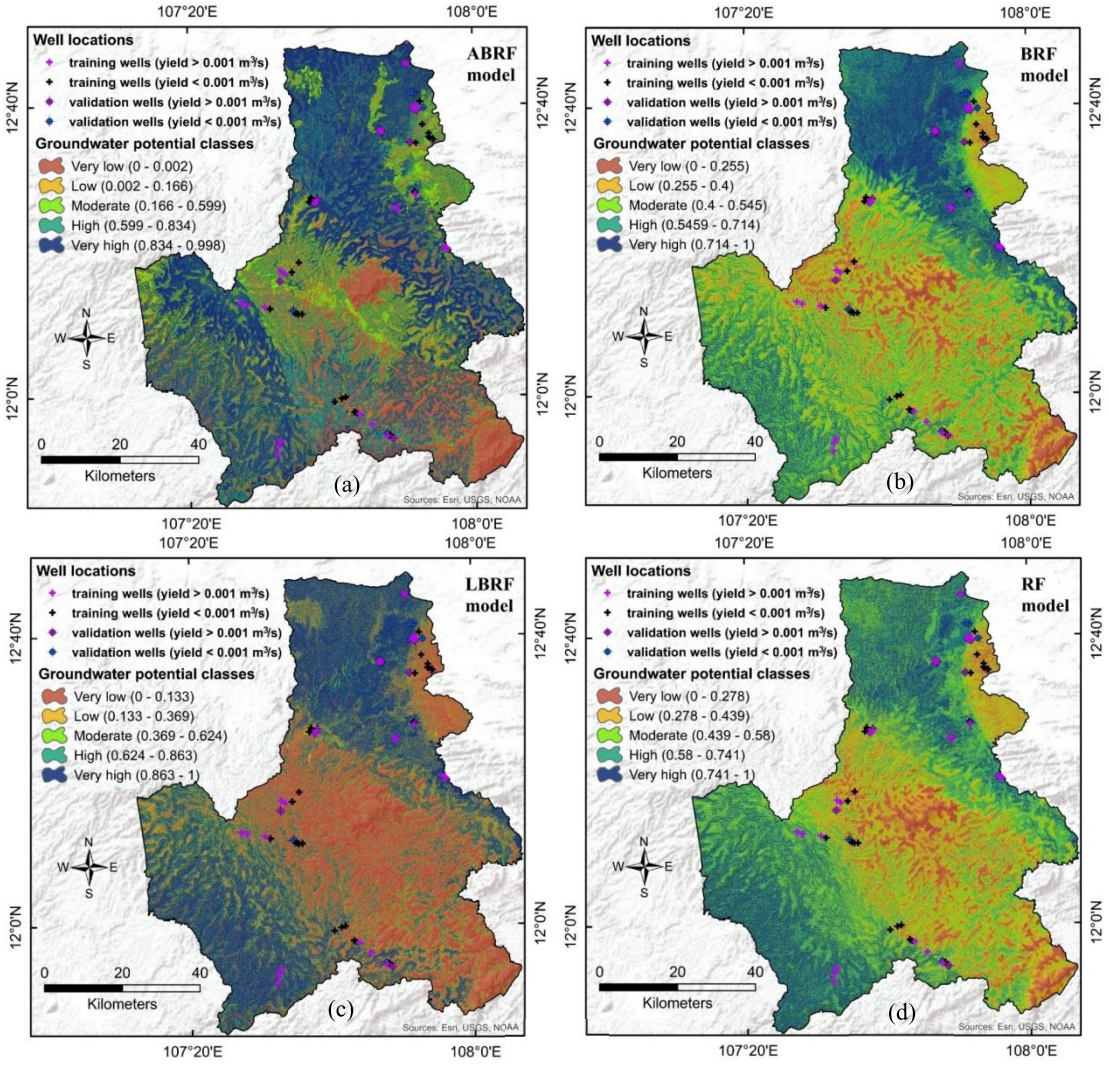


Figure 5. Maps of groundwater potential generated from the models: (a) ABRF, (b) BRF, (c) LBRF, (d) RF

5. Discussions

Overexploitation of ground resources due to increasing population, industrial growth, and urbanization necessitates a systematic assessment of groundwater potential for optimum utilization by proper management (Seenipandi et al., 2013). The development of accurate groundwater potential zone maps (Vadiati et al., 2018) using new technologies such as advanced hybrid/ensemble ML methods is an essential step.

