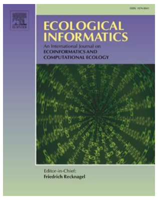
Contents lists available at ScienceDirect

Ecological Informatics

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



Naïve Bayes ensemble models for groundwater potential mapping

Binh Thai Pham [a,b,*], Abolfazl Jaafari [c,*], Tran Van Phong [d], Davood Mafi-Gholami [e], Mahdis Amiri [f], Nguyen Van Tao [d], Van-Hao Duong [g], Indra Prakash [h]

[a] *University of Transport Technology, 54 Trieu Khuc, Thanh Xuan, Hanoi 100000, Viet Nam*
[b] *Civil and Environmental Engineering Program, Graduate School of Advanced Science and Engineering, Hiroshima University, 1-4-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8527, Japan*
[c] *Research Institute of Forests and Rangelands, Agricultural Research, Education, and Extension Organization (AREEO), Tehran 1496813111, Iran*
[d] *Institute of Geological Sciences, Vietnam Academy of Science and Technology, 84 Chua Lang, Dong Da, Hanoi, Viet Nam*
[e] *Department of Forest Sciences, Faculty of Natural Resources and Earth Sciences, Shahrekord University, Shahrekord 8818634141, Iran*
[f] *Department of Watershed and Arid Zone Management, Gorgan University of Agricultural Sciences and Natural Resources, Gorgan 4918943464, Iran*
[g] *Hanoi University of Mining and Geology. No 18, Vien street, Bac Tu Liem district, Hanoi, Viet Nam*
[h] *DDG (R) Geological Survey of India, Gandhinagar 382010, India*

ARTICLE INFO

Keywords:
Machine learning
Ensemble modeling
Naïve Bayes
Bagging
AdaBoost
Rotation Forest

ABSTRACT

Groundwater potential maps are important tools for the sustainable management of water resources, especially in agricultural producing countries like Vietnam. Here, we describe the development and application of a spatially explicit ensemble modeling framework that allows for analyzing spatially explicit data for estimating groundwater potential across the Kon Tum Province, Vietnam. Based on this framework, the Naïve Bayes (NB) method was integrated with the Bagging (B), AdaBoost (AB), and Rotation Forest (RF) ensemble learning techniques to develop three ensemble models, namely BNB, ABNB, and RFNB. A suite of well yield data and thirteen explanatory variables (i.e., elevation, aspect, slope, curvature, river density, topographic wetness index, sediment transport index, soil type, geology, land use, rainfall, and flow direction and accumulation) were incorporated into the modeling processes over the independent training and validation levels of the single NB model and its three ensembles. Several performance metrics (i.e., area under the receiver operating characteristic curve (AUC), root mean square error (RMSE), accuracy, sensitivity, specificity, negative predictive value, and positive predictive value) demonstrated that the three ensemble models successfully surpassed the single NB model in groundwater potential mapping. The ensemble RFNB model with AUC = 0.849, accuracy = 83.33%, sensitivity = 100%, specificity = 75%, and RMSE = 0.406 exhibited the most accurate performance for mapping groundwater potential in the Kon Tum Province, followed by the ABNB (AUC = 0.844), BNB (AUC = 0.815), and single NB (AUC = 0.786) models, respectively. Further, the correlation based feature selection method identified elevation, slope, land use, rainfall, and STI as the most useful explanatory variables for explaining the distribution of groundwater potential in the Kon Tum Province. The methodology proposed in this case study and the produced potential maps enable managers to align water use patterns with the shared benefits and costs of different users and to develop strategies for sustainable groundwater exploitation, preservation, and management.

Abbreviations: Naïve Bayes, (NB); Bagging, (B); AdaBoost, (AB); Rotation forest, (RF); Bagging-Naïve Bayes, (BNB); AdaBoost-Naïve Bayes, (ABNB); Rotation Forest-Naïve Bayes, (RFNB); area under curve, (AUC); artificial neural network, (ANN); adaptive neuro fuzzy inference system, (ANFIS); support vector machine, (SVM); support vector regression, (SVR); boosted regression tree, (BRT); classification and regression tree, (CART); multivariate adaptive regression spline, (MARS); correlation based feature selection method, (CBFS); Vietnam Academy for Water Resources, (VAWR); topographic wetness index, (TWI); sediment transport index, (STI); digital elevation model, (DEM); Vietnam Meteorological Organization, (VMO); Water Resources Planning and Investigation of Vietnam, (WRPI); average merit, (AM); principal component analysis, (PCA); positive predictive value, (PPV); negative predictive value, (NPV); receiver operating characteristics, (ROC) curve; root mean square error, (RMSE); true positive, (TP); false positive, (FP); false negative, (FN); true negative, (TN); Reduced Pruning Error Tree, (RPET); Random Subspace, (RSS)..

* Corresponding authors.
E-mail addresses: binhpt@utt.edu.vn (B.T. Pham), jaafari@rifr-ac.ir (A. Jaafari).

https://doi.org/10.1016/j.ecoinf.2021.101389
Received 8 May 2021; Received in revised form 26 June 2021; Accepted 26 July 2021
Available online 31 July 2021

1. Introduction

Groundwater stores about 30% of the global freshwater, supplies about 40% of the water for global irrigated farmlands, and constitutes the main source of drinking water for more than two billion people around the world (Jasechko et al., 2017; Siebert et al., 2010). With the current ever-increasing freshwater demand, overexploitation has caused a widespread decline in groundwater table (Pandey et al., 2020) which in turn exacerbated land degradation and desertification worldwide (Guo and Shen, 2016; Ma et al., 2020; MacDonald et al., 2021; Rodell et al., 2018). Further, the ongoing climate change mirrored in prolonged drought occurrences is expected to amplify the already water scarcity in many regions around the world. To counteract these threats and to develop prudent strategies for sustainable groundwater exploitation, preservation, and management, managers and scientists need to urgently tackle the critical problems caused by overexploitation of groundwater aquifers. Among different solutions and strategies, spatially explicit mapping of groundwater potential is prominent that enables managers to align water use patterns with the shared benefits and costs of different users. Groundwater potential that indicates the probability of groundwater occurrence/presence or the content of groundwater reservoir in a region (Khosravi et al., 2018; Rahmati and Melesse, 2016) is an essential decision support tool, particularly for regions vulnerable to the impacts of climate change and where aquifers are already depleted due to excessive groundwater extraction and/or inappropriate aquifer recharge (Gaur et al., 2011; He et al., 2021), as well as data-scarce regions where hydraulic and pumping data are not available. For these regions, groundwater potential mapping using field-based yield data that characterize extraction volume and the velocity of groundwater at several measurement points is important (Lee et al., 2020).

Some common groundwater measurement methods such as hydrogeological techniques, field hydraulic conducting surveys, and exploratory drilling are not only very time consuming and prohibitive/costly, but also are limited to small-scale, sparse data obtained from point measurements (Becker, 2006) that have made it difficult to map groundwater potential. In recent years, spatially explicit data obtained via remote sensing have been suggested and used to overcome the limitation of local information. Using remote sensing techniques, point-based local-scale data can be extended to large, spatially distributed data sets that are appropriate for analysis using different geographic information systems (GISs) and data processing methods such as machine learning algorithms (Avand et al., 2020; Choubin and Rahmati, 2021; Razavi-Termeh et al., 2019; Tolche, 2021).

