Vietnam Journal of Earth Sciences, 46(2), 272-281, https://doi.org/10.15625/2615-9783/20316

Vietnam Journal of Earth Sciences
https://vjs.ac.vn/index.php/jse

Groundwater potential zoning using Logistics Model Trees based novel ensemble machine learning model

Tran Xuan Bien[1], Pham The Trinh[2,3], Luu Thuy Duong[1], Tran Van Phong[4,5], Vuong Hong Nhat[6], Hiep Van Le[7], Dam Duc Nguyen[7], Indra Prakash[8], Pham Thanh Tam[9], Binh Thai Pham[7*]

[1]*Hanoi University of Natural Resources and Environment, Hanoi, Vietnam*
[2]*Tay Nguyen University, 567 Le Duan, Buon Ma Thuot, DakLak Province, Vietnam*
[3]*Department of Science and Technology, DakLak Province, Vietnam*
[4]*Institute of Geological Sciences, Vietnam Academy of Science and Technology, Hanoi, Vietnam*
[5]*Graduate University of Science and Technology, VAST, Hanoi, Vietnam*
[6]*Institute of Geography, Vietnam Academy of Science and Technology, Hanoi, Vietnam*
[7]*University of Transport Technology, 54 Trieu Khuc, Thanh Xuan, Ha Noi, Vietnam*
[8]*DDG(R) Geological Survey of India, Gandhinagar, Gujarat 382010, India*
[9]*Thai Nguyen University of Agriculture and Forestry, Thai Nguyen, Vietnam*

Received 12 December 2023; Received in revised form 12 December 2023; Accepted 11 March 2024

ABSTRACT

In this work, the main aim is to map the potential zones of groundwater in Central Highlands (Vietnam) using a novel ensemble machine learning model, namely CG-LMT, which is a combination of two advanced techniques, namely Cascade Generalization (CG) and Logistics Model Trees (LMT). For this, a total of 501 wells data and a set of twelve affecting factors were gathered and selected to generate training and testing datasets used for building and validating the model. Validation of the models was implemented utilizing various quantitative indices, including ROC curve. Results of the present study indicated that the novel ensemble model performed well for groundwater potential mapping and modeling (AUC = 0.742), and its predictive capability is even better than a single LMT model (AUC = 0.727). Thus, the CG-LMT is a promising tool for accurately predicting potential groundwater areas. In addition, the potential map of groundwater generated from the CG-LMT model is a helpful tool for better-studying water resource management in the area.

Keywords: Machine learning, groundwater potential mapping, Logistics Model Trees, cascade generalization, Vietnam.

1. Introduction

Groundwater is a significant source of freshwater in many regions, supporting agriculture, domestic consumption, and industrial uses. Given the rising water demands worldwide and the ever-decreasing availability of surface water resources, groundwater remains an essential alternative (Li et al., 2021). Identifying areas with rich groundwater potential ensures sustainable water resource management and aids in efficiently planning drilling wells and water esource infrastructure. Groundwater potential mapping is a critical aspect of hydrogeology that aids in identifying areas with potential for groundwater accumulation and extraction (Goswami and Ghosal, 2022). Groundwater potential mapping

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

involves evaluating several factors and using various tools to determine the locations most suitable for groundwater extraction.

In recent decades, machine learning (ML) has been considered a powerful tool in hydrogeology, particularly in mapping groundwater potential (Hai et al., 2022; Lee et al., 2020). Its powerfulness lies in the computational algorithms, which can learn and discover patterns effectively, draw insights, and improve predictions or decisions through experience and data analysis. Many ML-based models were potentially applied to map groundwater in various regions of the world (Mosavi et al., 2021; Prasad et al., 2020). Popular ML models used for groundwater potential mapping are Random Forests (Naghibi et al., 2017), Support Vector Machine (Lee et al., 2018), Artificial Neural Networks (Lee et al., 2012), Decision Trees (Sachdeva and Kumar, 2021), and K-Nearest Neighbors (Naghibi et al., 2018). Dey et al. (2023) applied and compared various ML models (decision tree, random forest, K-nearest neighbors, XGBoost, and support vector machine) for the potential zoning of groundwater. Anh et al. (2023) combined a Support vector machine with random search and Bayesian optimization methods to improve the effectiveness of potential groundwater mapping. Morgan et al. (2023) used random forests for zoning potential groundwater areas. In general, the ML models indicate potential tools with high accuracy for mapping groundwater potential.