In the model study, twelve influencing factors (e.g., soil, TWI, infiltration, curvature, land use, aspect, river density, slope, SPI, STI, elevation, and rainfall) were considered. Their importance was evaluated using the One Rule methods value to show that rainfall (average daily), soil, and SPI, with values of 68.039, 65.659, and 62.878, respectively, are the critical factors for mapping groundwater potential (Table 2). River density (53.969) is the most negligible significant factor. Generally, factors with the highest AM values are

the most critical and essential variables for developing groundwater models (Vadiati et al., 2018). However, other factors with lower AM values can also be helpful for model development. Thus, based on the evaluation of parameters, we decided to utilize all factors for the groundwater modeling. A review of the relevant literature shows that the importance of groundwater influencing factors depends on the local ground conditions and geo-environmental conditioning factors, which may vary from one region to another. However, the identification of the minor critical variables by the One-R (AM value) in the present study is in line with Moeck et al. (2020) and Pourghasemi et al. (Pourghasemi et al., 2020) and Nguyen et al. (Nguyen et al., 2020b). Examples of differences in the selection of the most critical factors influencing different results can be observed in many studies. Chen et al. (2019a) presented that elevation, SPI, and lithology were the most important factors for mapping groundwater potential in the Ningtiaota region (China). Avand et al. (2020b) selected land use, lithology, and rainfall as the essential factors in the Yasuj-Dena area, Iran. The rainfall factor is directly helpful in recharging the groundwater and is thus very important in many studies (Zhang et al., 2019) as in the present study.

Validation results of models revealed that the LBRF method has high predictive accuracy with the highest values of PPV (72.73%), NPV (84.62%), SST (80.00%), SPF (78.57%), ACC (79.17%), AUC (0.776) and K (0.578) indices, and lowest RMSE (0.459), followed by ABRF, BRF and RF methods. Thus, the predictive capability of the LBRF model is the best in mapping groundwater potential compared to other models. In general, the results show that hybrid ML models have improved the predictive capability of a single RF model. This is in line with the results of the previous works of other researchers (Nguyen et al., 2020b). The advantage of the RF algorithm compared to other individual algorithms is the ability to deal with large amounts of data without removing covariates (Naghibi et al., 2017). During the training phase, the RF algorithm uses the maximum set of specific trees. Thus it can produce a large number of classification trees for better modeling performance (Catani et al., 2013). Moreover, there are other advantages of hybrid models more than a single simple model, such as Naïve Bayes tree integrated with Random Subspace (Shirzadi et al., 2017), ADT combined with AdaBoost (Tien Bui et al., 2019), FLDA integrated with Bagging (Miraki et al., 2019), RF integrated with Random Subspace (Binh Thai et al., 2019). Compared with other similar studies (Ha et al., 2021; Nguyen et al., 2020d) observed that the LBRF algorithm has a higher performance than another hybrid model such as ABQDA (AUC = 0.741) and similar performance to the RABANN model (AUC = 0.776). Therefore, our new study of groundwater potential mapping using RF and its ensemble/ hybrid frameworks (ABRF, BRF, and LBRF) confirm the possible use of new ensemble models in obtaining better predictive accuracy in groundwater modeling for the development of accurate maps of groundwater potential (Fig. 6). The map constructed by the ABRF model indicated that 42.47% of the area is covered by a very high potential class, representing that this part of the study area has a higher potential for groundwater productivity, which is consistent with the results of Pham et al. (2019).