Spatial pattern analysis and mapping of groundwater potential based on the spatially explicit data is an attractive alternative to difficult conventional approaches (Kalhor et al., 2019). In recent years, many researchers have investigated and measured groundwater potential in different regions around the world. In these studies, remotely sensed data and GIS techniques have been used and different methods from machine learning have been evaluated (Agarwal et al., 2019; Nguyen et al., 2020). Unlike the traditional field-based methods adopted for groundwater potential mapping, machine learning methods find patterns of groundwater reservoirs in diverse hydrogeological settings and use these patterns to predict where another groundwater can exist (Pham et al., 2019a). In recent years, many accurate predictive models have been derived from machine learning methods that allowed for handling missing data, outliers, and various types of geo-environmental variables (DeSimone et al., 2020) and provided accurate estimations of groundwater potential for many regions around the world. Artificial neural network (ANN), adaptive neuro fuzzy inference system (ANFIS), random forest (RF), support vector machine (SVM), support vector regression (SVR), boosted regression tree (BRT), classification and regression tree (CART), and multivariate adaptive regression spline (MARS) are among the popular and proficient machine learning methods for groundwater potential mapping (Choubin and Rahmati, 2021; Fadhillah et al., 2021; Motevalli et al., 2019; Naghibi et al., 2017a; Naghibi and Moradi Dashtpagerdi, 2017; Yen et al., 2021; Zhu and Abdelkareem, 2021). As a result of these machine learning applications, groundwater potential mapping is now faster, easier, cheaper, and more accurate than ever before. Concurrently, the development and validation of new models have become mainstream in this field of science, with an increasing number of researchers adopting various optimization algorithms and ensemble learning techniques to develop hybrid and ensemble predictive models to improve the capability of a base model for groundwater potential mapping (Nguyen et al., 2020; Pham et al., 2019a). A hybrid or an ensemble model indicates a model that combines at least two methods to yield more accurate results than a standalone method. Examples of the most recent efforts for hybrid and ensemble modeling of groundwater potential can be found in recent works published in the literature (Al-Fugara et al., 2020; Motevalli et al., 2019; Naghibi et al., 2019; Nguyen et al., 2020; Pham et al., 2019a).

Motivated by a desire to better understand how different methods/models interpret spatially explicit information for proving accurate and unbiased predictions of groundwater potential, we modeled groundwater potential in the Kon Tum Province, Vietnam, using the Naïve Bayes machine learning method that relies on the AdaBoost, Bagging, and Rotation Forest ensemble learning techniques. Thirteen explanatory variables are incorporated into the Naïve Bayes method to characterize the geoenvironmental conditions of the study area. Various validation metrics are used to investigate how well the models conform to ground truth data of groundwater potential and to measure the uncertainty and limitations of predictions. By proposing three Naïve Bayes-based ensemble models, our study contributes to the suite of research that seeks to provide managers with decision support systems for sustainable groundwater exploitation, preservation, and management.

2. Research area

The Kon Tum Province covers an area of 9676.5 Km^2 in the central part (13°55′10“ N to 15°27’15” N and 107°20′15“ E 108°32’30” E) of Vietnam (Fig. 1). The province has a population of approximately 530,000 and is predominantly mountainous with an average elevation and slope of 600 m and 46 degrees, respectively. The climate of the province is monsoon tropical that is characterized by two separate seasons, i.e., rainy and dry. The rainy season typically extends from April to November, whereas the dry season extends from December to March of the next year. The average annual rainfall is about 2121 mm. The province suitability permits the cultivation of a variety of industrial crops and medicinal plants such as coffee, passion fruit, and Ngoc Linh ginseng. The effects of two rainy and dry seasons on water resources make social and economic activities more dependent on groundwater extraction. The groundwater source in the Kon Tum province has potential and industrial reserves of a C2 level (Vietnamese classification standard): 100 thousand m^3/day, especially at a depth of 60–300 m with relatively large reserves.

3. Modeling methodology

Spatial modeling of groundwater potential in the Kon Tum Province, Vietnam, involves the following six main steps (Fig. 2):

1) Collecting the locations of groundwater reservoirs (wells) through multiple field surveys. we randomly divided these locations into two groups such that 70% (42 wells) were used for model training and building and the remaining locations (18 wells = 30%) were reserved for model validation (Al-Fugara et al., 2020; Nguyen et al., 2019; Pham and Prakash, 2019).
2) Collecting spatially explicit data related to a set of geoenvironmental variables that are supposed to directly or indirectly influence the distribution of groundwater potential within the Kon Tum Province.

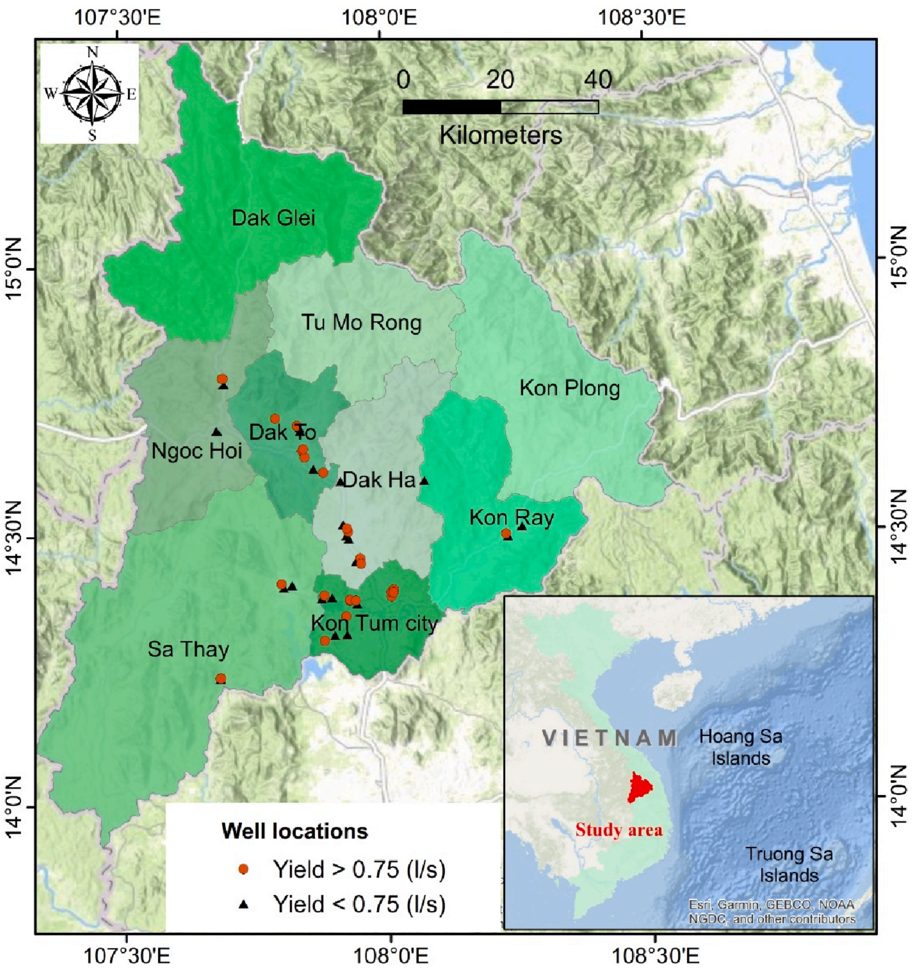


Fig. 1. Location of the Kon Tum Province (study area).

