Geoscience Frontiers 12 (2021) 101154

HOSTED BY

Contents lists available at ScienceDirect

Geoscience Frontiers

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

Research Paper

Locally weighted learning based hybrid intelligence models for groundwater potential mapping and modeling: A case study at Gia Lai province, Vietnam

Hoang Phan Hai Yen [a], Binh Thai Pham [b], Tran Van Phong [c], Duong Hai Ha [d,*], Romulus Costache [e,f], Hiep Van Le [b], Huu Duy Nguyen [g], Mahdis Amiri [h], Nguyen Van Tao [c], Indra Prakash [i]

[a] Department of Geography, School of Social Sciences Education, Vinh University, 182 Le Duan, Vinh, Nghe An, Vietnam
[b] University of Transport Technology, Hanoi 100000, Viet Nam
[c] Institute of Geological Sciences, Vietnam Academy of Science and Technology (VAST), 84 Chua Lang, Dong Da, Hanoi, Viet Nam
[d] Institute for Water and Environment, Hanoi 100000, Viet Nam
[e] Research Institute of the University of Bucharest, 90-92 Sos. Panduri, 5th District, Bucharest, Romania
[f] National Institute of Hydrology and Water Management, Bucureşti-Ploieşti Road, 97E, 1st District, 013686 Bucharest, Romania
[g] Faculty of Geography, VNU University of Science, Vietnam National University, 334 Nguyen Trai, Hanoi 100000, Viet Nam
[h] Department of Watershed & Arid Zone Management, Gorgan University of Agricultural Sciences & Natural Resources, Gorgan 4918943464, Iran
[i] DDG (R) Geological Survey of India, Gandhinagar 382010, India

ARTICLE INFO

Article history:
Received 28 August 2020
Received in revised form 22 January 2021
Accepted 28 January 2021
Available online 20 February 2021

Handling Editor: E. Shaji

Keywords:
Locally weighted learning
Hybrid models
Groundwater potential
GIS
Vietnam

ABSTRACT

The groundwater potential map is an important tool for a sustainable water management and land use planning, particularly for agricultural countries like Vietnam. In this article, we proposed new machine learning ensemble techniques namely AdaBoost ensemble (ABLWL), Bagging ensemble (BLWL), Multi Boost ensemble (MBLWL), Rotation Forest ensemble (RFLWL) with Locally Weighted Learning (LWL) algorithm as a base classifier to build the groundwater potential map of Gia Lai province in Vietnam. For this study, eleven conditioning factors (aspect, altitude, curvature, slope, Stream Transport Index (STI), Topographic Wetness Index (TWI), soil, geology, river density, rainfall, land-use) and 134 wells yield data was used to create training (70%) and testing (30%) datasets for the development and validation of the models. Several statistical indices were used namely Positive Predictive Value (PPV), Negative Predictive Value (NPV), Sensitivity (SST), Specificity (SPF), Accuracy (ACC), Kappa, and Receiver Operating Characteristics (ROC) curve to validate and compare performance of models. Results show that performance of all the models is good to very good (AUC: 0.75 to 0.829) but the ABLWL model with AUC = 0.89 is the best. All the models applied in this study can support decision-makers to streamline the management of the groundwater and to develop economy not only of specific territories but also in other regions across the world with minor changes of the input parameters.

1. Introduction

Historically, groundwater has been an inseparable part of human life. This is explained by the fact that a significant portion of water demand, especially in the drinking and agricultural sectors, is met all over the world by the groundwater resources (Bhuvaneswaran and Ganesh, 2019). Excessive usage of groundwater resources corroborated with some factors, such as insufficient rainfall, high population density and low surface water volume, which have led to an increase in the groundwater demand around the world (Sajedi-Hosseini et al., 2018). In general, the groundwater levels are declining in recent decades due to overuse and improper protection measures. Therefore, accurate spatial forecasting of groundwater resources is vital for the sustainable use, conservation and management strategies (Koh et al., 2020). Quality of groundwater is of great importance due to its various domains of usage. Changes in groundwater quality in an area are consequences of physical and chemical parameters that are highly influenced by geological formations and anthropogenic activities (Subramani et al., 2005; Šimanský et al., 2018; Sato et al., 2019). One of the most important human activities that lead to physical and chemical pollution of groundwater is the urban and industrial sewage, whose expansion is widely correlated with the population growth, urbanization and lifestyle changes (Rahman, 2008; Van Hoang et al., 2020). Moreover, frequent environmental changes, due to natural and human factors, are a significant threat to the groundwater quality. The human activities such as agriculture and industries lead to an excessive pollution with physical and chemical elements such as chloride, lead, nitrate, arsenic and

* Corresponding author.
E-mail address: hahaiduongcwe@yahoo.com (D.H. Ha).

https://doi.org/10.1016/j.gsf.2021.101154

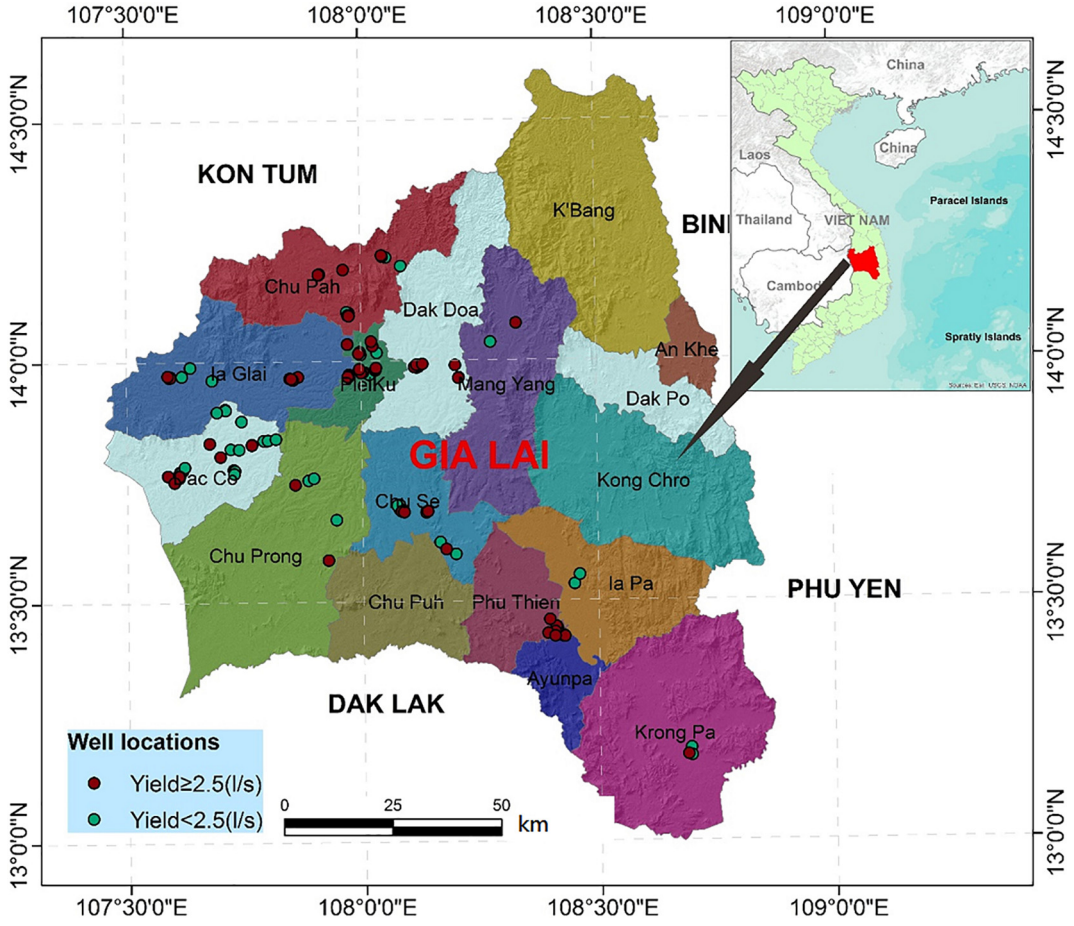

Fig. 1. Research area location in Vietnam.

ammonia (Khattak et al., 2020; Trung et al., 2020). The precision of the researches related to groundwater potential depends mainly on the selection of appropriate influencing factors and on the use of precise methods for spatial modeling (Arabameri et al., 2019; Sivasankar et al., 2019). Therefore, selection of the suitable influencing factors for the identification of areas with high groundwater storage potential is vital for an efficient water resource management.

Spatial techniques such as Remote Sensing (RS) and Geographic Information Systems (GIS) in conjunction with ground observation data are considered to be the most accurate and efficient methods for describing and understanding the spatial distribution of groundwater potential (Mosavi et al., 2020a). These RS and GIS techniques demonstrated their reliability as well as their ability to accurately predict the groundwater potential (Andualem and Demeke, 2019; Malakootian et al., 2020). A few researchers have combined statistical and probabilistic methods: weight of evidence, frequency ratio and evidential belief function with the GIS techniques to recognize the groundwater potential across specific regions (Mogaji and Lim, 2018; Arabameri et al., 2019). Other researchers have used Random Forest (RF), Support Vector Machines (SVM), Multivariate Adaptive Regression Spline (MARS), Locally Weighted Learning, Bagging, Dagging, AdaBoost, Multi Boost, Rotation Forest and Booting Regression Tree (BRT) (Bui et al., 2016; Naghibi et al., 2017a; Golkarian et al., 2018; Sameen et al., 2019) to assess the groundwater potential. In recent years, many researchers are applying hybrid algorithms in groundwater studies for better accurate prediction and mapping of groundwater potential. Pham et al. (2019a) used the Bagging-Decision Stump model to study the potential of groundwater in Vadodara district, Gujarat, India. Miraki et al. (2019) used a new machine learning hybrid model which is a combination of Random Forest and Random Subspace ensemble in order to estimate the groundwater potential in Kurdistan province from Iran. Arabameri et al. (2020) used a novel ensemble of Multi-criteria Decision Making and Artificial Intelligence techniques in groundwater potential mapping in Iran. Chen et al. (2019) demonstrated effectiveness of hybrid models in groundwater potential mapping within Shanxi Province of China using Novel hybrid integration approach of bagging-based Fisher's linear discriminate function. Some researchers have used logistic regression-based multi-adaptive boosting ensemble (Rizeei et al., 2019), and metaheuristic based neural fuzzy (Kordestani et al., 2019) in groundwater potential mapping.

In this paper, a single machine learning model namely Locally Weighted Learning (LWL) and four ensemble models: AdaBoost - Locally Weighted Learning (ABLWL), Bagging - Locally Weighted Learning (BLWL), Multi Boost - Locally Weighted Learning (MBLWL) and Rotation Forest - Locally Weighted Learning ensemble (RFLWL) with LWL as a base classifier were used for ground water potential mapping and to model the spatial distribution of the groundwater potential in the Gai Lai province, Vietnam. It is worth to mention that all these five proposed algorithms have been used for the first time to map, evaluate and model the groundwater potential of an area. Performance and predictive ability of models were evaluated using several statistical indices namely Area Under Receiver Operating Characteristic (ROC) curve (AUC), Kappa, Accuracy (ACC), Specificity (SPF), Sensitivity (SST), Negative predictive value (NPV), and Positive predictive value (PPV). Weka packages and GIS software were used for data processing and modeling.