The potential intermediate class occupies about 16% of the area, and the remaining 22.87% falls into the shallow likely class. The reliability analysis of the maps was performed using the frequency ratio method. The results show that most of the high-performance well yield areas can be correlated with very high groundwater potential classes of the

developed maps (Fig. 6). This indicates that groundwater potential zones classified based on the hybrid models are reliable and accurate, which is in line with the other similar studies (Nguyen et al., 2020a; Pham et al., 2019).

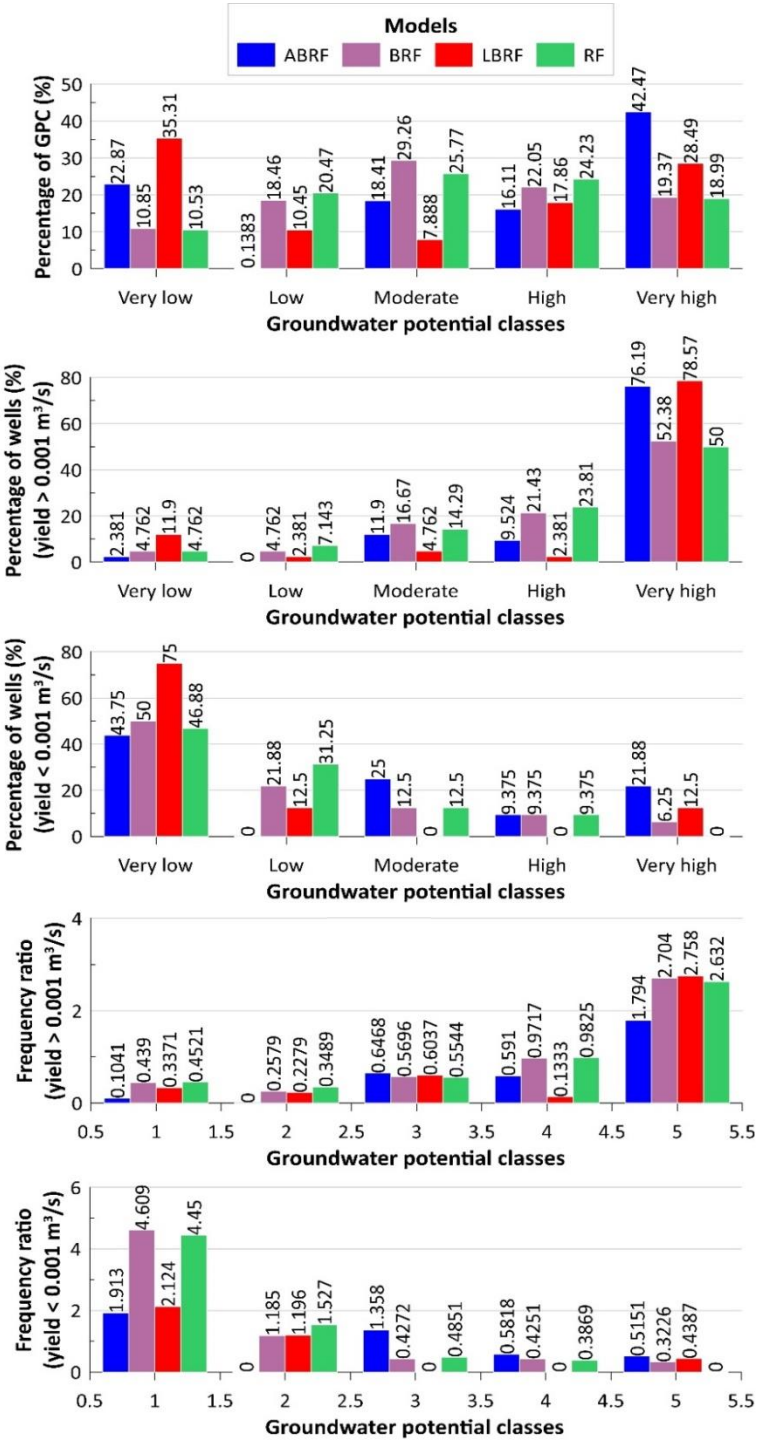


Figure 6. Analysis of groundwater potential maps

6. Concluding remarks

Water scarcity in this decade is a significant challenge mainly due to inadequate mapping, planning, and management of groundwater resources. The main aim of the present work was to map the groundwater potential of the Dak Nong Province, Vietnam using RF and its ensemble framework, namely BRF, ABRF, and LBRF. For this, twelve groundwater potential conditioning factors, namely topography (aspect, TWI, curvature, slope, and elevation), soil, land-use, hydrology (infiltration and river density, SPI, STI), and rainfall were used to develop the models. Well, yield data was used to develop and validate different groundwater potential zones. One-Method was utilized to prioritize the importance of groundwater potential affecting factors. In the present work, rainfall is the essential factor (AM: 68.039), whereas river density is the least important (AM: 53.969) in building the models.

The performance of the proposed models was evaluated and compared using various statistical indexes on training and testing phases for mapping groundwater potential. The results showed that the hybrid models (LBRF, BRF, and ABRF) outperform the single RF model. Among those, the LBRF has the highest precision in predicting the potential of groundwater with a value of AUC_{LBRF}= 0.776, followed by the accuracy of the ABRF (AUC= 0.766), BRF (AUC = 0.745), and RF (AUC= 0.734) models. Therefore, the hybrid model LBRF is a promising tool for assessing the groundwater potential in other regions.