In recent years, hybrid ML models have been known as more advanced for constructing better groundwater potential maps (Arabameri et al., 2021; Yariyan et al., 2022). Therefore, the main aim of the present study is to develop a novel hybrid ML model, CG-LMT, which is a hybridization of CG optimization and a single LMT classifier for improving the performance of potential groundwater mapping at the Central Highlands (Vietnam). The main difference between the present and published works is that first-time CG optimization and a single LMT classifier were combined to develop a novel CG-LMT for groundwater study in Vietnam. Various validation metrics, such as the receiver operating characteristic (ROC) curve, were utilized to validate the models. ArcGIS software was used for data preparation and generation while Weka software was used for groundwater modeling.

2. Material and methods

2.1. Data utilized

2.1.1. Description of the study area

This study focuses on Vietnam's Central Highlands area, located in the southern segment of Vietnam's central territories (Fig. 1). Covering approximately 54,700 square kilometers, the region is inhabited by nearly 4.6 million people (Bien et al., 2023). It features various elevated plateaus with elevations ranging from 500 meters to 1500 meters.

The climatic conditions of the Central Highlands are varied and can be divided into three unique sub-areas: the Northern Central Highlands, which includes Kon Tum and Gia Lai; the core Central Highlands, made up of Dak Lak and Dak Nong; and the Southern Central Highlands, consisting solely of Lam Dong Province. The core Central Highlands experiences the warmest temperatures and sits at the lowest elevations. The area generally has two primary weather seasons: a wet period from May through October and a dry spell that lasts from November to April, with March and April being notably hot and arid.

2.1.2. Groundwater wells and affecting factors

This study gathered data from 501 wells through the Vietnamese National Center for Water Resources Planning and Investigation. The yields varied among these wells, with 287 producing under 2 l/s, while 214 produced over 2 l/s (Bien et al., 2023). We divided the well records into two segments: 70% was used for constructing and training the model, and the remaining 30% served for its validation.