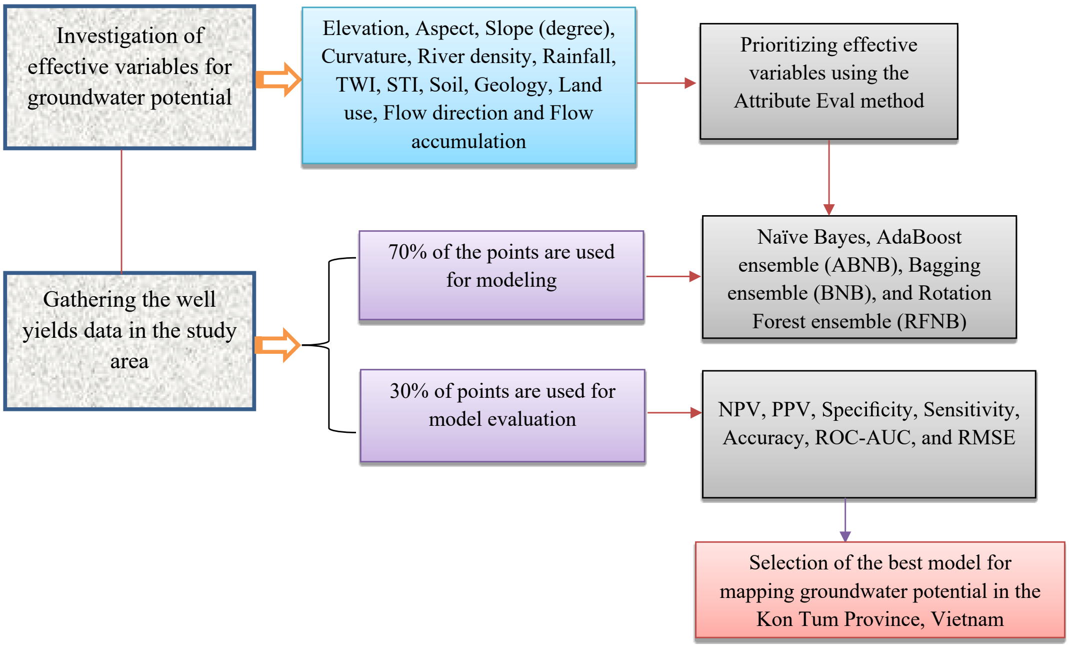

Investigation of effective variables for groundwater potential
Elevation, Aspect, Slope (degree), Curvature, River density, Rainfall, TWI, STI, Soil, Geology, Land use, Flow direction and Flow accumulation
Prioritizing effective variables using the Attribute Eval method
Naïve Bayes, AdaBoost ensemble (ABNB), Bagging ensemble (BNB), and Rotation Forest ensemble (RFNB)
Gathering the well yields data in the study area
70% of the points are used for modeling
30% of points are used for model evaluation
NPV, PPV, Specificity, Sensitivity, Accuracy, ROC-AUC, and RMSE
Selection of the best model for mapping groundwater potential in the Kon Tum Province, Vietnam

Fig. 2. Modeling strategy for groundwater potential mapping in the Kon Tum Province, Vietnam.

3) Selecting the most significant variables influencing groundwater potential using the correlation based feature selection (CBFS) method.
4) Spatially explicit modeling of groundwater potential using ensemble models. This step that is the core of our methodology combines the Naïve Bayes classifier with the AdaBoost, Bagging, and Rotation Forest techniques that results in three ensemble models: ABNB, BNB, and FRNB.
5) Validating the training (i.e., goodness-of-fit) and validation (i.e., predictive ability) performances of the models using various metrics.
6) Generating groundwater potential maps using the models' outputs. The sixth and final step of our methodology was conducted via two main phases: 1) whole pixels of the study region were fed to the validated models to produce the potential indices for the entire study region. 2) The generated potential indices were reclassified and ground into five classes to visually display the groundwater potential maps.

3.1. Data collection

3.1.1. Well yields

Sixty drilled wells data were collected from the national projects conducted by the Vietnam Academy for Water Resources (VAWR) and were used in this study. Out of these, 30 wells with groundwater yield of >0.75 l/s were considered as "groundwater" class, and the remaining wells with groundwater yield of <0.75 l/s were considered as "non-groundwater" class for the modeling process. The dataset was randomly divided into two separate sets such that 70% of samples were used for model training and the remaining samples (30%) were considered for model evaluation.

3.1.2. Groundwater explanatory variables

Selection of the variables that may directly or indirectly influence the distribution of groundwater potential is a significant step in groundwater potential modeling and significantly affects the performance of the predictive models and the quality of results. Based on the previous works reported in the literature (Choubin and Rahmati, 2021; Díaz-Alcaide and Martínez-Santos, 2019; Forootan and Seyedi, 2021; Mallick et al., 2021; Ozdemir, 2011; Rahmati and Melesse, 2016; Singh et al., 2019; Tolche, 2021), multiple field surveys and measurements, and data availability, we ended up with thirteen variables for this study: elevation, aspect, slope, topographic wetness index (TWI), sediment transport index (STI), curvature, river density, flow direction, flow accumulation, rainfall, land use, soil type, and geology (Table 1 and Fig. 3).

Table 1
Overview of the explanatory variables used in this study.

Variable	Scale	Source
Elevation (m)	30 m	USGS DEM
Slope degree	30 m	DEM
Aspect	30 m	DEM
TWI	30 m	DEM
STI	30 m	DEM
Curvature	30 m	DEM
Flow direction	30 m	DEM
Flow accumulation	30 m	DEM
River density (km/km^2)	30 m	DEM
Rainfall (mm)	30 m	VMO
Land use	1:50000	WRPI
Soil type	30 m	WRPI
Geology	1:300000	WRPI

DEM: digital elevation model that was obtained from the United States Geological Survey (https://earthexplorer.usgs.gov/), VMO: Vietnam Meteorological Organization, WRPI: Water Resources Planning and Investigation of Vietnam.

Elevation considerably affects local conditions of the landscape for groundwater distribution. Groundwater reservoirs often follow the altitude gradient and tend to accumulate under the low-elevated portions of the landscape (Ozdemir, 2011). Aspect was used as an important explanatory variable for generating groundwater potential maps because this variable is correlated with evapotranspiration and describes the direction of water flow that influences groundwater recharge and storage (Singh et al., 2019). Slope was selected as another explanatory variable because of its association with the hydrology processes that determine the runoff direction and infiltration capacity of the landscape (Magesh et al., 2012). Curvature is an indication of the amount of water that accumulates at the ground surface and its infiltration. STI reflects the capacity for sediment transfer and shows the amount of erosion and depositions that can affect the infiltration and recharge (Conforti et al., 2011; Naghibi and Moradi Dashtpagerdi, 2017). TWI quantifies the effect of topography on the hydrologic process, thereby on infiltration and recharge (Naghibi et al., 2017b). Soil types largely determine the amount of water infiltration to the ground surface to recharge the groundwater aquifers (Naghibi et al., 2017a; Zhang et al., 2019). Geology exerts a significant on surface permeability, thereby on the groundwater recharge capacity (Tolche, 2021). Rainfall is a very important variable for measuring the variability of groundwater recharge and storage, and thereby the modeling of the groundwater potential (Jenifer and Jha, 2017). River density is another influential variable for changing the capacity of groundwater recharge and storage. The areas with dense drainage networks are typically associated with a highly runoff rate, and accordingly less recharge rate and groundwater potential (Oikonomidis et al., 2015). Land-use type is an important factor considered to depict the groundwater potential. Land use (i.e., human activity) affects the potential for groundwater occurrence by changing evapotranspiration, runoff (Singh et al., 2019), recharge rates, and water demands (Lerner and Harris, 2009).

3.2. Variable selection

In machine learning modeling, feature selection allows for identifying and removing irrelevant and redundant factors to build a small set of factors that explain the dataset better than the initial set of factors. To identify the most influential factors on groundwater potential, we employed the correlation based feature selection (CBFS) method (Nguyen et al., 2020). This method utilizes Pearson's correlation approach to measure the influence of each independent factor on a response variable. Using this method, we calculated the correlation between every factor with the yield locations in the study area and ranked the factors based on their average merit (AM) for the predictive modeling of groundwater potential. We perform the CBFS method in the open-source Weka software.