The limitation of the study is that geological factors have not been considered in the model study, as the surface is covered mainly by sandy soil. In future research, sub-surface geology will be regarded to refine the input parameters of the models.

Acknowledgments

We thank Vietnam Academy for Water Resources for providing the data under the state project Code ĐTĐL.CN-69/21 to carry out this research.

Conflict of interest: The authors declare that there is no conflict of interest.

References

An T.-K., Kim M.-H., 2010. A new diverse AdaBoost classifier, 2010 International Conference on Artificial Intelligence and Computational Intelligence. IEEE, 359-363.

Arkoprovo B., Adarsa J., Prakash S.S., 2012. Delineation of groundwater potential zones using satellite remote sensing and geographic information system techniques: a case study from Ganjam district. Orissa. India. Research Journal of Recent Sciences, 1(9), 59-66.

Avand M., et al., 2020a. A Tree-based Intelligence Ensemble Approach for Spatial Prediction of Potential Groundwater. International Journal of Digital Earth, 13(12), 1408-1429.

Avand M., et al., 2019. A Comparative Assessment of Random Forest and k- Nearest Neighbor Classifiers for Gully Erosion Susceptibility Mapping. Water, 11(10), 2076.

Avand M., et al., 2020b. A tree-based intelligence ensemble approach for spatial prediction of potential groundwater. International Journal of Digital Earth, 1-22.

Barzegar R., Asghari Moghaddam A., Adamowski J., Nazemi A., 2019. Delimitation of groundwater zones under contamination risk using a bagged ensemble of optimized DRASTIC frameworks. Environmental Science and Pollution Research, 26, 1-15.

Binh Thai P., et al., 2019. A Novel Intelligence Approach of a Sequential Minimal Optimization-Based Support Vector Machine for Landslide

Susceptibility Mapping, Sustainability, 11(22), 6323.

Bonham-Carter G.F., 2014. Geographic information systems for geoscientists: modelling with GIS. Elsevier.

Bourque C.P.-A., Bayat M., 2015. Landscape variation in tree species richness in northern Iran forests. PloS one, 10(4), e0121172.

Breiman L., 1996. Bagging predictors. Machine Learning, 24, 123-140.

Breiman L., 2001. Random forests. Machine Learning, 45, 5-32.