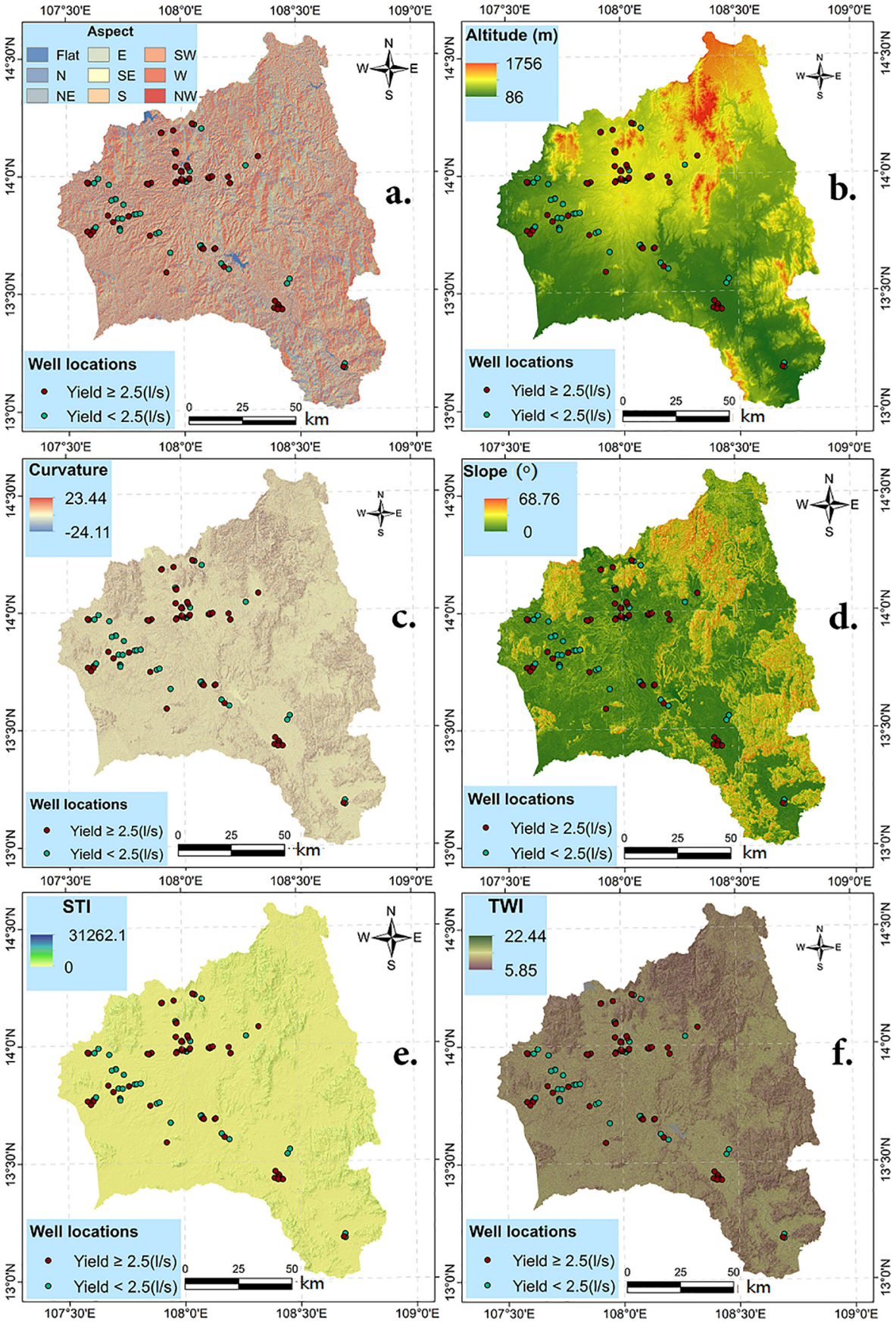

Fig. 2. Spatial extent and values of groundwater potential predictors. (a) Aspect; (b) altitude; (c) curvature; (d) slope; (e) STI; (f)TWI.

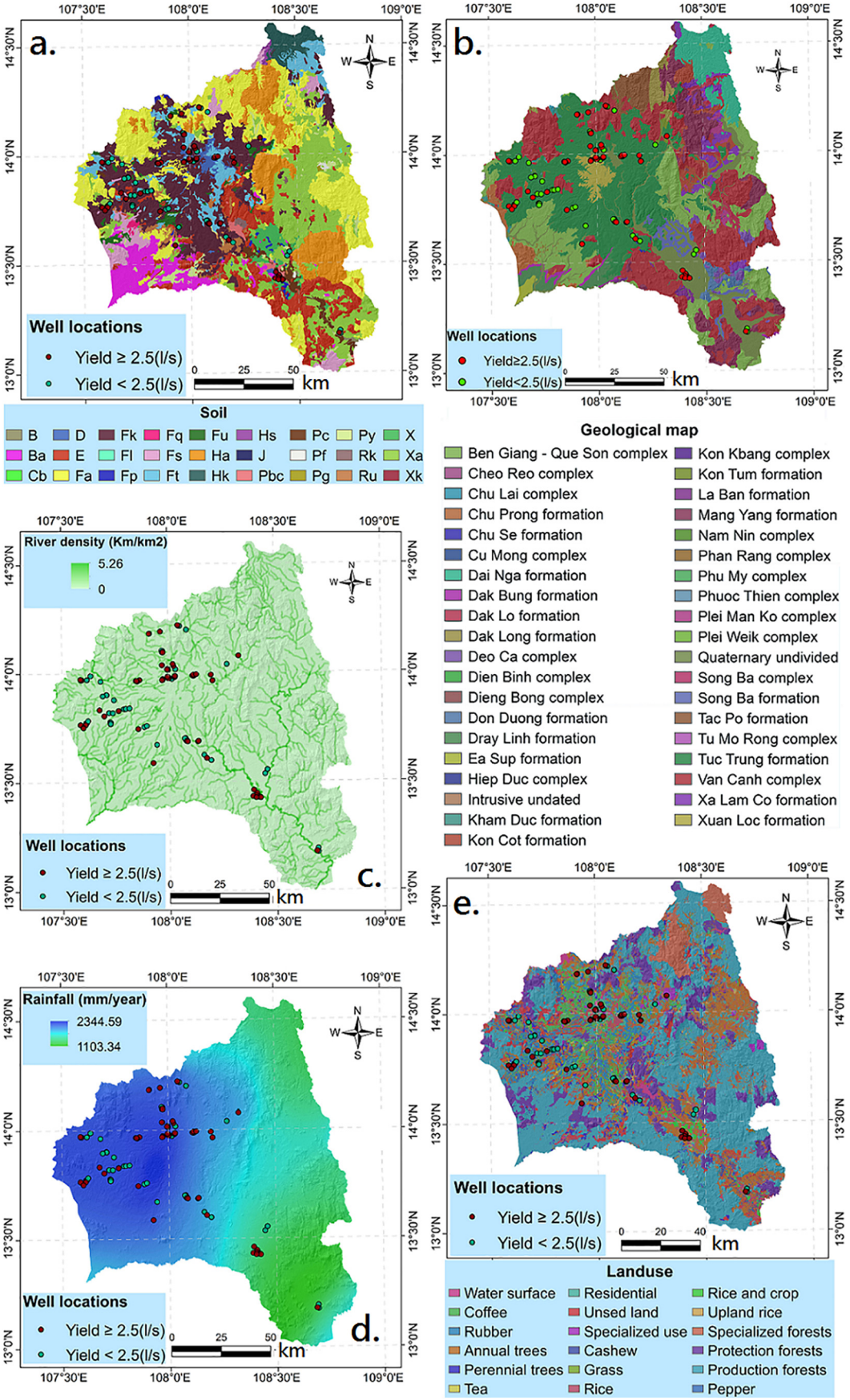

Fig. 3. Spatial extent and values of groundwater potential predictors. (a) Soil; (b) geological map; (c) river density; (d) land use; (e) rainfall.

Table 1
Information of soil types used in this study.

No.	Code	Description	No.	Code	Description
1	B	Loamy sand or gravelly loamy sand	14	Ha	Red yellow humus on acid magma rock
2	Ba	Faded soil on acid magma and sand	15	J	Grab soil
3	Cb	Coastal Beach	16	Pbc	Sour alluvial soil
4	D	Land sloping valley by the convergence	17	Pc	Alluvial soil
5	E	Erosion's soil inert	18	Pf	The alluvial soil has red and yellow sloping layers
6	Fa	Red yellow soil on acid magma	19	Pg	Alluvium glay soil
7	Fk	Red-brown soil on basalt	20	Py	Stream alluvial soil
8	Fl	Red-yellow soil changes due to wet rice cultivation	21	Rk	Black soil on basalt accretion products
9	Fp	Brown-yellow soil on ancient alluvial gold	22	Ru	Permeable brown soil on foam basalt products
10	Fq	Pale yellow soil on sand stone	23	X	Gray soil on ancient alluvium
11	Fs	Yellow-red soil on clay and metamorphic rocks	24	Xa	Gray soil on acid magma and sand stone
12	Ft	Purple-brown soil on basalt	25	Xk	Gray glay soil
13	Fu	Brown-yellow soil on basalt			

2. Materials and methods

2.1. Study area and data

2.1.1. Study area

Gia Lai (12°58'20" to 14°36'30" north latitude; 107°27'23" to 108°54'40" east longitude) is a mountainous province located in the northern part of the Central Highlands (Fig. 1) at an average elevation of 700–800 m above sea level. Topography of the area is rugged which includes mountains, highlands and valleys. Elevation of the area reduces from north to south. Geologically the area is occupied by basalt rocks of the three main eruptive episodes: Miocene, Pliocene and Pleistocene (Phuc et al., 2018). Cenozoic basalts have provided natural resources such as groundwater, forestry and bauxite (Tuan et al., 2019). Soil types in the area belongs to seven main groups: alluvial soil, gray soil, black soil, red soil, red yellow loam soil, and inert soil. In this region, main land use includes crops of rubber, coffee and pepper plants. In addition, land use pattern also includes forest land in the study area (Quyen et al., 2014).

This province belongs to the tropical highland monsoon climate region with heavy rainfall (Minh et al., 2018). Climate in this region is divided into two distinct seasons: rainy and dry seasons. In particular, the rainy season usually starts from May and ends in October. The dry season is from November to April next year (Ly and Thuy, 2019). The annual average temperature ranges between 22° and 25 °C. The average rainfall 1200–1750 mm is in the East Truong Son Region, and 2200–2500 mm in the West Truong Son Region. This area has rich surface water resources of about 23 billion m^3, which are distributed in the large river basins such as Se San River, Ba River, Srep Pook River.

The province of Gialai is often affected by the drought problem. In 2019, drought caused thousands of hectares of agricultural losses. The total economical damages are estimated to about 2 millions US dollars. Surface water is unevenly distributed in the province which is not sufficient for the sustainability and development of the area. Therefore, construction of the groundwater potential map for the development of groundwater resources is considered a mandatory task for the proper water resource management in this province.

2.1.2. Data used

2.1.2.1. Well yields. Well yields data is important for assessing groundwater potential of the area. Proper well inventory data is required to be used in the ground water potential mapping based on the Machine Learning models (Chen et al., 2020; Nguyen et al., 2020a). Quality of the results of the groundwater potential modeling is directly dependent on the accuracy of the wells locations (Rahmati et al., 2019). In this study, we have used the data of 134 georeferenced wells obtained from Vietnam Academy for Water Resources and the field mission carried out in the years 2018 and 2019.

2.1.2.2. Groundwater influencing factors. The selection of groundwater influencing factors is a crucial step in the workflow developed to model the groundwater potential. In this study 11 influencing factors were selected and divided into different groups: Topography (aspect, altitude, slope), geology, soil, hydrology (river density, curvature, STI, TWI), climatic conditions (rainfall) and anthropogenic activities (landuse). The topography and hydrology factors were extracted from Aster Digital Elevation Model (DEM) of 30 m resolution. The climatic factors were obtained from WorldClim v2 database. The land use map, soil map and geological map were collected at a scale of 1:50000 from the Resource and Environment department of Gia Lai province.