3.3. Methods used

3.3.1. Naïve Bayes

Naïve Bayes (NB) is a probabilistic method inspired by Bayesian theory that utilizes numerous simple possibilities to make decisions based on the main assumption that all variables conditionally independent of one another (Soria et al., 2011). In fact, the efficiency of the NB method comes from this assumption that significantly simplifies the representation and estimation of conditional probability (Ouyang et al., 2021; Zhao and Li, 2020). Using the Bayesian classification scheme, NB can alleviate the complexity associated with the conditional probability estimation by making a conditional independence assumption that reduces the number of variables to be estimated. This assumption is made stimulation via the demand to assess the involving possibilities of the training data. In reality, most compositions of adjective/qualities amounts are neither not existent in the training data nor not existent in adequate numeral (Chen et al., 2020; Zhang et al., 2020). As a result, a straight guess of every related to poly-variable possibility will not be

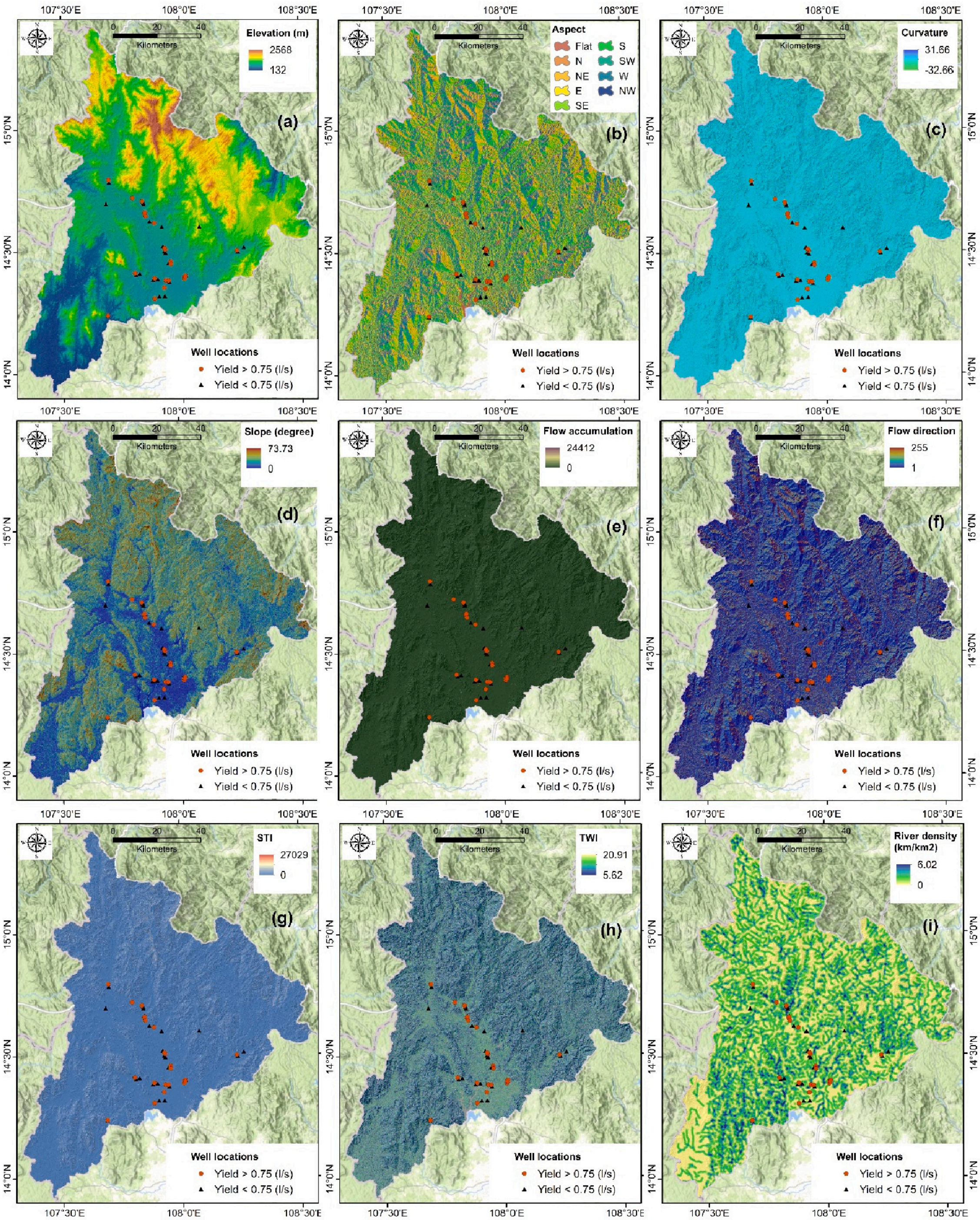


Fig. 3. Thematic maps of the explanatory variables used for groundwater potential mapping in the Kon Tum Province.

trustworthy. NB confused this trouble by its contingent independence assumption. NB can be applied to both binary and multiclass classification problems. Despite this presumption of rigid independence/liberty, the NB classifier is perfectly worthy in many real world applications due to its simplicity (Chen et al., 2020). To apply NB for groundwater potential modeling and mapping, let X = (x_1, x_2, …x_n) be a vector of independent explanatory variables. Thus, the potential for groundwater presence ($P(C_j|x_1, \ldots, x_n)$) is estimated by:

$$P(C_j \mid X) = \frac{P(C_j)P(X \mid C_j)}{P(X)} \tag{1}$$

3.3.2. AdaBoost

The boosting technique is to know several weak classifiers and incorporate them in some procedures with the aim of teaching an individual powerful classifier. AdaBoost, short for Adaptive Boosting, is an ensemble learning algorithm to build a “potent” classification as a longitudinal/linear composition and deal with dual classification (Freund and Schapire, 1997). This method can adaptively modify the weakness of classifiers during a modeling process. AdaBoost uses the same set of training samples to train different weak classifiers and then combines the classifiers to develop a strong and reliable single classifier. To do so, AdaBoost changes data distribution that leads to changing the weight of each sample in the dataset. The new weight-modified dataset is fed to the lower classifier for training. The classifiers resulted from each training course are combined together to develop the ultimate classifier. The AdaBoost ensemble technique has been acknowledged in various applications due to its ability to reduce many biases and errors of the modeling process that results in an improved predictive modeling (Kovalnogov et al., 2021; Sun et al., 2011; Wu et al., 2020).

3.3.3. Bagging

Bagging (Bootstrap aggregating) ensemble technique is a simple group learning model that provides basic models for manufacturing and association/aggregation (Breiman, 1996). Bagging has been proposed with the aim of reducing variance without increasing bias error excessively. Bagging is an effective method for modeling many environmental issues (Adhvaryu and Panchal, 2012; Yang et al., 2020). Bagging utilizes the bootstrap technique to choose the samples arbitrary with the replacement approach to generate numerous sets of samples, which also termed bootstrapped subsets. Each one of the subsets trains a base classifier individually until their outputs are merged into a single strong classifier using the majority voting approach (Medvedeva et al., 2021b).

3.3.4. Rotation forest

Compared to the AdaBoost and Bagging ensemble learning techniques, Rotation forest (RF) is a fairly recent technique proposed by Rodriguez et al. (2006). The core idea of RF is to use bootstrap sampling and principal component analysis (PCA) to generate a classifier. RF randomly divides a training dataset M into K subsets and then applies PCA with the bootstrap sampling to each subset to generate a rotation sparse matrix. Then, a classifier is developed on the features recurrently predicted by the matrix. By combining the output of the multiple classifiers, the ultimate classifier is developed. Previous works prove that RF is a favorable choice for ensemble modeling of many environmental problems (Naghibi et al., 2019; Pham et al., 2016).