Bui Q.-T., et al., 2020. Verification of novel integrations of swarm intelligence algorithms into deep learning neural network for flood susceptibility mapping. Journal of Hydrology, 581, 124379.

Catani F., Lagomarsino D., Segoni S., Tofani V., 2013. Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. Natural Hazards and Earth System Sciences, 13, 2815.

Chai T., Draxler R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7, 1247-1250.

Chen W., et al., 2018. GIS-based groundwater potential analysis using novel ensemble weights-of-evidence with logistic regression and functional tree models. Science of the Total Environment, 634, 853-867.

Chen W., Pourghasemi H.R., Zhao Z., 2017. A GIS-based comparative study of Dempster-Shafer, logistic regression and artificial neural network models for landslide susceptibility mapping. Geocarto International, 32, 367-385.

Chen W., et al., 2019a. Novel Hybrid Integration Approach of Bagging-Based Fisher's Linear Discriminant Function for Groundwater Potential Analysis. Natural Resources Research, 28(4), 1239-1258.

Chen W., et al., 2019b. Novel hybrid integration approach of bagging-based fisher's linear discriminant function for groundwater potential analysis. Natural Resources Research, 28, 1239-1258.

El-Hoz M., Mohsen A., Iaaly A., 2014. Assessing groundwater quality in a coastal area using the GIS technique. Desalination and Water Treatment, 52, 1967-1979.

Elmahdy S.I., Mohamed M.M., 2014. Groundwater potential modelling using remote sensing and GIS: a case study of the Al Dhaid area, United Arab Emirates. Geocarto International, 29, 433-450.

Ganapuram S., Kumar G.T.V., Krishna I.V.M., Kahya E., Demirel M.C., 2009. Mapping of groundwater potential zones in the Musi basin using remote sensing data and GIS. Advances in Engineering Software, 40, 506-518.

Ha D.H., et al., 2021. Quadratic discriminant analysis based ensemble machine learning models for groundwater potential modeling and mapping. Water Resources Management, 35, 4415-4433.

Hadzima-Nyarko M., Trinh S.H., 2022. Prediction of compressive strength of concrete at high heating conditions by using artificial neural network-based Bayesian regularization. Journal of Science and Transport Technology, 2, 9-21.

Hess S., 2005. Advanced discrete choice models with applications to transport demand.

Janizadeh S., et al., 2019. Prediction Success of Machine Learning Methods for Flash Flood Susceptibility Mapping in the Tafresh Watershed, Iran. Sustainability, 11, 5426.

Jha M.K., Chowdhury A., Chowdary V., Peiffer S., 2007. Groundwater management and development by integrated remote sensing and geographic information systems: prospects and constraints. Water Resources Management, 21, 427-467.

Jha M.K., Kamii Y., Chikamori K., 2009. Cost-effective approaches for sustainable groundwater management in alluvial aquifer systems. Water resources management, 23, 219.

Jou R.-C., Hensher D.A., Hsu T.-L., 2011. Airport ground access mode choice behavior after the introduction of a new mode: A case study of

Taoyuan International Airport in Taiwan. Transportation Research Part E: Logistics and Transportation Review, 47, 371-381.

Kumar R., 2022. Prediction and sensitivity analysis of self compacting concrete slump flow by random forest algorithm. Journal of Science and Transport Technology, 2, 32-43.

Le H.-A., Nguyen T.-A., Nguyen D.-D., Prakash I., 2020. Prediction of soil unconfined compressive strength using Artificial Neural Network Model. Vietnam Journal of Earth Sciences, 42, 255-264.

Lee S., Hong S.-M., Jung H.-S., 2018. GIS-based groundwater potential mapping using artificial neural network and support vector machine models: the case of Boryeong city in Korea. Geocarto international, 33, 847-861.

Lee S., Song K.-Y., Kim Y., Park I., 2012. Regional groundwater productivity potential mapping using a geographic information system (GIS) based artificial neural network model. Hydrogeology Journal, 20, 1511-1527.