Aspect is considered to be an important factor in constructing the groundwater potential model because it shows the relationship with the direction of water flow (Solomon and Quiel, 2006; Costache, 2019a; Nguyen et al., 2020b). In this study, the aspect map has been divided into nine classes: Flat, N, NE, E, SE, S, SW, W, NW (Fig. 2a). Altitude plays an important role in the prediction model of groundwater potential. Usually the groundwater level follows topography altitude line, and has also tendency to accumulate under the low altitude surfaces (Ozdemir, 2011; Chen et al., 2018). In this study area, altitude ranges from 86 m to 1756 m (Fig. 2b). The curvature shows the ability of the water to accumulate at the ground surface and thus in the infiltration (Pham et al., 2019b). In the study area, the curvature ranges from −24.11 to 23.44 (Fig. 2c). Slope factor was selected due to its relationship with the hydrology processes. The slope reflects the altitudinal changes over the distance unit and is considered a significant element in determining runoff direction and seepage (infiltration) capacity (Ercanoglu and Gokceoglu, 2002). In the study area slope varies from 0 to 68° (Fig. 2d). STI shows the sediment transfer capacity and characterizes the process of erosion and depositions (Conforti et al., 2011). In this study, the STI value ranges from 0 to 31,262 (Fig. 2e). TWI is an important factor in quantifying control of topography on hydrologic process thus in infiltration and run off (Rahmati et al., 2016; Naghibi et al., 2017b). In this study, the TWI value ranges from 5.85 to 2.43 (Fig. 2f).

Soil exhibits the ability of the water from the ground surface to infiltrate and to recharge the groundwater reservoirs (Oanh and Van Lap, 2016; Naghibi et al., 2017a). The soil map in this study has been divided into 27 types (Fig. 3a and Table 1). The geology factor plays an important role in the surface permeability, and therefore, in the groundwater recharge capacity (Ayazi et al., 2010; Al-Abadi, 2015). The geology map has been classified into 39 types (Fig. 3b).

River density influences the recharge capacity of groundwater (Ayazi et al., 2010). River density in the study area varies from 0 to 5.2 km/km^2.

Rainfall is considered to be a very important factor for the modeling of the groundwater potential due to its direct impact on the recharge capacity (Fig. 3c). Precipitation directly helps in the recharge of the groundwater potential zones (Naghibi and Pourghasemi, 2015; Oikonomidis et al., 2015). Rainfall in the study area varies from 1103 mm to 2344 mm (Fig. 3e).

Land use is considered as an important anthropogenic factor in groundwater study. The land use map has been classified into 18 different categories (Fig. 3d). Constructed (concrete/ bitumen etc.) surfaces unless properly design lead to an increase of surface runoff and thus decrease capacity of ground surface in recharging groundwater reservoirs by infiltration (Hasegawa et al., 2017; Hoa et al., 2019).

2.2. Methods

2.2.1. Locally weighted learning (LWL)

Locally Weighted Learning methods are a class of function approximation techniques, where a local model is created for each point of interest based on neighbouring data, instead of building a global model for the whole function space. LWL methods are non-parametric.

In the LWL, local regression can be used for multi-objective regression subject as follows: i) implementing a locally weighted regression procedure for each aim agent, or ii) sketching a weightlifting method that directly controls multiple goal data. The major advantage of LWL is the placement of the data in a small neighborhood (Adeli et al., 2017). The standard form of LWL algorithm is described by the following mathematical equation (Schaal et al., 2000):

$$y = f(x) + \varepsilon \tag{1}$$

where $x \in \Re_n$ is a n-dimensional input vector, the noise term has mean zero, $E\{\varepsilon\} = 0$, and the output is one-dimensional.

2.2.2. AdaBoost

Boosting algorithms belong to the family of group learning methods, which can improve the grouping accuracy by synthesizing many "weak/feeble" basic learners to generate a "strong/powerful" committee (Jiang et al., 2019). In AdaBoost the decision tree is often used as a weak learner, while its predictions are merged in order to produce the final outcomes (Jiang et al., 2019). The major advantage of AdaBoost model is the dissolving two-category problems, multi-category single-tag problems, multi- category multi-tag problems, groups of single-tag problems, and regression problems (Hong et al., 2018), on the other hand, the main purpose of group classification is to decrease the category (error rate) of a weak class by presenting and collecting multiple classifications (Rajesh and Dhuli, 2018). The output of the AdaBoost final classifier is defined by the following equation (Zheng and Peng, 2019):

$$H(x) = \sum_{t=1}^{J} \alpha_t h_t(x) \tag{2}$$

where the AdaBoost creates the hypothesis $h_t = \{-1, +1\}$ and generates the weight of the hypothesis α_t correctly. H signifies the combined hypothesis while $t = 1, 2, \ldots, J$ represents the cycle of training.

iteration of base classifier.

2.2.3. Bagging ensemble

Bagging (known as the Bootstrap Collection) represents a simple group of learner which collects basic models that are created using bootstrap samples (Subasi et al., 2020). According to Breiman (1996), the bootstrap samples are used to drive the single classifiers. If we consider $R = (a, b)$ a vector included in the initial training sample, where $a = a_i$, $i = 1, 2, \ldots, n$ represent the groundwater potential predictors (n is the sum of predictors) and $b = b_i \in (1$ – wells locations, 0 – non-wells locations). According to Bui et al. (2016), the predictor of *Bagging ensemble* model $\phi(a, T)$ will predict a y class following the relation below:

$$Q(b_i|a) = P[\phi(a, T) = b_i] \tag{3}$$

Further will be defined the probability $P(b_i, a)$ where a generates the values of b_i, while the form of the probability (P_{cor}) in which the predictor generate a correct classification for x is the following:

$$P_{cor} = \sum_{b_i} Q(b_i|a)\, P(b_i|a) \tag{4}$$

Finally, the probability associated to the final correct classification can be written as:

$$p = \int \left[\sum_{b_i} Q(b_i|a) P(b_i|a) \right] P_a d(a) \tag{5}$$

where $P_a d(a)$ represents the distribution probability of a.

2.2.4. Multi Boost ensemble

The Multi Boost method represents a combination of Ada Boost (boosting) and Wagging (a variant of bagging) models, with the ability to decrease the variance and bias, and also to avoid over-fitting subject in the spatial modeling (Bui et al., 2016). This algorithm was first introduced by Webb (2000). Wagging is a component of Bagging that uses a variety of training instances which creates several weights intended to decrease the bias of the AdaBoost method.

The application of Multi Boost ensemble requires three stages (Wang et al., 2020): (i) in a first stage, from the training sample a subset is selected in a random way in order to create the initial base classifier; (ii) in the second stage, the precision performance is involved in the adjustment of the instance weight; (iii) in the final stage, in order to train a new classifier, from the instance weighted a new subset is selected. Let consider $S(a_i, b_i)$ a training dataset, where $a_i \in R$ and $b_i \in$ (wells locations and non-wells locations) (Bui et al., 2016), the next equation will provide the value of the final classifier (Webb, 2000):

$$C' = \underset{b \in B}{\operatorname{argmax}} \sum_{t:C_t(a)=b} log \frac{1}{D_t} \tag{6}$$

$$\varepsilon_t = \frac{\sum a_j \in s : C_{j(a_j) \neq b_i} \text{weight}\,(a_j)}{m} \tag{7}$$

where C' is equal to C having associated the weights assigned to be 1, C_t is a base learner (C'), and ε_t is the weighted error which characterizes the training dataset.

2.2.5. Rotation Forest ensemble

Rotation Forest is an ensemble machine learning algorithm which generates classifier based on their feature extraction procedure (Ozcift and Gulten, 2011). Within the Rotation Forest ensemble, the split of the feature set in many subsets is followed by a Principal Component Analysis (PCA) which is applied to each new subset (Hong et al., 2018). A classification is constructed on features that are frequently generated by the predicted matrix, while the final result will be achieved by combining the output of multiple classifications. Detailed description of mathematical background of Rotation Forest is as below.

Let X = W be a set of training examples, Y = P be the relevant class tags, and F be its attributes. If we suppose that there are N = M training instances and n attributes in a microarray data set, X = W is an N × N matrix. Let Y = P be a set of category tags (ω1…, ωc) where Y = P takes values. Further, if the attribute set is randomly divided into K sub-sets of proximate size, then there are L decision trees in a Rotation Forest defined by D1…, DL, where, L and K are two parameters that

must be predetermined. The construction of the training set for the classifier is carried out according to the following steps (Hong et al., 2018):

(1) Divide F randomly into subdivided K sets. Separated subcategories are selected to maximize the chance of high diversity. For easiness, assume that K is an n parameter, so that each subset of the attribute contains the following property A = M = n / K.
(2) Let F_{ij} be a subset of the attributes for training classes Di, and let X_{ij} be the X data set for the F_{ij} attributes. For each proper subset, a subset other than the class is randomly chosen from X_{ij}. In the next step, a subset of bootstrap of objects is drawn with 75% of the data set to form a new training set called X_{ij}. Further, a linear transformation is performed on X_{ij} to be able to produce the constituent components in a C_{ij} matrix, which is determined using particles: the specific values are zero, therefore, it is possible that all M vectors are possible.
(3) Create a sparse rotation matrix Ri with the gain coefficients in matrix Cij, as following:

$$R_i = \begin{bmatrix} a_{i,1}^{(1)}, a_{i,1}^{(2)}, \ldots, a_{i,1}^{(M_1)} & 0 & \cdots & 0 \\ 0 & a_{i,1}^{(1)}, a_{i,1}^{(2)}, \ldots, a_{i,1}^{(M_2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{i,K}^{(1)}, a_{i,K}^{(2)}, \ldots, a_{i,K}^{(M_K)} \end{bmatrix} \quad (8)$$

The matrix is rearranged with the help of the values of coefficients which are associated to each class, their values being calculated through the average combination in the given test sample which is mathematically described by the following equation Van Hoang:

$$m_j(g) = \frac{1}{L} + \sum_{i=1}^{L} d_{ij}(gR_i^a), j = 1, \ldots, c \quad (9)$$

where $d_{ij}(xR_i^a)$ is the probability given by the classifier D_i for the hypothesis that x belongs to the class j.