3.4. Model development and validation

To develop the models, we used the open-source WEKA software. The parameter settings used in model development are summarized in Table 2.

To validate the ensemble models developed, we used positive predictive value (PPV), negative predictive value (NPV), specificity, sensitivity, accuracy, receiver operating characteristics (ROC) curve, and root mean square error (RMSE). To calculate these validation metrics, four statistical indices, i.e., true positive (TP), false positive (FP), false negative (FN), and true negative (TN), were derived from the confusion matrices that are middle results of performing the models within the WEKA software. These validation metrics are among the most widely used metrics in machine learning modeling (Hou et al., 2020; Kovalnogov et al., 2020; Medvedeva et al., 2021a; Zuo et al., 2015; Zuo et al., 2020). In our previous work (Pham et al., 2019a), we have described these metrics in the context of groundwater potential modeling and provided their mathematical formula.

Table 2
Optimal parameters of the models.

Parameter	Models			
	NB	ABNB	BNB	RFNB
Batch size	100	100	100	100
Number of decimal places	2	2	2	2
Number of iterations	–	12	10	8
Seed	–	1	1	1
Weight of threshold	–	100	–	–
Number of execution slots	–	–	1	1
Maximum of group	–	–	–	3
Minimum of group	–	–	–	3
Removed percentage	–	–	–	50
Projection filter	–	–	–	PCA

4. Results and discussion

4.1. Variable importance

Up to now, no universal guideline has been proposed for the selection of a proper set of variables that best explains the geo-environmental characteristics affecting groundwater potential. Researchers typically follow previous works to set a list of potential input variables, then decide upon the final set of variables based on the general characteristics of their study region and availability of data. With the recent advance in remote sensing and data processing techniques, researchers have inclined to use as many variables as possible, working on an assumption that the more the input variables, the higher model accuracy (Medvedeva et al., 2020). However, different variables exert different levels of impact on groundwater potential and it is very likely that some variables are irrelevant to the modeling process, highlighting the need for evaluating the variables in terms of their utility and importance for model building. The results of the CBFS method that we used for measuring the importance of the selected variables revealed that all thirteen variables more or less affect groundwater potential across the study area (Table 3). Elevation with the average merit (AM) of 0.63 was identified as the most influential variable, followed by the slope (AM = 0.358), land use (AM = 0.339), and rainfall (AM = 0.315) variables. Since all thirteen variables achieved AM $\neq$ 0, they were used in the modeling process.

An investigation into the previous works shows that the effect of different variables on groundwater potential is largely site-specific and cannot be exactly extrapolated to other regions. For instance, while groundwater potential in the Chilgazi region of Iran is significantly associated with TWI and distance from rivers (Tien Bui et al., 2019), elevation was the most influential variable on groundwater potential in the Ningtiaota region of China (Hou et al., 2018). For the DakNong Province of Vietnam, Nguyen et al. (2020) identified elevation and rainfall as the most and least important variables, respectively. In contrast, Oikonomidis et al. (2015) reported rainfall as the most influential variable on groundwater potential in Thessaly, Greece. In a national-scale groundwater potential mapping for New Zealand, Singh et al. (2019) identified lithology as the most useful variable. Different variable ranking has also been reported by Ozdemir (2011), Kordestani et al. (2019), Tolche (2021), and Mallick et al. (2021) for different regions around the world, indicating the importance of variables for groundwater potential assessment depends on the geo-environmental and topo-hydrological characteristics of the study area.

4.2. Model performance

Using various performance metrics, the single NB model and its

Table 3
Variable ranks extracted using the CBFS method.

Rank	Variable	Average merit (AM)
1	Elevation	0.63
2	Slope	0.358
3	Land use	0.339
4	Rainfall	0.315
5	STI	0.218
6	Curvature	0.169
7	Flow direction	0.146
8	River density	0.15
9	Aspect	0.138
10	Flow accumulation	0.125
11	TWI	0.122
12	Geology	0.096
13	Soil	0.038

Table 4
Training performance of the models.

Metric	NB	ABNB	BNB	RFNB
PPV (%)	48.28	62.07	58.12	58.12
NPV (%)	96.15	96.77	100	96.77
SST (%)	93.33	94.74	100	93.33
SPF (%)	66.67	73.17	100	66.67
ACC (%)	73.33	80	75.41	73.33
RMSE	0.499	0.386	0.475	0.483

derived ensemble models were validated and compared in terms of identifying the general pattern of groundwater potential in the training phase of the modeling process and predicting future groundwater occurrence in the validation phase. Over the training phase, the ensemble ABNB model attained the highest PPV (62.07%) and ACC (80%) and the lowest RMSE (0.386) (Table 4). These results can be further interpreted to mean that this model correctly classified 62.07% of groundwater pixels in the potential class and 80% of all training pixels. In terms of the NPV, SST, SPF metrics, the BNB model achieved the highest possible values indicating that this model correctly classified 100% of all pixels into the potential class, classified 100% of all pixels in the non-potential class, and classified 100% of groundwater pixels in the potential class. Overall, the three ensemble models outperformed the single NB model over the training phase of groundwater potential modeling.

To evaluate the predictive ability and generalizability of the proposed models, they were tested with the validation dataset (i.e., unseen data) over the validation phase. Based on a variety of performance metrics (Table 5), we found that the single NB model was outperformed by the three ensemble models. Compared to the ensemble models, the single NB model achieved the lowest values of PPV (60%), NPV (90%), SST (90%), SPF (70%), and ACC (76.67%) and the highest modeling error (RMSE = 0.446). Among the three ensemble models, the RFNB model gained the highest value of PPV (66.82%), NPV (100%), SST (100%), SPF (75%), and ACC (83.33%) and the lowest modeling error (RMSE = 0.406). Previous research has also reported on the asymmetric performance rates for different machine learning models when they are applied to the different training and validation datasets (Avand et al., 2020; Elzain et al., 2021; Naghibi et al., 2019; Pham et al., 2016) and attributed these differences to the dissimilar computational basis of the models. For example, Breiman (1996) reported that Bagging performs excellent if an unstable base classifier is utilized. According to Quinlan (1996), the overfitting problem is the main reason for AdaBoost's low performance.

Table 5
Validation performance of the models.

Metric	NB	ABNB	BNB	RFNB
PPV (%)	60	66	66.67	66.82
NPV (%)	90	93.33	100	100
SST (%)	90	90.91	100	100
SPF (%)	70	71.68	73.43	75
ACC (%)	76.67	80	80	83.33
RMSE	0.446	0.439	0.417	0.406

Although we used several performance metrics to measure the goodness-of-fit and predictive abilities of the models, AUC is the most reliable and trusted metric that was derived from the ROC curve and used to evaluate the overall performance of the models proposed for mapping groundwater potential in the Kun Tom Province. The AUC values obtained from the training phase (Fig. 4a) of the modeling process ranked the single NB model as the best model (AUC = 0.889), followed by the ABNB (AUC = 0.883), BNB (AUC = 0.849), RFNB (AUC = 0.801) models. In the matter of the predictive ability that was measured over the validation phase (Fig. 4b), the AUC values exhibited that the RFNB model with AUC = 0.849 provided the most accurate prediction of groundwater potential, followed by the ABNB, BNB, and NB models that reached the AUCs of 0.844, 0.815, and 0.786, respectively.