Li H., Huang H., Liu J., 2010. Parameter Estimation of the Mixed Logit Model and Its Application. Journal of Transportation Systems Engineering and Information Technology, 10, 73-78.

Liaw A., Wiener M., 2002. Classification and regression by randomForest. R news, 2, 18-22.

Maclin R., Opitz D., 1998. An Empirical Evaluation of Bagging and Boosting. Proceedings of the National Conference on Artificial Intelligence.

Mafi-Gholami D., Zenner E.K., Jaafari A., Bakhtiari H.R., Bui D.T., 2019. Multi-hazards vulnerability assessment of southern coasts of Iran. Journal of environmental management, 252, 109628.

McClish D., 1989. Analyzing a portion of the ROC Curve. Medical decision making : an international journal of the Society for Medical Decision Making, 9, 190-5.

Miraki S., et al., 2019. Mapping Groundwater Potential Using a Novel Hybrid Intelligence Approach. Water Resources Management, 33, 281-302.

Moeck C., et al., 2020. A global-scale dataset of direct natural groundwater recharge rates: A review of variables, processes and relationships. Science of The Total Environment, 717, 137042.

Moghaddam D.D., Rezaei M., Pourghasemi H., Pourtaghie Z., Pradhan B., 2015. Groundwater spring potential mapping using bivariate statistical model and GIS in the Taleghan watershed, Iran. Arabian Journal of Geosciences, 8, 913-929.

Mokarram M., Roshan G., Negahban S., 2015. Landform classification using topography position index (case study: salt dome of Korsia-Darab plain, Iran). Modeling Earth Systems and Environment, 1(4), 1-7.

Morariu D., Vintan L., Tresp V., 2005. Meta-classification using SVM classifiers for text documents. Intl. Jrnl. of Applied Mathematics and Computer Sciences, 1(1), 15-20.

Mousavi S.M., Golkarian A., Naghibi S.A., Kalantar B., Pradhan B., 2017. GIS-based groundwater spring potential mapping using data mining boosted regression tree and probabilistic frequency ratio models in Iran. Aims Geosci, 3, 91-115.

Mul M.L., Mutiibwa R.K., Foppen J.W.A., Uhlenbrook S., Savenije H.H.G., 2007. Identification of groundwater flow systems using geological mapping and chemical spring analysis in South Pare Mountains, Tanzania. Physics and Chemistry of the Earth, Parts A/B/C, 32, 1015-1022.

Naghibi S.A., Ahmadi K., Daneshi A., 2017. Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. Water Resources Management, 31, 2761-2775.

Naghibi S.A., Dashtpagerdi M.M., 2017. Evaluation of four supervised learning methods for groundwater spring potential mapping in Khalkhal region (Iran) using GIS-based features. Hydrogeology Journal, 25, 169-189.

Naghibi S.A., Pourghasemi H.R., Dixon B., 2016. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. Environmental monitoring and assessment, 188, 44.

Nguyen P., et al., 2020a. Soft Computing Ensemble Models Based on Logistic Regression for Groundwater Potential Mapping. Applied Sciences, 10, 2469.

Nguyen P., et al., 2020b. Groundwater Potential Mapping Combining Artificial Neural Network and Real AdaBoost Ensemble Technique: The DakNong Province Case- study, Vietnam. International Journal of Environmental Research and Public Health, 17, 2473.

Nguyen P.T., et al., 2020c. Soft Computing Ensemble Models Based on Logistic Regression for Groundwater Potential Mapping. Applied Sciences, 10, 2469.

Nguyen P.T., et al., 2020d. Groundwater potential mapping combining artificial neural network and real AdaBoost ensemble technique: the DakNong province case-study, Vietnam. International Journal of Environmental Research and Public Health, 17, 2473.

Nhu V.-H., et al., 2020. Shallow Landslide Susceptibility Mapping by Random Forest Base Classifier and its Ensembles in a Semi-Arid Region of Iran. Forests, 11, 421.