2.2.6. Validation indices

Models performance evaluation, through statistical metrics, is an important step in the modeling process (Wu et al., 2004; Costache, 2019b; Miraki et al., 2019). This step is essential and critical for distinguishing the precision of potential groundwater models. To validate the predictability of models, it is desirable to evaluate them using both training and testing datasets (Khosravi et al., 2018; Costache et al., 2020a). In the current research, statistical metrics used are: Area Under Receiver Operating Characteristic (ROC) curve (AUC), Kappa, Accuracy (ACC), Specificity (SPF), Sensitivity (SST), Negative predictive value (NPV), and Positive predictive value (PPV) (Golkarian et al., 2018; Avand et al., 2020; Costache et al., 2020b; Qi et al., 2020). In groundwater potential modeling, PPV refers to the ratio of properly categorized high potential groundwater pixels to all categorized pixels as high potential groundwater pixels, while NPV mention to the ratio of properly grouped pixels (Nguyen et al., 2020b). They are classified as low potential groundwater pixels. The SST is the ratio of high potential groundwater pixels correctly classified from all properly classified pixels as high potential groundwater pixels plus those that are erroneously classified as low potential groundwater pixels. SPF represents the ratio of properly classified low potential groundwater pixels among all properly classified pixels as low potential groundwater pixels plus those that are incorrectly classified as high potential groundwater pixels. The ACC shows the overall performance of a forecast model and is computed as the ratio of high potential groundwater and low potential groundwater pixels that are properly classified (Pham et al., 2019c; Mosavi et al., 2020b; Nguyen et al., 2020c). The evaluation index formulas are given below (Jaafari et al., 2018; Costache and Bui, 2020; Qi et al., 2020):

$$PPV = \frac{TP}{TP + FP} \quad (10)$$

$$NPV = \frac{TN}{TN + FN} \quad (11)$$

$$SST = \frac{TP}{TP + FN} \quad (12)$$

$$SPF = \frac{TN}{TN + FP} \quad (13)$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (14)$$

$$Kappa = \frac{P_{obs} - P_{exp}}{1 - P_{exp}} \quad (15)$$

$$AUC = (\sum TP + \sum TN)/P + N \quad (16)$$

where TP is True Positive (+), TN is True Megative (−), FP is False Positive (+), FN is False Negative (−), P is the all number of groundwater pixels, N is the total value of non-groundwater pixels, N is the value of instances in the dataset, P_{obs} and P_{exp} are the measured and envisaged constants, respectively.

The Receiver Operating Characteristic (ROC) curve, is the most common statistical method that is used to estimate proficiency and reliability of a binary prediction model (Bui et al., 2016; Jaafari et al., 2018). The Area Under the ROC curve (AUC) is used to assess the predictive ability of models (Phong et al., 2019; Yariyan et al., 2020).

2.2.7. Correlation based feature selection (CFS)

Disjointed and unneeded/superfluous variables should be eliminated from the modeling process to reduce the redundant information and to avoid the model bias (Dempster, 2008). This process helps to increase the accuracy of the applied models. In terms of groundwater potential modeling, a predictor is of a major importance if it is spatially correlated with the groundwater presence and uncorrelated with the other predictors (Hall, 2000). According to Costache (2019c), CFS model has the ability to eliminate the information that is noisy, irrelevant and redundant. In the case of CFS algorithm, the higher value will be assigned to the predictors that are highly correlated with the groundwater potential. These values will be calculated with the following relation (Costache, 2019c):

$$CFS = \frac{k \times r_{cf}}{\sqrt{k + k(k-1)r_{ff}}} \quad (17)$$

where *CFS* is the correlation between each groundwater predictor with the presence of wells points, k is the number of groundwater predictors, r_{cf} the average correlation between groundwater predictors and wells points, and r_{ff} the average inter-correlation between groundwater predictors.

3. Main steps of the methodological workflow

Methodology of this study is presented in Fig. 4 with following steps:

(i) Preparation of inventory of high potential groundwater locations and low potential groundwater locations: In this step, the well data was divided into two parts (50:50), including high potential groundwater locations (well yield ≥ 2.5 l/s) defined as "1" class in the dataset and low potential groundwater locations (well yield < 2.5 l/s) defined as "0" class in the dataset. Out of this, 70% of high and low potential groundwater locations were used to generate the training dataset used to construct the models and maps while 30% remaining high and low potential

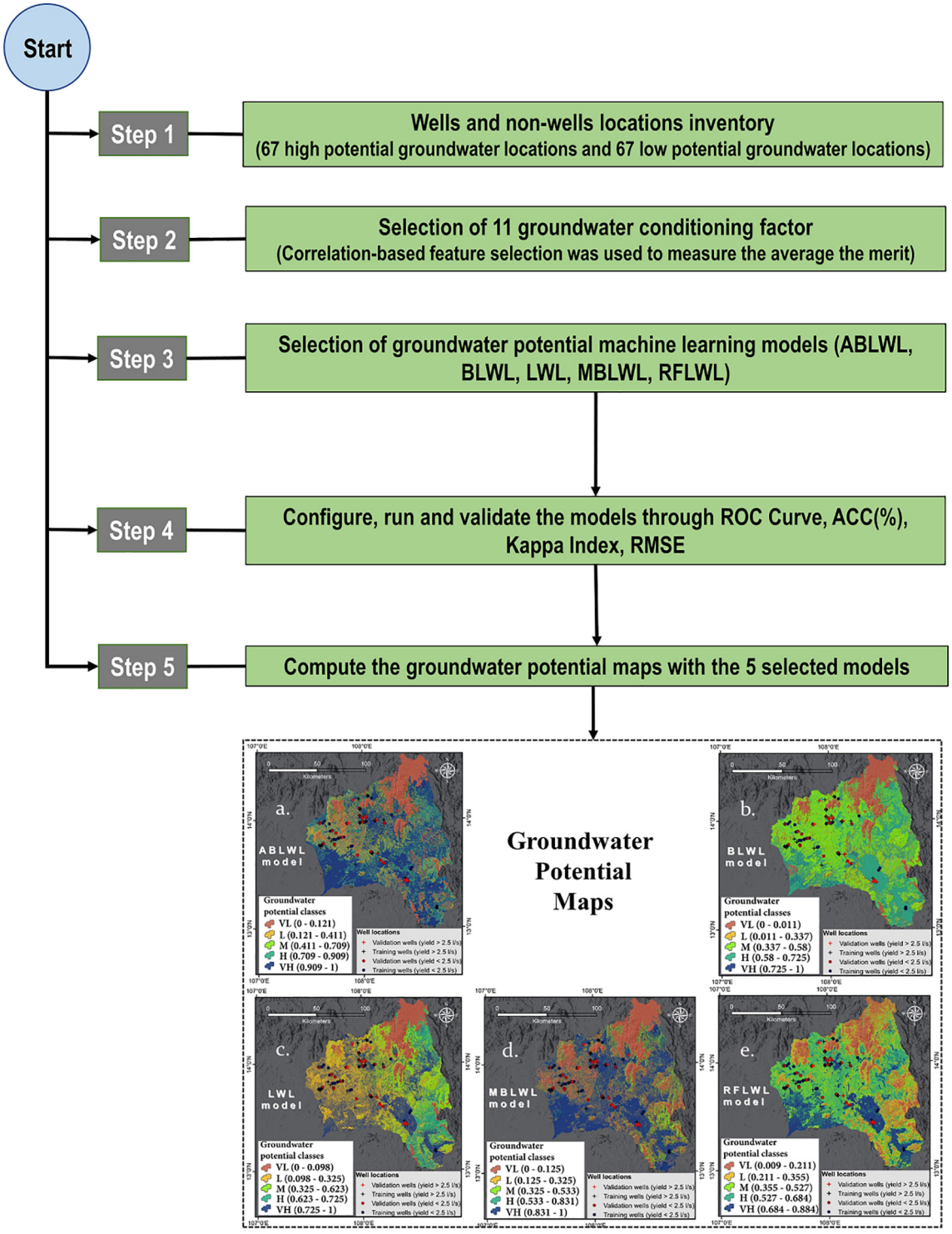

Fig. 4. Flowchart of the methodological steps applied in the present research.

groundwater locations were used to generate testing dataset used to validate the models and maps. This process was carried out by using the tools and functions available in GIS applications.

(ii) Validation and selection of the groundwater predictors using CFS model (their values were normalized between 0 and 1): in this step, a total of 11 groundwater potential factors was initially selected to sample with well data for generations of the initial training and testing datasets. Thereafter, the initial training datasets was used to validate the importance of the groundwater potential factors using CFS feature selection method on which the important factors were selected to generate the final datasets for the groundwater potential modeling.

(iii) Development of groundwater potential models: Four ensemble models namely ABLWL, BLWL, MBLWL, RFLWL were developed and one single model LWL was used for groundwater

Table 2
Hyper- parameters used to construct the models in this study.

No.	Hyper-parameter	Models				
		LWL	ABLWL	BLWL	MBLWL	RFLWL
1	KNN	–1	–	–	–	–
2	Batch size	100	100	100	100	100
3	Nearest neighbour search algorithm	Linear NN search	–	–	–	–
4	Number of decimal places	2	2	2	2	2
5	Weighting kernel	0	–	–	–	–
6	Number of interations	–	9	1	10	4
7	Seed	–	1	1	1	1
8	Weight of threshold	–	100	–	100	
9	Number of execution slots	–	–	1	–	1
10	Number of subcommittees	–	–	–	3	–
11	Maximum size of a group	–	–	–	–	3
12	Minimum size of a group	–	–	–	–	3
13	Percentage of instances to be removed	–	–	–	–	50

Table 3
Select attribute feature using correlation-based feature selection method.

Rank	Average merit	Error (AM)	Average rank	Error (AR)	Factor
1	0.496	0.02	1	0	Slope
2	0.444	0.021	2	0	Elevation
3	0.352	0.013	3.1	0.3	Geology
4	0.286	0.026	4.4	0.66	STI
5	0.287	0.018	4.5	0.5	River density
6	0.222	0.021	6	0	Soil
7	0.103	0.021	7.4	0.66	Rainfall
8	0.075	0.027	8.3	1.19	TWI
9	0.05	0.022	9.5	1.12	Landuse
10	0.05	0.022	9.5	1.02	Curvature
11	0.024	0.02	10.3	0.78	Aspect

potential assessment and mapping. In the hybrid/ensemble models, AdaBoost, Bagging, MultiBosst, and Rotation Forest ensembles were used to optimize the training dataset in the ABLWL, BLWL, MBLWL, RFLWL models, respectively while LWL was used as a base classifier. In this step, training dataset was used to build the models. Hyper-parameters used to construct the models are presented in Table 2.

(iv) Validation of groundwater potential models: In this step, various quantitative validation indices such as AUC, Kappa, ACC, SPF, SST, NPV, and PPV have been applied on both training and testing datasets for validating and comparing the goodness of fit and the predictive capability of the models, respectively.

(v) Construction of groundwater potential maps: In this step, the results of training the models were used to construct the maps of which the Natural Break classification method was used to classify the potential classes of the groundwater on the maps. Validation of the maps was also done by using statistical frequency ratio analysis.

4. Results

4.1. Factor importance

One of the important steps in the groundwater potential model is the selection of the relevant groundwater predictors. The goal of this step is to remove unnecessary factors and focus on the most important factors which have ability to reduce noise and to increase the models accuracy (Hoa et al., 2019; Nguyen et al., 2019; Bui et al., 2020). The Table 3 presents Average Merit (AM) value of 11 influence factor based on Correlation Attribute Evaluation method. The results show that all the factors have an Average Merit score which allows them to significantly contribute modeling process of the groundwater potential in Gia Lai province. However, the slope (0.496), elevation (0.444), Geology (0.352) and STI (0.286) factors are the most important predictors. This finding is in agreement with previous studies (Chen et al., 2020; Nguyen et al., 2020c).

4.2. Model validation and comparison

In this study, various statistical indices were used to validate the model's performance (Table 4 and Fig. 5). In terms of training data, the highest accuracy was achieved by MBLWL model (88.71%), followed by ABLWL (84.68%), RFLWL (82.26%), LWL (80.65%) and BLWL (73.39%). In terms of Kappa index, performance of the MBLWL model is the best (0.773), followed by ABLWL (0.691), RFLWL (0.641), LWL (0.614), and BLWL (0.458). For validation data, the ABLWL model achieved the highest accuracy (77.36%), followed by BLWL (69.81%), MBLWL (69.81%), RFLWL (69.81%) and LWL (67.92%). The scores of Kappa Index revealed that ABLWL model is the best (K = 0.544), being followed by MBLWL (0.392), BLWL (0.389), FRLWL (0.389) and the LWL (0.358). Additionally, the ROC Curve (Fig. 5) shows that in terms of training sample, the ABLWL model has the highest performance among the five applied models with an AUC-ROC of 0.968. The second in the hierarchy of performances was the model MBLWL (0.967), followed by RFLWL (0.919), LWL (0.917) and BLWL (0.87). For the validation data, the ABLWL model is the most efficient, with an AUC-ROC of 0.829, followed by the MBLWL (0.786), BLWL (0.776), RFLWL (0.759) and LWL (0.746) (Fig. 5).