The need to apply ensemble learning techniques to avoid overfitting was particularly relevant to our study because the excellent training performance of the single NB model decreased to a relatively poor validation performance that indicates an overfitted training performance (Avand et al., 2020). While the models achieved different ranks (i.e., performance) in the training and validation phases, the ROC method, which has been used as the main performance metric (Elzain et al., 2021), revealed that the ensemble models effectively improved the predictive ability of the single NB model. Coupling the NB model with the three ensemble learning techniques, the predictive ability of the NB model enhanced by up to 7.4%. In line with our modeling outputs, previous researches achieved improved model performance using the ensemble modeling approach (Huang and Gao, 2017; Melville and Mooney, 2005).

Although recent studies have shown the effectiveness of ensemble modeling approaches, these techniques exhibited diverse levels of performance when they were applied to different problems in different regions around the world. For instance, Pham et al. (2019b) reported that the Reduced Pruning Error Tree (RPET) method combined with the Rotation Forest and Bagging performed better than its combination with the MultiBoost and Random Subspace (RSS) for landslide prediction, whereas Nhu et al. (2020) reported on an improved performance of the same model for gully erosion prediction using the RSS ensemble learning technique. This is also the case for the works that used ensemble techniques for flood susceptibility prediction (Pham et al., 2020; Pham et al., 2021; Shahabi et al., 2020). These results allow for the conclusion that local variability can significantly alter the performance of different predictive models derived from different machine learning and

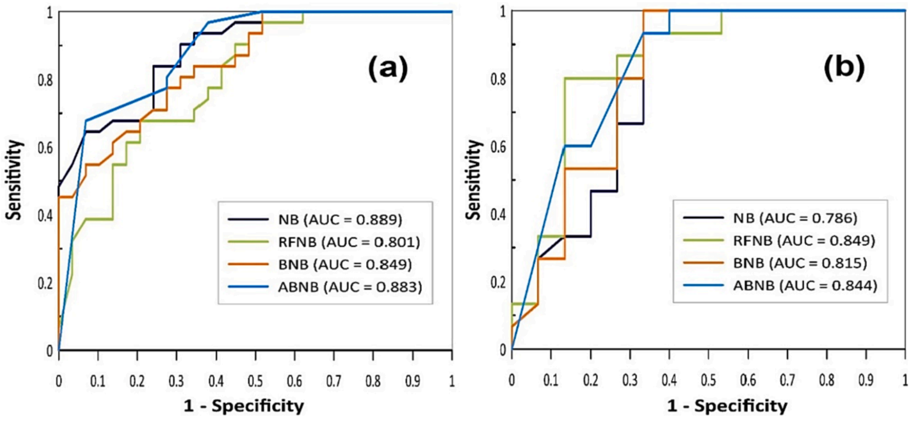


Fig. 4. ROC curve of the models over the training phase (a) and validation phase (b).

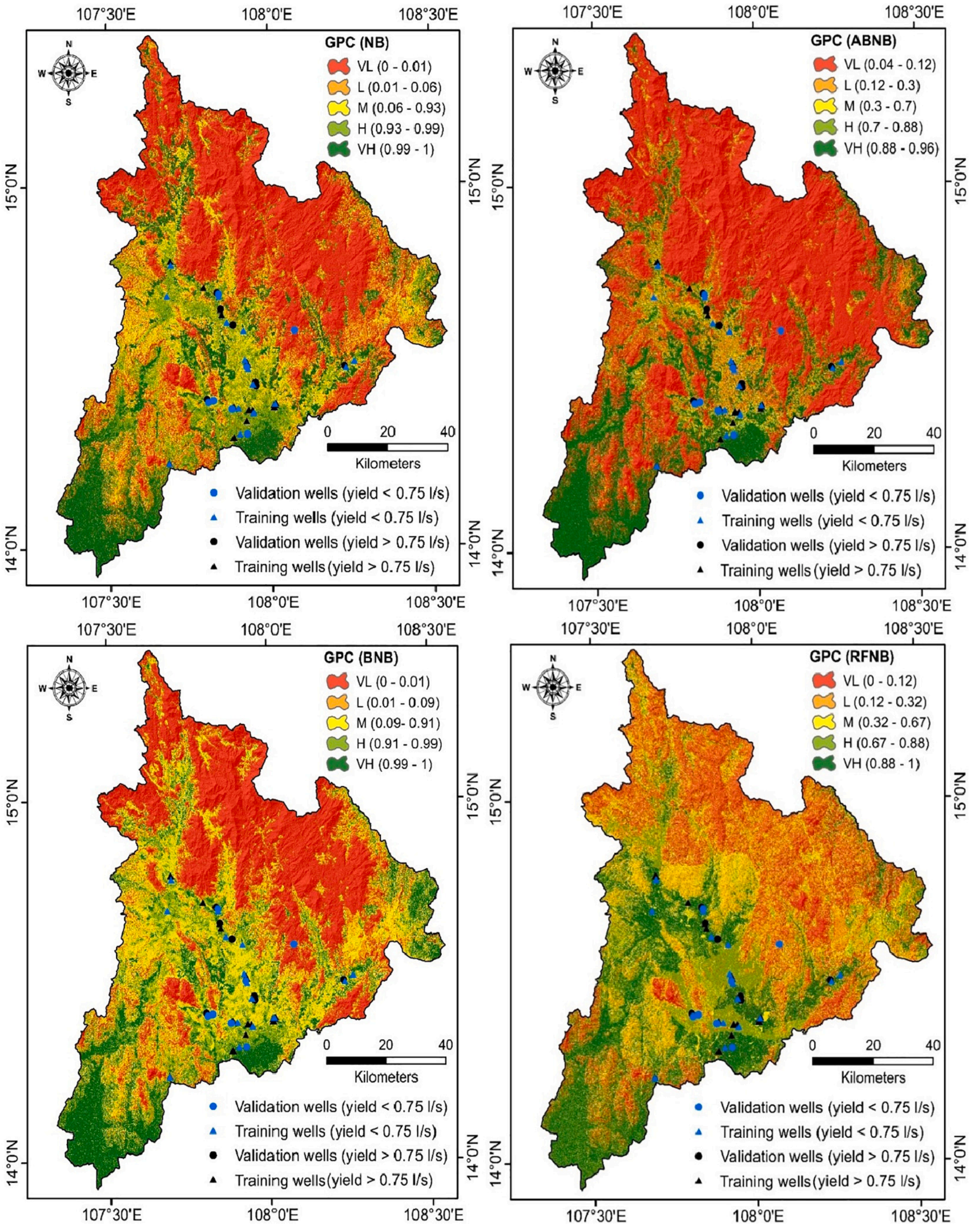


Fig. 5. Groundwater potential maps produced by the NB model and its ensembles. In each map, GPC stands for groundwater potential classes.

ensemble learning techniques. This apparently dictates more modeling effort in different regions to achieve robust predictive models.

4.3. Groundwater potential maps

Four spatially explicit maps of distribution groundwater potential across the study area were the end product of the modeling process proposed in this study (Fig. 5). In favor of a broad range of applications (e.g., land use planning, groundwater management, and water resource allocation), the distribution maps were generated to display very low, low, moderate, high, and very high potential to groundwater occurrence across the Kon Tum Province. Fig. 6 shows a detailed analysis of the produced maps. Based on this analysis, the five categories (very low to very high) covered 45.22, 7.60, 18.66, 12.54, and 15.98% of the province, respectively, on the NB model map. The ABNB model classified 58.28, 11.86, 5.57, 3.13, and 21.16% of the province into very low to very high potential categories, respectively. The BNB model classified 38.01, 6.81, 29.33, 9.94, and 15.91% of the province into very low to very high potential categories, respectively. The RFNB model classified 20.89, 29.33, 20.59, 15.98, and 21.16% of the province into very low to very high potential categories, respectively. In general, the high and very high potential categories of all four maps are predominantly associated with the low-elevated, gentle slope portions of the province that have a low density of drainage networks.