Oh H.-J., Kim Y.-S., Choi J.-K., Park E., Lee S., 2011. GIS mapping of regional probabilistic groundwater potential in the area of Pohang City, Korea. Journal of Hydrology, 399, 158-172.

Opitz D., Maclin R., 1999. Popular Ensemble Methods: An Empirical Study, 11, 169-198.

Pham B.T., et al., 2020. GIS Based Hybrid Computational Approaches for Flash Flood Susceptibility Assessment. Water, 12, 683.

Pham B.T., et al., 2019. Hybrid computational intelligence models for groundwater potential mapping. Catena, 182, 104101.

Pourghasemi H.R., et al., 2020. Using machine learning algorithms to map the groundwater recharge potential zones. Journal of Environmental Management, 265, 110525.

Rahmati O., Pourghasemi H.R., Melesse A.M., 2016. Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: a case study at Mehran Region, Iran. Catena, 137, 360-372.

Rizeei H.M., Pradhan B., Saharkhiz M.A., Lee S., 2019. Groundwater aquifer potential modeling using an ensemble multi-adoptive boosting logistic regression technique. Journal of Hydrology, 579, 124172.

Saraf A., et al., 2004. GIS based surface hydrological modelling in identification of groundwater recharge zones. International Journal of Remote Sensing, 25, 5759-5770.

Seenipandi D., Chandrasekar N., Magesh N.S., 2013. Identification of potential groundwater recharge zones in Vaigai upper basin, Tamil Nadu, using GIS-based analytical hierarchical process (AHP) technique. Arabian Journal of Geosciences, 7(4), 1385-1401.

Senanayake I., Dissanayake D., Mayadunna B., Weerasekera W., 2016. An approach to delineate groundwater recharge potential sites in Ambalantota, Sri Lanka using GIS techniques. Geoscience Frontiers, 7, 115-124.

Shahabi H., et al., 2019a. A Semi-Automated Object-Based Gully Networks Detection Using Different Machine Learning Models: A Case Study of Bowen Catchment, Queensland, Australia. Sensors, 19, 4893.

Shahabi H., et al., 2019b. A Semi-Automated Object-Based Gully Networks Detection Using Different Machine Learning Models: A Case Study of Bowen Catchment, Queensland, Australia. Sensors, 19, 4893.

Shirzadi A., et al., 2017. Shallow landslide susceptibility assessment using a novel hybrid intelligence approach. Environmental Earth Sciences, 76, 60.

Souissi D., et al., 2018. Mapping groundwater recharge potential zones in arid region using GIS and Landsat approaches, southeast Tunisia. Hydrological Sciences Journal, 63, 251-268.

Tien Bui D., et al., 2019. A hybrid computational intelligence approach to groundwater spring potential mapping. Water, 11, 2013.

Train K., 2009. Discrete Choice Methods With Simulation, Cambridge university press.

Tran A.-T., Le T.-H., Nguyen H.M., 2022. Forecast of surface chloride concentration of concrete utilizing ensemble decision tree boosted. Journal of Science and Transport Technology, 2, 44-56.

Vadiati M., Adamowski J., Beynaghi A., 2018. A brief overview of trends in groundwater research: Progress towards sustainability? Journal of Environmental Management, 223, 849-851.

Visa S., Ramsay B., Ralescu A., Knaap E., 2011. Confusion Matrix-based Feature Selection, 710(1), 120-127.

Zaidi F.K., Nazzal Y., Ahmed I., Naeem M., Jafri M.K., 2015. Identification of potential artificial groundwater recharge zones in Northwestern Saudi Arabia using GIS and Boolean logic. Journal of African Earth Sciences, 111, 156-169.

Zhang B., et al., 2019. Potential hazards to a tunnel caused by adjacent reservoir impoundment. Bulletin of Engineering Geology and the Environment, 78, 397-415.