4.3. Groundwater potential maps

The maps of the groundwater potential were constructed after the training procedure and validation of the five models. These processes were carried out in two main steps: i) all the pixels of the entire study area were included in the models to generate the indices of the groundwater potential; ii) reclassification of these indices using the natural break method in ArcGIS software (Bui et al., 2020; Nguyen et al., 2020a). The groundwater potential maps in this study were divided into five categories: very low, low, moderate, high, and very high.

By analyzing the groundwater potential resulted from ABLWL model (Fig. 6a), it can be seen that 21.2% of the study area is in the very low potential area, 9.882% in the low area, 3.87 in the moderate area, 1 7.83% in the high zone, and 47.22% in the very high zone. It can be observed that

Table 4
Statistical metrics used for the evaluation of model's performance.

	Training					Validating				
	ABLWL	BLWL	MBLWL	RFLWL	LWL	ABLWL	BLWL	MBLWL	RFLWL	LWL
PPV (%)	70	45	83.33	63.33	85	61.54	38.46	50	38.46	65.38
NPV (%)	98.44	100	93.75	100	76.56	92.59	100	88.89	100	70.37
SST (%)	97.67	100	92.59	100	77.27	88.89	100	81.25	100	68
SPF (%)	77.78	65.98	85.71	74.42	84.48	71.43	62.79	64.86	62.79	67.86
ACC (%)	84.68	73.39	88.71	82.26	80.65	77.36	69.81	69.81	69.81	67.92
K index	0.691	0.458	0.773	0.641	0.614	0.691	0.389	0.392	0.389	0.358

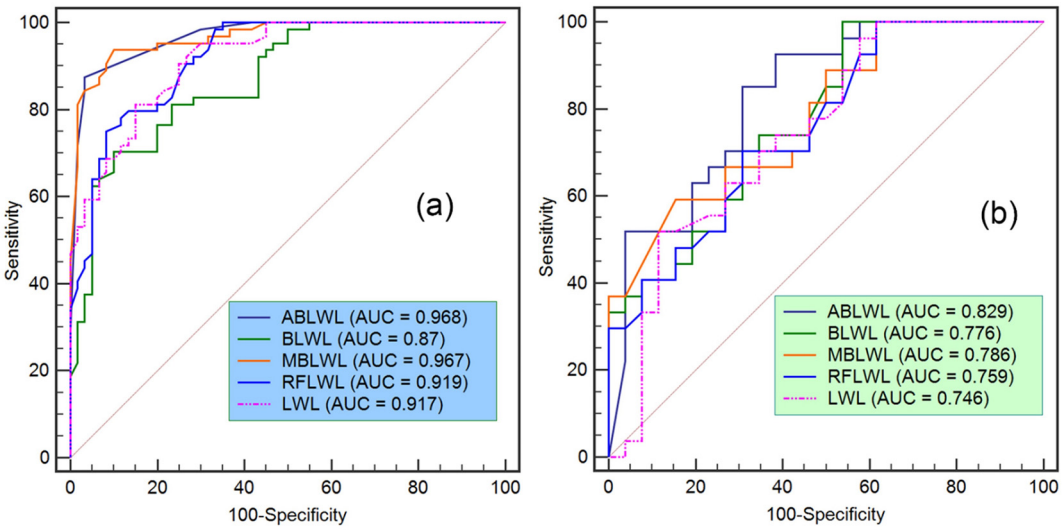

Fig. 5. ROC curve of models. (a) Training dataset; (b) validation dataset.

the high and very high groundwater potential is located especially in the southern-western and southern part of the study area, while the low and very low groundwater potential can be found in the northern side of research zone. In the case of the BLWL model (Fig. 6b), 18.02% of Gia Lai province is located in very low, 2.198% in low, 13.64 in high, 2.4 in very high groundwater potential. It is worth to note that the central part of the study region is characterized by a medium groundwater potential, while the very low potential is spread especially in the extreme northern region. In terms of LWL model (Fig. 6c), 12.08% of the study area is characterized by a very low groundwater potential, 33.6% in low groundwater potential, 13.64% in moderate groundwater potential, 12.54% in high groundwater potential, while 28.15% of the study area has a very high potential for groundwater presence. The application of this model revealed a high concentration of low groundwater potential in the western half of the study area, while the eastern half is mainly characterized by the presence of medium, high and very high potential. If we analyze the results of MBLWL model (Fig. 6d), it can be observed that 35.05% of the study area is in the very low potential area, 2.255% has a low groundwater potential, 12.24% is characterized by a moderate potential, 2.754% has high potential for groundwater presence, while around 47.7% area is included in the very high potential areas. The very high potential zone covers very large areas in the southern and central parts of the study area. It should be noted that the transition between the areas with very high potential to the areas with very low potential is within very short distances. In terms of the RFLWL model (Fig. 6e), 16.51% of the study area is located in the area with a very low groundwater potential, 14.92% in low potential area, 24.18% in the high potential area, while around of 16.6% of the study area has a very high groundwater potential (Fig. 7). All maps show that the groundwater potential is concentrated in the southern and southwestern part of the study area. The large spread of the medium groundwater potential is across the entire study area.

5. Discussion

The groundwater is one of the most important resources for the economic development of a country (Khosravi et al., 2019; Tien Bui et al., 2019). Therefore, precise identification of high groundwater potential area/zone is very important for the best management of this valuable resource. It should be mentioned here that the majority of the previous studies focused on groundwater potential mapping using different approaches (Shahid et al., 2000; Oh et al., 2011; Arabameri et al., 2020). However, more effort is needed to improve the predictive ability of the models to generate better groundwater potential maps. Currently, with the development of Machine Learning (ML) approaches several models have been developed and applied in the groundwater studies. These ML methods have been able to explain the complex relationships between the groundwater predictors and location of the wells and or groundwater springs (Chen et al., 2019; Jaafari et al., 2019; Nguyen et al., 2020c). It is worth to mention that these techniques can also be adapted to be applied even in large areas, with the availability of limited data (Rahmati et al., 2018).

On the analysis of the previous studies related to the estimation of different natural phenomena using ML models, we can see that these techniques provide different results if they are applied in different regions. For example, in terms of landslide susceptibility, Adaboost is more efficient in combination with ADtree than the Bagging technique (Wu et al., 2020). Similarly, Adaboost model is better in combination with HyperPipes than Bagging for the construction of the landslide potential map (Tran et al., 2020). From these studies, we can conclude that the predictive ability of machine learning depends on the local geo-environment conditions and the input variables. According to Bui et al. (2020) the higher amount of the input data could lead to more accurate results.

In this study, the different ML techniques including single LWL and four novel hybrid models namely ABLWL, BLWL, MBLWL and RFLWL were used to determine the regional distribution of groundwater potential. Statistical indices like PPV, NPV, SST, SPF, ACC, Kappa, and ROC were used to validate and compare performance of the models. All models achieved good (AUC: 0.7–0.8) and very good (AUC: 0.8–0.9) results (Choubin et al., 2019) and were successfully used to construct the groundwater potential map of the study area., In terms of AUC value, performance of the hybrid model ABLWL (0.829) is the best followed by MBLWL (0.786), BLWL (0.776), RFLWL (0.759) and single LWL (0.746). The results highlight that hybrid ML models performed better than the stand-alone one, as hybrid models produces low error values and also reduces bias and overfitting issues (Abdollahi and Ebrahimi, 2020).

Locally Weighted Learning is built by combining several local models which are also characterized by the advantage of the improvements of different individual algorithms (Flentge, 2006). Therefore, LWL as a base classifier in case of the hybrid/ ensemble models has a high contribution in increasing prediction capability of individual models like Adaboost, Bagging, Multiboost, Rotation forest. The AdaBoost model is

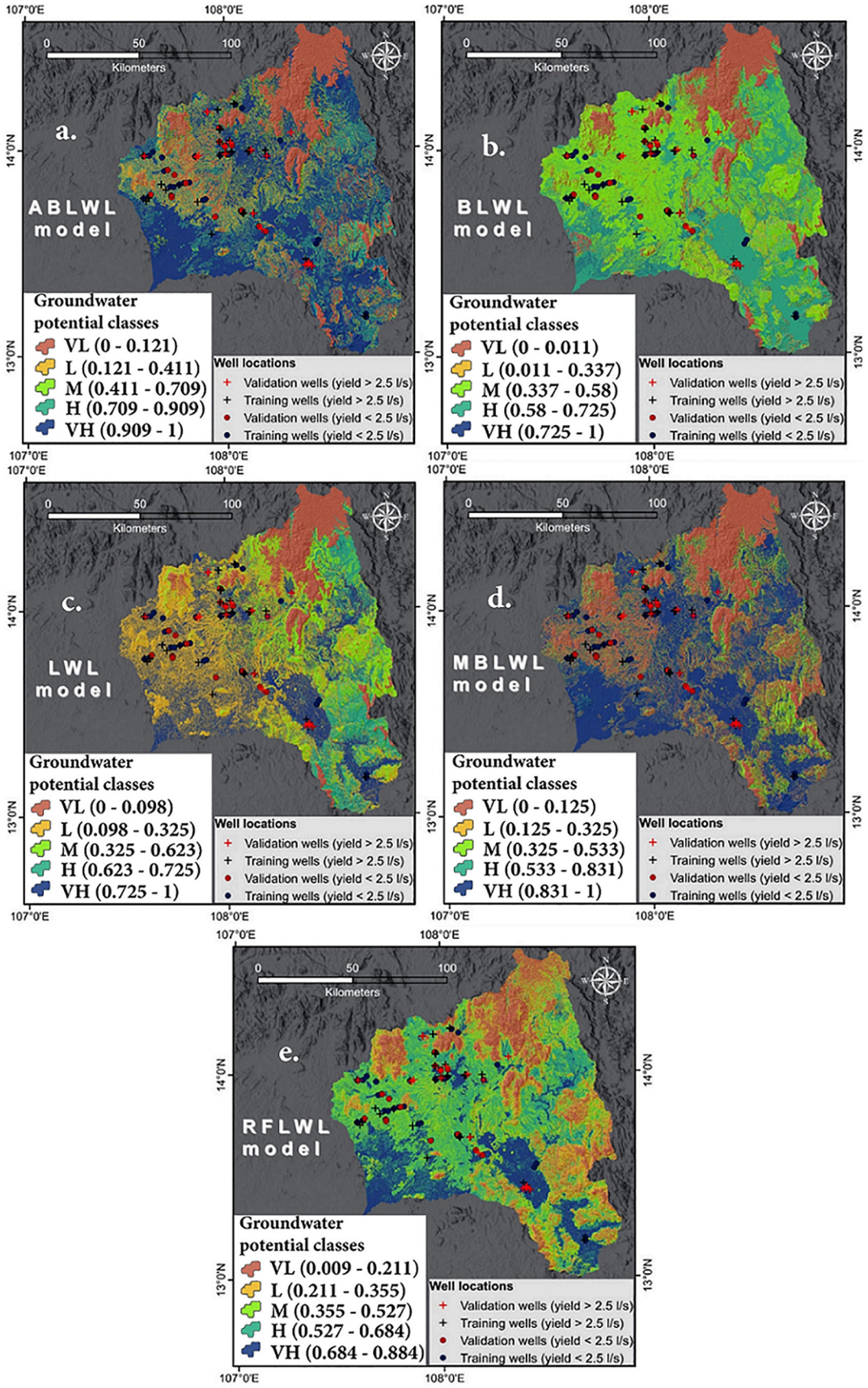

Fig. 6. Spatial extent and values of groundwater potential indices. (a) ABLWL; (b) BLWL; (c) LWL; (d) MBLWL; (e) RFLWL.