5. Concluding remarks

Groundwater aquifers need to be treated as scarce resources, with much stronger attention to managing demand. Identification of the regions with high groundwater potential is a significant step in the

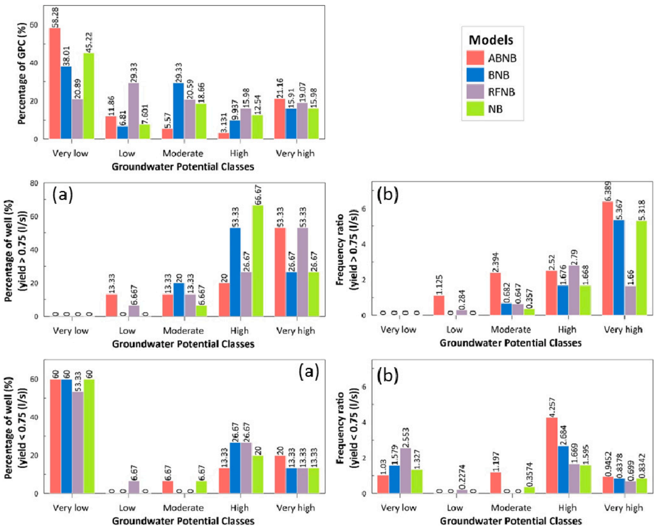


Fig. 6. Quantitative analysis of the groundwater potential maps. Training samples (a), Validation samples (b).

management of groundwater aquifers. Within the open-source Weka software, we developed an ensemble modeling framework that enabled us to analyze remotely-sensed data using an integration of the NB model with the Bagging, AdaBoost, and Rotation Forest ensemble learning technique to produce an accurate estimate of groundwater potential across the Kon Tum Province, Vietnam. The potential maps depicted the geographic locations and geoenvironmental conditions of those parts of the study area potentially suitable to groundwater presence. In addition to providing spatially explicit maps of groundwater potential for the Kon Tum Province, the main contributions of our study to literature and for water resource management include: (1) ranking thirteen geo-environmental variables in terms of their significance for groundwater potential modeling, (2) demonstrating the efficiency of ensemble modeling for estimating groundwater potential, and (3) delineating the Kon Tum Province into several classes of groundwater potential in favor of a wide range of management purposes. These contributions might be beneficial for (1) facilitating the translation of geoenvironmental information for data-scarce regions, (2) strengthening the pathways between decision-makers, stakeholders, and researchers, (3) developing strategies for compressing the high water-consumption users, (4) developing smart monitoring systems, and (5) promoting water-saving technologies for different users. Future research can include other ensemble learning techniques and Bayes rules for a fairer comparison between the capabilities of different methods.

Funding

This study received no external funding.

Declaration of Competing Interest

None.

References

Adhvaryu, P., Panchal, M., 2012. A review on diverse ensemble methods for classification. IOSR J. Comp. Eng. 1, 27–32.

Agarwal, G., Saade, S., Shahid, M., Tester, M., Sun, Y., 2019. Quantile function modeling with application to salinity tolerance analysis of plant data. BMC Plant Biol. 19, 526.

Al-Fugara, A.K., et al., 2020. Novel hybrid models combining meta-heuristic algorithms with support vector regression (SVR) for groundwater potential mapping. Geocarto Int. 1–20.

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

Becker, M.W., 2006. Potential for satellite remote sensing of ground water. Groundwater 44, 306–318.

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

Chen, S., Webb, G.I., Liu, L., Ma, X., 2020. A novel selective naïve Bayes algorithm. Knowl.-Based Syst. 192, 105361.

Choubin, B., Rahmati, O., 2021. Groundwater potential mapping using hybridization of simulated annealing and random forest. In: Water Engineering Modeling and Mathematic Tools. Elsevier, pp. 391–403.

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, 881–898.

DeSimone, L.A., Pope, J.P., Ransom, K.M., 2020. Machine-learning models to map pH and redox conditions in groundwater in a layered aquifer system, Northern Atlantic Coastal Plain, eastern USA. J. Hydrol. 30, 100697.

Díaz-Alcaide, S., Martínez-Santos, P., 2019. Advances in groundwater potential mapping. Hydrogeol. J. 27, 2307–2324.

Elzain, H.E., et al., 2021. ANFIS-MOA models for the assessment of groundwater contamination vulnerability in a nitrate contaminated area. J. Environ. Manag. 286, 112162.

Fadhillah, M.F., Lee, S., Lee, C.-W., Park, Y.-C., 2021. Application of support vector regression and Metaheuristic optimization algorithms for groundwater potential mapping in Gangneung-si, South Korea. Remote Sens. 13, 1196.

Forootan, E., Seyedi, F., 2021. GIS-based multi-criteria decision making and entropy approaches for groundwater potential zones delineation. Earth Sci. Inf. 14, 333–347.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139.

Gaur, S., Chahar, B.R., Graillot, D., 2011. Combined use of groundwater modeling and potential zone analysis for management of groundwater. Int. J. Appl. Earth Obs. Geoinf. 13, 127–139.

Guo, Y., Shen, Y., 2016. Agricultural water supply/demand changes under projected future climate change in the arid region of northwestern China. J. Hydrol. 540, 257–273.

He, L., Shao, F., Ren, L., 2021. Sustainability appraisal of desired contaminated groundwater remediation strategies: an information-entropy-based stochastic multi-criteria preference model. Environ. Dev. Sustain. 23, 1759–1779.

Hou, E., Wang, J., Chen, W., 2018. A comparative study on groundwater spring potential analysis based on statistical index, index of entropy and certainty factors models. Geocarto Int. 33, 754–769.

Hou, C.C., Simos, T.E., Famelis, I.T., 2020. Neural network solution of pantograph type differential equations. Math. Meth. Appl. Sci. 43, 3369–3374.

Huang, J., Gao, J., 2017. An ensemble simulation approach for artificial neural network: an example from chlorophyll a simulation in Lake Poyang, China. Ecol. Informat. 37, 52–58.

Jasechko, S., et al., 2017. Global aquifers dominated by fossil groundwaters but wells vulnerable to modern contamination. Nat. Geosci. 10, 425–429.

Jenifer, M.A., Jha, M.K., 2017. Comparison of analytic hierarchy process, catastrophe and entropy techniques for evaluating groundwater prospect of hard-rock aquifer systems. J. Hydrol. 548, 605–624.

Kalhor, K., Ghasemizadeh, R., Rajic, L., Alshawabkeh, A., 2019. Assessment of groundwater quality and remediation in karst aquifers: a review. Groundw. Sustain. Dev. 8, 104–121.

Khosravi, K., et al., 2018. A comparison study of DRASTIC methods with various objective methods for groundwater vulnerability assessment. Sci. Total Environ. 642, 1032–1049.

Kordestani, M.D., et al., 2019. Groundwater potential mapping using a novel data-mining ensemble model. Hydrogeol. J. 27, 211–224.

Kovalnogov, V.N., Simos, T.E., Tsitouras, C., 2020. Ninth-order, explicit, two-step methods for second-order inhomogeneous linear IVPs. Math. Meth. Appl. Sci. 43, 4918–4926.

Kovalnogov, V.N., Simos, T.E., Tsitouras, C., 2021. Runge–Kutta pairs suited for SIR-type epidemic models. Math. Meth. Appl. Sci. 44, 5210–5216.

Lee, S., Hyun, Y., Lee, S., Lee, M.-J., 2020. Groundwater potential mapping using remote sensing and GIS-based machine learning techniques. Remote Sens. 12, 1200.

Lerner, D.N., Harris, B., 2009. The relationship between land use and groundwater resources and quality. Land Use Policy 26, S265–S273.