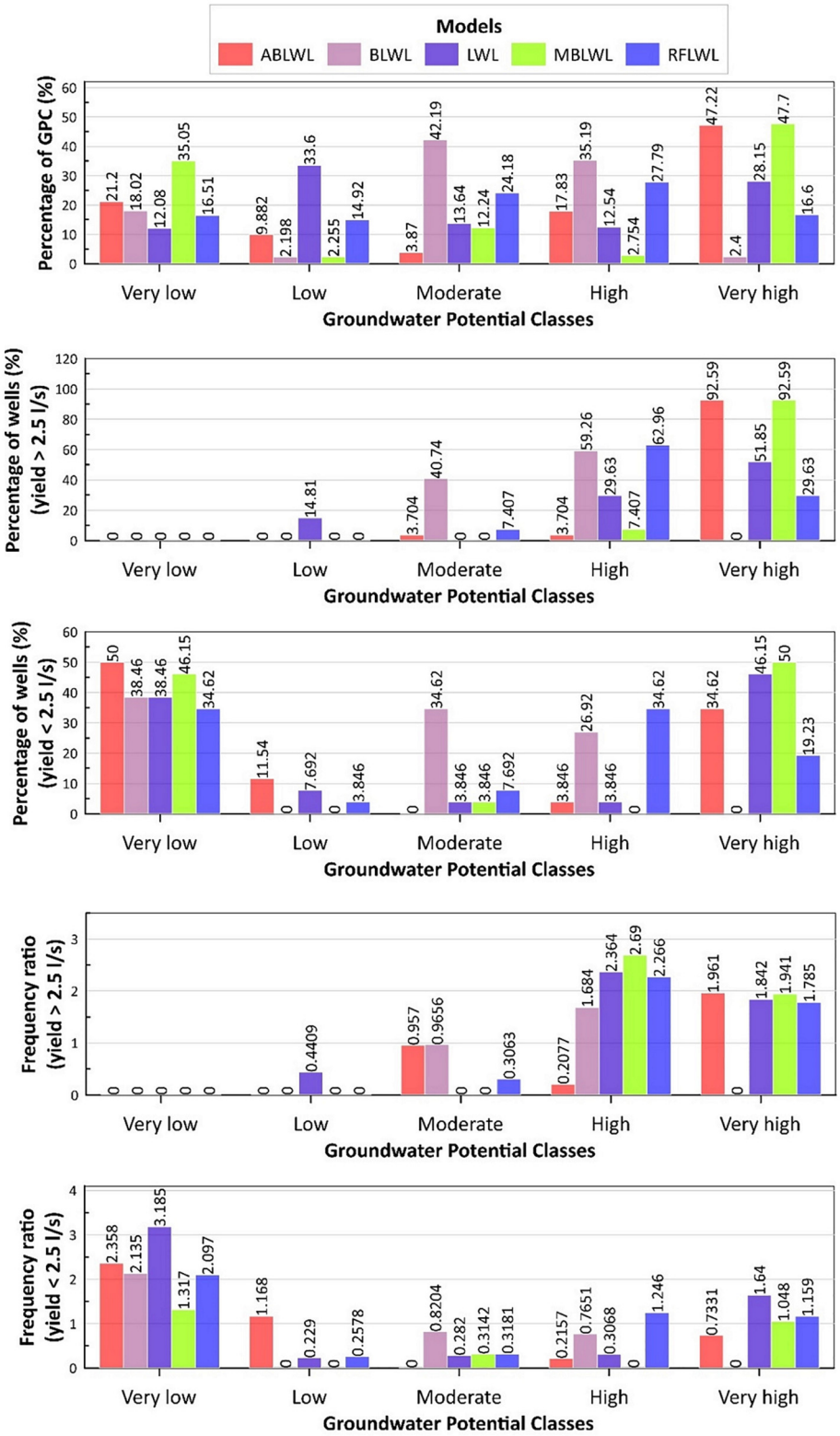

Fig. 7. Groundwater potential maps validation using the percentage of wells and their frequency ratio within the groundwater potential classes.

considered to be one of the algorithms able to improve the prediction ability of the individual model and to maintain stability of the model. In addition, along with the advantage of its simplicity, this algorithm can combine weak classifications (Shin et al., 2009; Cui et al., 2019; Tien Bui et al., 2019). According to the literature review, the training data in the MultiBoost algorithm is divided into multiple subsets in order to reduce model bias (Pham et al., 2017, 2019c). In the Bagging algorithm, the noise of input data is reduced by decreasing the sensitivities of individual classifications using the Bootstrap sampling method. It should be noted that, in terms of groundwater potential prediction,

RF can efficiently resolve unbalanced and overfitting data (Pham et al., 2019d).

6. Concluding remarks

- It is important to delineate high groundwater potential zone for the proper land use planning and water resource management of an area. In this study, four new advanced hybrid ML models (ABLWL, BLWL, MBLWL, RFLWL) and one single model LWL were applied for groundwater potential modeling and mapping at the Gia Lai province, Vietnam.
- All the proposed models performed well but ABLWL achieved the highest accuracy (AUC = 0.829) in terms of groundwater potential evaluation. Thus, ABLWL model can be used for the development of accurate groundwater potential map of the study area.
- The map developed by novel hybrid model (ABLWL) can be used by the decision-makers for the sustainable development and management of groundwater resources in conjunction with surface water for the systematic land use planning and economic development not only of the study area but also other parts of the world considering local geo-environmental conditions.
- Limitation of the study is that we have not considered detailed subsurface conditions in the models besides dynamic climate change effect affecting local and global hydrology. Model development is a continuous process and thus these factors need to be considered in future. Moving forward, other hybrid models with different algorithms need to be developed to improve further performance of the new models for the development of groundwater potential maps.

Declaration of competing interest

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

Acknowledgment

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 105.08-2019.03.

References

Sato, H., Shibasaki, N., Lap, N., Thi Kim Oanh, T., Lan, N., 2019. Characteristics on distribution of chemical composition in groundwater along the Mekong and Bassac (Hậu) river, Vietnam. Vietnam J. Earth Sci. 41 (3), 272–288. https://doi.org/10.15625/0866-7187/41/3/13969.

Abdollahi, H., Ebrahimi, S.B., 2020. A new hybrid model for forecasting Brent crude oil price. Energy 200, 117520. https://doi.org/10.1016/j.energy.2020.117520.

Adeli, E., Meng, Y., Li, G., Lin, W., Shen, D., 2017. Joint sparse and low-rank regularized multitask multi-linear regression for prediction of infant brain development with incomplete data. Med. Image Comput. Comput. Assist. Interv. 10433, 40–48.

Al-Abadi, A.M., 2015. Groundwater potential mapping at northeastern Wasit and Missan governorates, Iraq using a data-driven weights of evidence technique in framework of GIS. Environ. Earth Sci. 74 (2), 1109–1124. https://doi.org/10.1007/s12665-015-4097-0.

Andualem, T.G., Demeke, G.G., 2019. Groundwater potential assessment using GIS and remote sensing: a case study of Guna tana landscape, upper blue Nile Basin, Ethiopia. J. Hydrol. Reg. Stud. 24, 100610. https://doi.org/10.1016/j.ejrh.2019.100610.

Arabameri, A., Rezaei, K., Cerda, A., Lombardo, L., Rodrigo-Comino, J., 2019. GIS-based groundwater potential mapping in Shahroud plain, Iran. A comparison among statistical (bivariate and multivariate), data mining and MCDM approaches. Sci. Total Environ. 658, 160–177. https://doi.org/10.1016/j.scitotenv.2018.12.115.

Arabameri, A., Lee, S., Tiefenbacher, J.P., Ngo, P.T.T., 2020. Novel ensemble of MCDM-artificial intelligence techniques for groundwater-potential mapping in arid and semi-arid regions (Iran). Remote Sens. 12 (3), 490. https://doi.org/10.3390/rs12030490.

Avand, M., Janizadeh, S., Tien Bui, D., Pham, V.H., Ngo, P.T.T., Nhu, V.-H., 2020. A tree-based intelligence ensemble approach for spatial prediction of potential groundwater. Int. J. Digit. Earth 13 (12), 1408–1429. https://doi.org/10.1080/17538947.2020.1718785.

Ayazi, M.H., Pirasteh, S., Arvin, A., Pradhan, B., Nikouravan, B., Mansor, S., 2010. Disasters and risk reduction in groundwater: Zagros Mountain Southwest Iran using geoinformatics techniques. Disaster Adv. 3 (1), 51–57.

Bhuvaneswaran, C., Ganesh, A., 2019. Spatial assessment of groundwater vulnerability using DRASTIC model with GIS in Uppar odai sub-watershed, Nandiyar, Cauvery Basin, Tamil Nadu. Groundw. Sustain. Dev. 9, 100270. https://doi.org/10.1016/j.gsd.2019.100270.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140.

Bui, D.T., Ho, T.-C., Pradhan, B., Pham, B.-T., Nhu, V.-H., Revhaug, I., 2016. GIS-baseed modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, Bagging, and MultiBoost ensemble frameworks. Environ. Earth Sci. 75 (14), 1101. https://doi.org/10.1007/s12665-016-5919-4.

Bui, Q.-T., Nguyen, Q.-H., Nguyen, X.L., Pham, V.D., Nguyen, H.D., Pham, V.-M., 2020. Verification of novel integrations of swarm intelligence algorithms into deep learning neural network for flood susceptibility mapping. J. Hydrol. 581, 124379. https://doi.org/10.1016/j.jhydrol.2019.124379.

Chen, W., Li, H., Hou, E., Wang, S., Wang, G., Panahi, M., Li, T., Peng, T., Guo, C., Niu, C., Xiao, L., Wang, J., Xie, X., Ahmad, B.B., 2018. GIS-based groundwater potential analysis using novel ensemble weights-of-evidence with logistic regression and functional tree models. Sci. Total Environ. 634, 853–867. https://doi.org/10.1016/j.scitotenv.2018.04.055.

Chen, W., Hong, H., Panahi, M., Shahabi, H., Wang, Y., Shirzadi, A., Pirasteh, S., Alesheikh, A., Khosravi, K., Panahi, S., Rezaie, S., Li, S., Jaafari, A., Tien Bui, D., Ahmad, B.B., 2019. Spatial prediction of landslide susceptibility using gis-based data mining techniques of anfis with whale optimization algorithm (woa) and grey wolf optimizer (gwo). Appl. Sci. 9 (18), 3755. https://doi.org/10.3390/app9183755.

Chen, W., Zhao, X., Tsangaratos, P., Shahabi, H., Ilia, I., Xue, W., Wang, X., Ahmad, B.B., 2020. Evaluating the usage of tree-based ensemble methods in groundwater spring potential mapping. J. Hydrol. 583, 124602. https://doi.org/10.1016/j.jhydrol.2020.124602.

Choubin, B., Rahmati, O., Tahmasebipour, N., Feizizadeh, B., Pourghasemi, H.R., 2019. Application of fuzzy analytical network process model for analyzing the gully erosion susceptibility. In: Pourghasemi, H., Rossi, M. (Eds.), Natural Hazards Gis-based Spatial Modeling Using Data Mining Techniques. Springer, Cham., pp. 105–125.

Conforti, M., Aucelli, P.P., Robustelli, G., Scarciglia, F., 2011. Geomorphology and GIS analysis for mapping gully erosion susceptibility in the Turbolo stream catchment (Northern Calabria, Italy). Nat. Hazards 56 (3), 881–898. https://doi.org/10.1007/s11069-010-9598-2.

Costache, R., 2019a. Flash-Flood Potential assessment in the upper and middle sector of Prahova river catchment (Romania). A comparative approach between four hybrid models. Sci. Total Environ. 659, 1115–1134. https://doi.org/10.1016/j.scitotenv.2018.12.397.

Costache, R., 2019b. Flood susceptibility assessment by using bivariate statistics and machine learning models-a useful tool for flood risk management. Water Resour. Manag. 33 (9), 3239–3256. https://doi.org/10.1007/s11269-019-02301-z.

Costache, R., 2019c. Flash-flood potential index mapping using weights of evidence, decision trees models and their novel hybrid integration. Stoch. Env. Res. Risk A. 33 (7), 1375–1402. https://doi.org/10.1007/s00477-019-01689-9.

Costache, R., Bui, D.T., 2020. Identification of areas prone to flash-flood phenomena using multiple-criteria decision-making, bivariate statistics, machine learning and their ensembles. Sci. Total Environ. 712, 136492. https://doi.org/10.1016/j.scitotenv.2019.136492.

Costache, R., Hong, H., Pham, Q.B., 2020a. Comparative assessment of the flash-flood potential within small mountain catchments using bivariate statistics and their novel hybrid integration with machine learning models. Sci. Total Environ. 711, 134514. https://doi.org/10.1016/j.scitotenv.2019.134514.

Costache, R., Popa, M.C., Bui, D.T., Diaconu, D.C., Ciubotaru, N., Minea, G., Pham, Q.B., 2020b. Spatial predicting of flood potential areas using novel hybridizations of fuzzy decision-making, bivariate statistics, and machine learning. J. Hydrol. 585, 124808. https://doi.org/10.1016/j.jhydrol.2020.124808.

Cui, J., Li, W., Fang, C., Su, S., Luan, J., Gao, T., Hu, L., Lu, Y., Chen, G., 2019. AdaBoost ensemble correction models for TDDFT calculated absorption energies. IEEE Access 7, 38397–38406. https://doi.org/10.1109/ACCESS.2019.2905928.

Dempster, A.P., 2008. Upper and lower probabilities induced by a multivalued mapping. In: Yager, R.R., Liu, L. (Eds.), Classic Works of the Dempster-Shafer Theory of Belief Functions. vol. 219. Springer, Berlin, Heidelberg, pp. 57–72.

Ercanoglu, M., Gokceoglu, C., 2002. Assessment of landslide susceptibility for a landslide-prone area (north of Yenice, NW Turkey) by fuzzy approach. Environ. Geol. 41 (6), 720–730. https://doi.org/10.1007/s00254-001-0454-2.

Flentge, F., 2006. Locally weighted interpolating growing neural gas. IEEE Trans. Neural. Netw. Learn Syst. 17 (6), 1382–1393. https://doi.org/10.1109/TNN.2006.879771.

Golkarian, A., Naghibi, S.A., Kalantar, B., Pradhan, B., 2018. Groundwater potential mapping using C5.0, random forest, and multivariate adaptive regression spline models in GIS. Environ. Monit. Assess. 190 (3), 149. https://doi.org/10.1007/s10661-018-6507-8.

Hall, M.A., 2000. Correlation-based feature selection for discrete and numeric class machine learning. Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., pp. 359–366.

Hasegawa, T., Fujimori, S., Ito, A., Takahashi, K., Masui, T., 2017. Global land-use allocation model linked to an integrated assessment model. Sci. Total Environ. 580, 787–796. https://doi.org/10.1016/j.scitotenv.2016.12.025.

Hoa, P.V., Giang, N.V., Binh, N.A., Hai, L.V.H., Pham, T.-D., Hasanlou, M., Bui, D.T., 2019. Soil salinity mapping using SAR Sentinel-1 data and advanced machine learning algorithms: a case study at Ben Tre province of the Mekong River Delta (Vietnam). Remote Sens. 11 (2), 128. https://doi.org/10.3390/rs11020128.

Hong, H., Liu, J., Bui, D.T., Pradhan, B., Acharya, T.D., Pham, B.T., Zhu, A.X., Chen, W., Ahmad, B.B., 2018. Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China). Catena 163, 399–413. https://doi.org/10.1016/j.catena.2018.01.005.

Jaafari, A., Zenner, E.K., Pham, B.T., 2018. Wildfire spatial pattern analysis in the Zagros Mountains, Iran: a comparative study of decision tree based classifiers. Ecol. Inform. 43, 200–211. https://doi.org/10.1016/j.ecoinf.2017.12.006.

Jaafari, A., Mafi-Gholami, D., Thai Pham, B., Tien Bui, D., 2019. Wildfire probability mapping: bivariate vs. multivariate statistics. Remote Sens. 11 (6), 618. https://doi.org/10.3390/rs11060618.

Jiang, H., Zheng, W., Luo, L., Dong, Y., 2019. A two-stage minimax concave penalty based method in pruned AdaBoost ensemble. Appl. Soft Comput. 83, 105674. https://doi.org/10.1016/j.asoc.2019.105674.

Khattak, S.A., Rashid, A., Tariq, M., Ali, L., Gao, X., Ayub, M., Javed, A., 2020. Potential risk and source distribution of groundwater contamination by mercury in district Swabi, Pakistan: application of multivariate study. Environ. Dev. Sustain. 1-19. https://doi.org/10.1007/s10668-020-00674-5.

Khosravi, K., Pham, B.T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., Prakash, I., Bui, D.T., 2018. A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. Sci. Total Environ. 627, 744–755. https://doi.org/10.1016/j.scitotenv.2018.01.266.

Khosravi, K., Barzegar, R., Miraki, S., Adamowski, J., Daggupati, P., Alizadeh, M.R., Pham, B.T., Alami, M.T., 2019. Stochastic modeling of groundwater fluoride contamination: introducing lazy learners. Groundwater 58 (5), 723–734. https://doi.org/10.1111/gwat.12963.

Koh, E.-H., Lee, E., Lee, K.-K., 2020. Application of geographically weighted regression models to predict spatial characteristics of nitrate contamination: implications for an effective groundwater management strategy. J. Environ. Manag. 268, 110646. https://doi.org/10.1016/j.jenvman.2020.110646.

Kordestani, M.D., Naghibi, S.A., Hashemi, H., Ahmadi, K., Kalantar, B., Pradhan, B., 2019. Groundwater potential mapping using a novel data-mining ensemble model. Hydrogeol. J. 27 (1), 211–224. https://doi.org/10.1007/s10040-018-1848-5.

Ly, P.T., Thuy, H.L.T., 2019. Spatial distribution of hot days in north central region, Vietnam in the period of 1980–2013. Vietnam J. Earth Sci. 41 (1), 36–45. https://doi.org/10.15625/0866-7187/41/1/13544.

Malakootian, M., Mohammadi, A., Faraji, M., 2020. Investigation of physicochemical parameters in drinking water resources and health risk assessment: a case study in NW Iran. Environ. Earth Sci. 79 (9), 195. https://doi.org/10.1007/s12665-020-08939-y.

Minh, P.T., Tuyet, B.T., Thao, T.T.T., Hang, L.T.T., 2018. Application of ensemble Kalman filter in WRF model to forecast rainfall on monsoon onset period in South Vietnam. Vietnam J. Earth Sci. 40 (4), 367–394. https://doi.org/10.15625/0866-7187/40/4/13134.

Miraki, S., Zanganeh, S.H., Chapi, K., Singh, V.P., Shirzadi, A., Shahabi, H., Pham, B.T., 2019. Mapping groundwater potential using a novel hybrid intelligence approach. Water Resour. Manag. 33 (1), 281–302. https://doi.org/10.1007/s11269-018-2102-6.

Mogaji, K.A., Lim, H.S., 2018. Application of Dempster-Shafer theory of evidence model to geoelectric and hydraulic parameters for groundwater potential zonation. NRIAG J. Astron. Geophys. 7 (1), 134–148. https://doi.org/10.1016/j.nrjag.2017.12.008.

Mosavi, A., Hosseini, F.S., Choubin, B., Goodarzi, M., Dineva, A.A., 2020a. Groundwater salinity susceptibility mapping using classifier ensemble and bayesian machine learning models. IEEE Access 8, 145564–145576. https://doi.org/10.1109/ACCESS.2020.3014908.

Mosavi, A., Sajedi-Hosseini, F., Choubin, B., Taromideh, F., Rahi, G., Dineva, A.A., 2020b. Susceptibility mapping of soil water erosion using machine learning models. Water 12 (7), 1995. https://doi.org/10.3390/w12071995.

Naghibi, S.A., Pourghasemi, H.R., 2015. A comparative assessment between three machine learning models and their performance comparison by bivariate and multivariate statistical methods in groundwater potential mapping. Water Resour. Manag. 29 (14), 5217–5236. https://doi.org/10.1007/s11269-015-1114-8.

Naghibi, S.A., Ahmadi, K., Daneshi, A., 2017a. Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. Water Resour. Manag. 31 (9), 2761–2775. https://doi.org/10.1007/s11269-017-1660-3.

Naghibi, S.A., Moghaddam, D.D., Kalantar, B., Pradhan, B., Kisi, O., 2017b. A comparative assessment of GIS-based data mining models and a novel ensemble model in groundwater well potential mapping. J. Hydrol. 548, 471–483. https://doi.org/10.1016/j.jhydrol.2017.03.020.

Nguyen, V.V., Pham, B.T., Vu, B.T., Prakash, I., Jha, S., Shahabi, H., Shirzadi, A., Ba, D.N., Kumar, R., Chatterjee, J.M., Tien Bui, D., 2019. Hybrid machine learning approaches for landslide susceptibility modeling. Forests 10 (2), 157. https://doi.org/10.3390/f10020157.

Nguyen, P.T., Ha, D.H., Jaafari, A., Nguyen, H.D., Van Phong, T., Al-Ansari, N., Prakash, I., Le, H.V., Pham, B.T., 2020a. Groundwater potential mapping combining artificial neural network and real adaboost ensemble technique: the Daknong province case- study, Vietnam. Int. J. Environ. Res. Public Health 17 (7), 2473. https://doi.org/10.3390/ijerph17072473.

Nguyen, P.T., Ha, D.H., Avand, M., Jaafari, A., Nguyen, H.D., Al-Ansari, N., Phong, T.V., Sharma, R., Kumar, R., Le, H.V., Le, H.P., Ho, L.S., Prakash, I., Pham, B.T., 2020b. Soft computing ensemble models based on logistic regression for groundwater potential mapping. Appl. Sci. 10 (7), 2469. https://doi.org/10.3390/app10072469.

Nguyen, P.T., Ha, D.H., Nguyen, H.D., Van Phong, T., Trinh, P.T., Al-Ansari, N., Le, H.V., Pham, B.T., Ho, L.S., Prakash, I., 2020c. Improvement of credal decision trees using ensemble frameworks for groundwater potential modeling. Sustainability 12 (7), 2622. https://doi.org/10.3390/su12072622.

Oanh, T.T.K., Van Lap, N., 2016. High arsenic concentration in groundwater related to sedimentary facies in the Mekong River Delta, Vietnam. Vietnam J. Earth Sci. 38 (2), 178–187. https://doi.org/10.15625/0866-7187/38/2/8600.