Ma, H., Zhu, Q., Zhao, W., 2020. Soil water response to precipitation in different micro-topographies on the semi-arid loess plateau, China. J. For. Res. 31, 245–256.

MacDonald, A.M., et al., 2021. Mapping groundwater recharge in Africa from ground observations and implications for water security. Environ. Res. Lett. 16, 034012.

Magesh, N.S., Chandrasekar, N., Soundranayagam, J.P., 2012. Delineation of groundwater potential zones in Theni district, Tamil Nadu, using remote sensing, GIS and MIF techniques. Geosci. Front. 3, 189–196.

Mallick, J., et al., 2021. Proposing receiver operating characteristic-based sensitivity analysis with introducing swarm optimized ensemble learning algorithms for groundwater potentiality modelling in Asir region, Saudi Arabia. Geocarto Int. 1–28.

Medvedeva, M.A., Simos, T.E., Tsitouras, C., 2020. Variable step-size implementation of sixth-order Numerov-type methods. Math. Meth. Appl. Sci. 43, 1204–1215.

Medvedeva, M., Simos, T.E., Tsitouras, C., Katsikis, V., 2021a. Direct estimation of SIR model parameters through second-order finite differences. Math. Meth. Appl. Sci. 44, 3819–3826.

Medvedeva, M.A., Katsikis, V.N., Mourtas, S.D., Simos, T.E., 2021b. Randomized time-varying knapsack problems via binary beetle antennae search algorithm: emphasis on applications in portfolio insurance. Math. Meth. Appl. Sci. 44, 2002–2012.

Melville, P., Mooney, R.J., 2005. Creating diversity in ensembles using artificial data. Information Fusion 6, 99–111.

Motevalli, A., et al., 2019. Inverse method using boosted regression tree and k-nearest neighbor to quantify effects of point and non-point source nitrate pollution in groundwater. J. Clean. Prod. 228, 1248–1263.

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

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, 2761–2775.

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.

Naghibi, S.A., et al., 2019. Application of rotation forest with decision trees as base classifier and a novel ensemble model in spatial modeling of groundwater potential. Environ. Monit. Assess. 191.

Nguyen, P.T., et al., 2019. Development of a novel hybrid intelligence approach for landslide spatial prediction. Appl. Sci. 9, 2824.

Nguyen, P.T., et al., 2020. 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, 2473.

Nhu, V.-H., et al., 2020. GIS-based gully Erosion susceptibility mapping: a comparison of computational ensemble data mining models. Appl. Sci. 10, 2039.

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.

Ouyang, L., Zhu, S., Ye, K., Park, C., Wang, M., 2021. Robust Bayesian hierarchical modeling and inference using scale mixtures of normal distributions. IISE Transact. 1–21.

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, 290–308.

Pandey, K., Kumar, S., Malik, A., Kuriqi, A., 2020. Artificial neural network optimized with a genetic algorithm for seasonal groundwater table depth prediction in Uttar Pradesh, India. Sustainability 12, 8932.

Pham, B.T., Prakash, I., 2019. Evaluation and comparison of LogitBoost ensemble, Fisher's linear discriminant analysis, logistic regression and support vector machines methods for landslide susceptibility mapping. Geocarto Int. 34, 316–333.

Pham, B.T., Tien Bui, D., Prakash, I., Dholakia, M.B., 2016. Rotation forest fuzzy rule-based classifier ensemble for spatial prediction of landslides using GIS. Nat. Hazards 83, 97–127.

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

Pham, B.T., et al., 2019b. Landslide susceptibility modeling using reduced error pruning trees and different ensemble techniques: hybrid machine learning approaches. Catena 175, 203–218.

Pham, B.T., et al., 2020. GIS based hybrid computational approaches for flash flood susceptibility assessment. Water 12, 683.

Pham, B.T., et al., 2021. Flood risk assessment using hybrid artificial intelligence models integrated with multi-criteria decision analysis in Quang Nam Province, Vietnam. J. Hydrol. 592, 125815.

Quinlan, J.R., 1996. Bagging, boosting, and C4. 5, AAAI/IAAI, 1, pp. 725–730.

Rahmati, O., Melesse, A.M., 2016. Application of Dempster–Shafer theory, spatial analysis and remote sensing for groundwater potentiality and nitrate pollution analysis in the semi-arid region of Khuzestan, Iran. Sci. Total Environ. 568, 1110–1123.

Razavi-Termeh, S.V., Sadeghi-Niaraki, A., Choi, S.-M., 2019. Groundwater potential mapping using an integrated Ensemble of Three Bivariate Statistical Models with random Forest and logistic model tree models. Water 11, 1596.

Rodell, M., et al., 2018. Emerging trends in global freshwater availability. Nature 557, 651–659.

Rodriguez, J.J., Kuncheva, L.I., Alonso, C.J., 2006. Rotation Forest: a new classifier ensemble method. IEEE Trans. Pattern Anal. Mach. Intell. 28, 1619–1630.

Shahabi, H., et al., 2020. Flood detection and susceptibility mapping using sentinel-1 remote sensing data and a machine learning approach: hybrid intelligence of bagging ensemble based on k-nearest neighbor classifier. Remote Sens. 12, 266.

Siebert, S., et al., 2010. Groundwater use for irrigation–a global inventory. Hydrol. Earth Syst. Sci. 14, 1863–1880.

Singh, S.K., Zeddies, M., Shankar, U., Griffiths, G.A., 2019. Potential groundwater recharge zones within New Zealand. Geosci. Front. 10, 1065–1072.

Soria, D., Garibaldi, J.M., Ambrogi, F., Biganzoli, E.M., Ellis, I.O., 2011. A 'non-parametric' version of the naive Bayes classifier. Knowl.-Based Syst. 24, 775–784.

Sun, J., Jia, M.-Y., Li, H., 2011. AdaBoost ensemble for financial distress prediction: an empirical comparison with data from Chinese listed companies. Expert Syst. Appl. 38, 9305–9312.

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

Tolche, A.D., 2021. Groundwater potential mapping using geospatial techniques: a case study of Dhungeta-Ramis sub-basin, Ethiopia. Geol. Ecol. Landscapes 5, 65–80.

Wu, Y., et al., 2020. Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping. CATENA 187, 104396.

Yang, C., Gao, F., Dong, M., 2020. Energy efficiency modeling of integrated energy system in coastal areas. J. Coast. Res. 103, 995–1001.

Yen, H.P.H., et al., 2021. Locally weighted learning based hybrid intelligence models for groundwater potential mapping and modeling: a case study at Gia Lai province, Vietnam. Geosci. Front. 12, 101154.

Zhang, K., et al., 2019. Ground observation-based analysis of soil moisture spatiotemporal variability across a humid to semi-humid transitional zone in China. J. Hydrol. 574, 903–914.

Zhang, H., et al., 2020. Development of novel in silico prediction model for drug-induced ototoxicity by using naïve Bayes classifier approach. Toxicol. in Vitro 65, 104812.

Zhao, C., Li, J., 2020. Equilibrium selection under the Bayes-based strategy updating rules. Symmetry 12, 739.

Zhu, Q., Abdelkareem, M., 2021. Mapping groundwater potential zones using a knowledge-driven approach and GIS analysis. Water 13, 579.

Zuo, C., Chen, Q., Tian, L., Waller, L., Asundi, A., 2015. Transport of intensity phase retrieval and computational imaging for partially coherent fields: the phase space perspective. Opt. Lasers Eng. 71, 20–32.

Zuo, X., Dong, M., Gao, F., Tian, S., 2020. The modeling of the electric heating and cooling system of the integrated energy system in the coastal area. J. Coast. Res. 103, 1022–1029.