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. J. Hydrol. 399 (3–4), 158–172. https://doi.org/10.1016/j.jhydrol.2010.12.027.

Oikonomidis, D., Dimogianni, S., Kazakis, N., Voudouris, K., 2015. A GIS/remote sensing-based methodology for groundwater potentiality assessment in Tirnavos area, Greece. J. Hydrol. 525, 197–208. https://doi.org/10.1016/j.jhydrol.2015.03.056.

Ozcift, A., Gulten, A., 2011. Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. Comput. Methods Prog. Biomed. 104 (3), 443–451. https://doi.org/10.1016/j.cmpb.2011.03.018.

Ozdemir, A., 2011. GIS-based groundwater spring potential mapping in the Sultan Mountains (Konya, Turkey) using frequency ratio, weights of evidence and logistic regression methods and their comparison. J. Hydrol. 411 (3–4), 290–308. https://doi.org/10.1016/j.jhydrol.2011.10.010.

Pham, B.T., Bui, D.T., Prakash, I., Dholakia, M., 2017. Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. Catena 149, 52–63. https://doi.org/10.1016/j.catena.2016.09.007.

Pham, B.T., Jaafari, A., Prakash, I., Singh, S.K., Quoc, N.K., Bui, D.T., 2019a. Hybrid computational intelligence models for groundwater potential mapping. Catena 182, 104101. https://doi.org/10.1016/j.catena.2019.104101.

Pham, B.T., Jaafari, A., Prakash, I., Bui, D.T., 2019b. A novel hybrid intelligent model of support vector machines and the MultiBoost ensemble for landslide susceptibility modeling. Bull. Eng. Geol. Environ. 78 (4), 2865–2886. https://doi.org/10.1007/s10064-018-1281-y.

Pham, B.T., Prakash, I., Dou, J., Singh, S.K., Trinh, P.T., Tran, H.T., Le, T.M., Van Phong, T., Khoi, D.K., Shirzadi, A., Tien Bui, D., 2019c. A novel hybrid approach of landslide susceptibility modelling using rotation forest ensemble and different base classifiers. Geocarto Int. 35 (12), 1267–1292. https://doi.org/10.1080/10106049.2018.1559885.

Pham, B.T., Prakash, I., Singh, S.K., Shirzadi, A., Shahabi, H., Bui, D.T., 2019d. Landslide susceptibility modeling using Reduced Error Pruning Trees and different ensemble techniques: hybrid machine learning approaches. Catena 175, 203–218. https://doi.org/10.1016/j.catena.2018.12.018.

Phong, T.V., Phan, T.T., Prakash, I., Singh, S.K., Shirzadi, A., Chapi, K., Ly, H.-B., Ho, L.S., Quoc, N.K., Pham, B.T., 2019. Landslide susceptibility modeling using different artificial intelligence methods: a case study at Muong Lay district, Vietnam. Geocarto Int., 1–24 https://doi.org/10.1080/10106049.2019.1665715.

Phuc, L.T., Tachihara, H., Honda, T., Tuat, L.T., Thom, B.V., Hoang, N., Chikano, Y., Yoshida, K., Tung, N.T., Danh, P.N., Hung, N.B., Duc, T.M., Vu, P.G.M., Hoa, N.T.M., Bien, H.T., Quy, T.Q., Minh, N.T., 2018. Geological values of lava caves in Krongno volcano geopark, Dak Nong, Vietnam. Vietnam J. Earth Sci. 40 (4), 299–319. https://doi.org/10.15625/0866-7187/40/4/13101.

Qi, C., Ly, H.-B., Chen, Q., Le, T.-T., Le, V.M., Pham, B.T., 2020. Flocculation-dewatering prediction of fine mineral tailings using a hybrid machine learning approach. Chemosphere 244, 125450. https://doi.org/10.1016/j.chemosphere.2019.125450.

Quyen, N.T.N., Liem, N.D., Loi, N.K., 2014. Effect of land use change on water discharge in Srepok watershed, Central Highland, Viet Nam. Int. Soil Water Conserv. Res. 2 (3), 74–86. https://doi.org/10.1016/S2095-6339(15)30025-3.

Rahman, A., 2008. A GIS based DRASTIC model for assessing groundwater vulnerability in shallow aquifer in Aligarh, India. Appl. Geogr. 28 (1), 32–53. https://doi.org/10.1016/j.apgeog.2007.07.008.

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. https://doi.org/10.1016/j.catena.2015.10.010.

Rahmati, O., Naghibi, S.A., Shahabi, H., Bui, D.T., Pradhan, B., Azareh, A., Rafiei-Sardooi, E., Samani, A.N., Melesse, A.M., 2018. Groundwater spring potential modelling: comprising the capability and robustness of three different modeling approaches. J. Hydrol. 565, 248–261. https://doi.org/10.1016/j.jhydrol.2018.08.027.

Rahmati, O., Moghaddam, D.D., Moosavi, V., Kalantari, Z., Samadi, M., Lee, S., Tien Bui, D., 2019. An automated python language-based tool for creating absence samples in groundwater potential mapping. Remote Sens. 11 (11), 1375. https://doi.org/10.3390/rs11111375.

Rajesh, K.N.V.P.S., Dhuli, R., 2018. Classification of imbalanced ECG beats using re-sampling techniques and AdaBoost ensemble classifier. Biomed. Signal Process. Control 41, 242–254. https://doi.org/10.1016/j.bspc.2017.12.004.

Rizeei, H.M., Pradhan, B., Saharkhiz, M.A., Lee, S., 2019. Groundwater aquifer potential modeling using an ensemble multi-adoptive boosting logistic regression technique. J. Hydrol. 579, 124172. https://doi.org/10.1016/j.jhydrol.2019.124172.

Sajedi-Hosseini, F., Malekian, A., Choubin, B., Rahmati, O., Cipullo, S., Coulon, F., Pradhan, B., 2018. A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. Sci. Total Environ. 644, 954–962. https://doi.org/10.1016/j.scitotenv.2018.07.054.

Sameen, M.I., Pradhan, B., Lee, S., 2019. Self-learning random forests model for mapping groundwater yield in data-scarce areas. Nat. Resour. Res. 28 (3), 757–775. https://doi.org/10.1007/s11053-018-9416-1.

Schaal, S., Atkeson, C.G., Vijayakumar, S., 2000. Real-time robot learning with locally weighted statistical learning. Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065). IEEE, pp. 288–293.

Shahid, S., Nath, S., Roy, J., 2000. Groundwater potential modelling in a soft rock area using a GIS. Int. J. Remote Sens. 21 (9), 1919–1924. https://doi.org/10.1080/014311600209823.

Shin, Y., Kim, D.W., Kim, J.Y., Kang, K.I., Cho, M.Y., Cho, H.H., 2009. Application of AdaBoost to the retaining wall method selection in construction. J. Comput. Civ. Eng. 23 (3), 188–192.

Šimanský, V., Horák, J., Juriga, M., Srank, D., 2018. Soil structure and soil organic matter in water-stable aggregates under different application rates of biochar. Vietnam J. Earth Sci. 40 (2), 97–108. https://doi.org/10.15625/0866-7187/40/2/11090.

Sivasankar, T., Lone, J.M., Sarma, K., Qadir, A., Raju, P., 2019. Estimation of above ground biomass using support vector machines and ALOS/PALSAR data. Vietnam J. Earth Sci. 41 (2), 95–104. https://doi.org/10.15625/0866-7187/41/2/13690.

Solomon, S., Quiel, F., 2006. Groundwater study using remote sensing and geographic information systems (GIS) in the central highlands of Eritrea. Hydrogeol. J. 14 (6), 1029–1041. https://doi.org/10.1007/s10040-006-0096-2.

Subasi, A., Kadasa, B., Kremic, E., 2020. Classification of the cardiotocogram data for anticipation of fetal risks using bagging ensemble classifier. Proc. Comput. Sci. 168, 34–39. https://doi.org/10.1016/j.procs.2020.02.248.

Subramani, T., Elango, L., Damodarasamy, S.R., 2005. Groundwater quality and its suitability for drinking and agricultural use in Chithar River Basin, Tamil Nadu, India. Environ. Geol. 47 (8), 1099–1110. https://doi.org/10.1007/s00254-005-1243-0.

Tien Bui, D., Shirzadi, A., Chapi, K., Shahabi, H., Pradhan, B., Pham, B.T., Singh, V.P., Chen, W., Khosravi, K., Bin Ahmad, B., 2019. A hybrid computational intelligence approach to groundwater spring potential mapping. Water 11 (10), 2013. https://doi.org/10.3390/w11102013.

Tran, Q.C., Minh, D.D., Jaafari, A., Al-Ansari, N., Minh, D.D., Van, D.T., Nguyen, D.A., Tran, T.H., Ho, L.S., Nguyen, D.H., Prakash, I., Le, H.P., Pham, B.T., 2020. Novel ensemble landslide predictive models based on the hyperpipes algorithm: a case study in the Nam Dam Commune, Vietnam. Appl. Sci. 10 (11), 3710. https://doi.org/10.3390/app10113710.

Trung, D.T., Nhan, N.T., Don, V.T., Hung, N.K., Kazmierczak, J., Nhan, P.Q., 2020. The controlling of paleo-riverbed migration on Arsenic mobilization in groundwater in the Red River Delta, Vietnam. Vietnam J. Earth Sci. 42 (2), 161–175. https://doi.org/10.15625/0866-7187/42/2/14998.

Tuan, N.T., Chi, T.T., Van, Y.T., Mung, V.T., 2019. Recreational and conservative valuation of Bien Ho landscape. Vietnam J. Earth Sci. 41 (2), 156–172. https://doi.org/10.15625/0866-7187/41/2/13729.

Van Hoang, N., Van, D.T., Hoa, P.L., 2020. Heavy metal contamination of soil based on pollution, geo-accumulation indices and enrichment factor in Phan Me coal mine area, Thai Nguyen province, Vietnam. Vietnam J. Earth Sci. 42 (2), 105–117.

Wang, G., Lei, X., Chen, W., Shahabi, H., Shirzadi, A., 2020. Hybrid computational intelligence methods for landslide susceptibility mapping. Symmetry 12 (3), 325. https://doi.org/10.3390/sym12030325.

Webb, G.I., 2000. MultiBoosting: a technique for combining boosting and wagging. Mach. Learn. 40 (2), 159–196. https://doi.org/10.1023/A:1007659514849.

Wu, B., Ai, H., Huang, C., Lao, S., 2004. Fast rotation invariant multi-view face detection based on real Adaboost. Proceedings of Sixth IEEE International Conference on Automatic Face and Gesture Recognition. IEEE, pp. 79–84.

Wu, Y., Ke, Y., Chen, Z., Liang, S., Zhao, H., Hong, H., 2020. Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping. Catena 187, 104396. https://doi.org/10.1016/j.catena.2019.104396.

Yariyan, P., Janizadeh, S., Van Phong, T., Nguyen, H.D., Costache, R., Van Le, H., Pham, B.T., Pradhan, B., Tiefenbacher, J.P., 2020. Improvement of best first decision trees using bagging and dagging ensembles for flood probability mapping. Water Resour. Manag. 34 (9), 3037–3053. https://doi.org/10.1007/s11269-020-02603-7.

Zheng, X.-X., Peng, P., 2019. Fault diagnosis of wind power converters based on compressed sensing theory and weight constrained Adaboost-SVM. J. Power Electron. 19 (2), 443–453. https://doi.org/10.6113/JPE.2019.19.2.443.