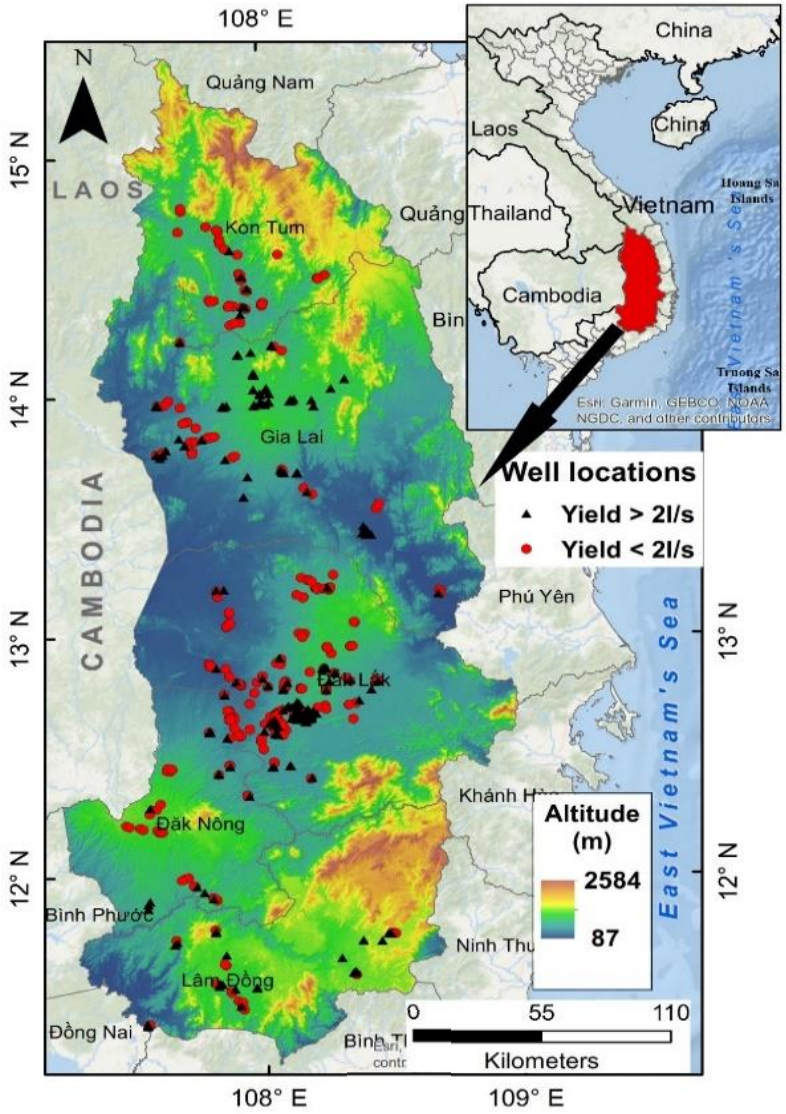

Figure 1. Location map of the study area and testing wells

In addition, a set of the affecting factors was selected for groundwater potential mapping in this study area, namely altitude, altitude difference, slope, curvature, aspect, land use/cover (LULC), flow accumulation, fault density, river density, rainfall, topographic wetness index (TWI), and geology. The main reason for selecting these affecting factors is based on the literature survey of the relevant published works (Bien et al., 2023; Ngo-Duc, 2023; Nguyen et al., 2024; Van Phong and Pham, 2023). Out of these, the factors: altitude, altitude difference, slope, curvature, aspect, river density, flow accumulation, and TWI were extracted from the SRTM Digital Elevation Model (DEM) with 90 m spatial resolution downloaded from USGS (https://earthexplorer.usgs.gov/). Geology and fault density maps were collected and generated from the General Department of Geology and Minerals of Vietnam (1:200.000). LULC map was extracted from ESA Sentinel-2 imagery at 10 m resolution. Maps of the affecting factors were constructed and shown in Fig. 2. The published work also shows data from the present study (Bien et al., 2023).

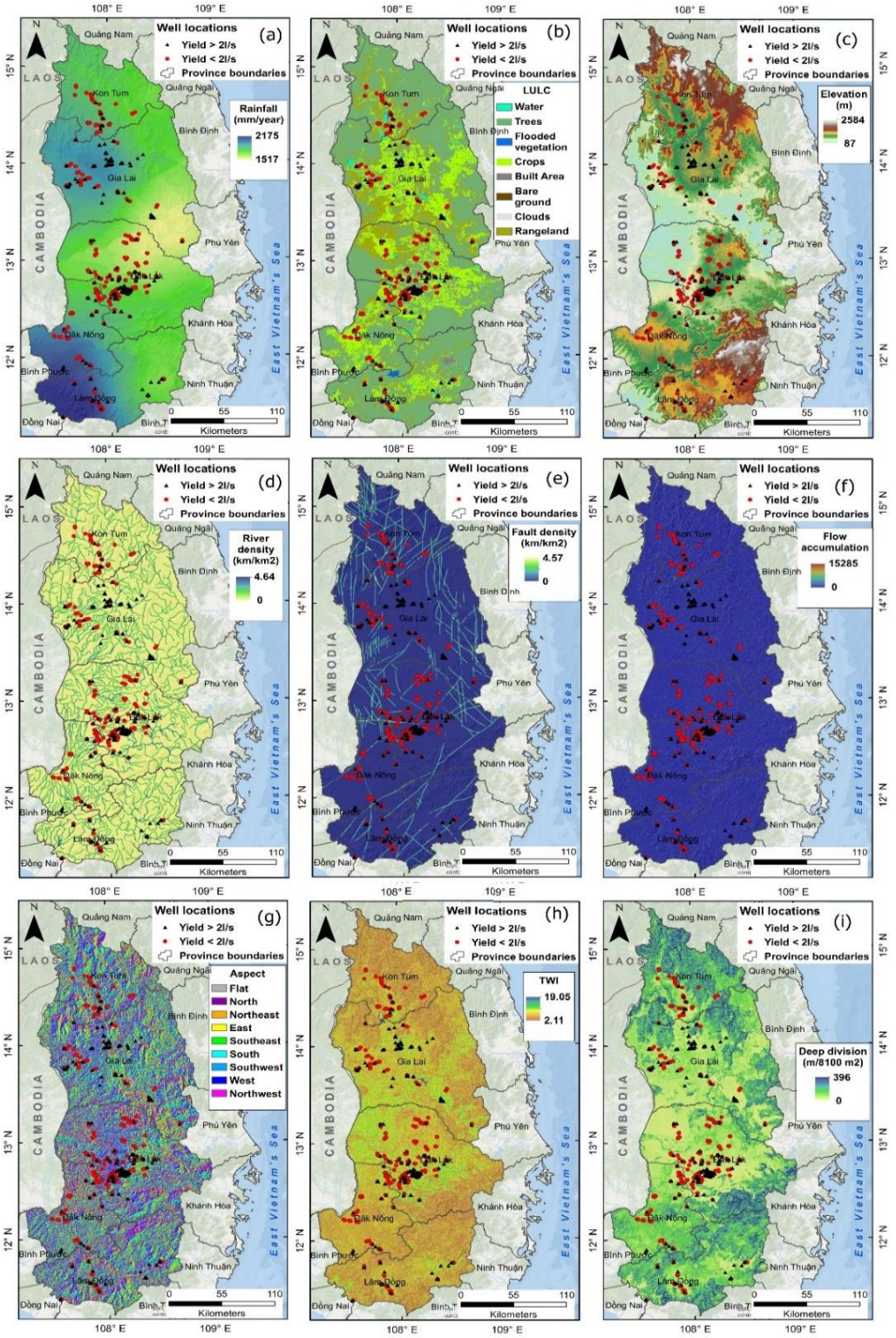

Figure 2. Maps of groundwater potential affecting factors (Bien et al., 2023)

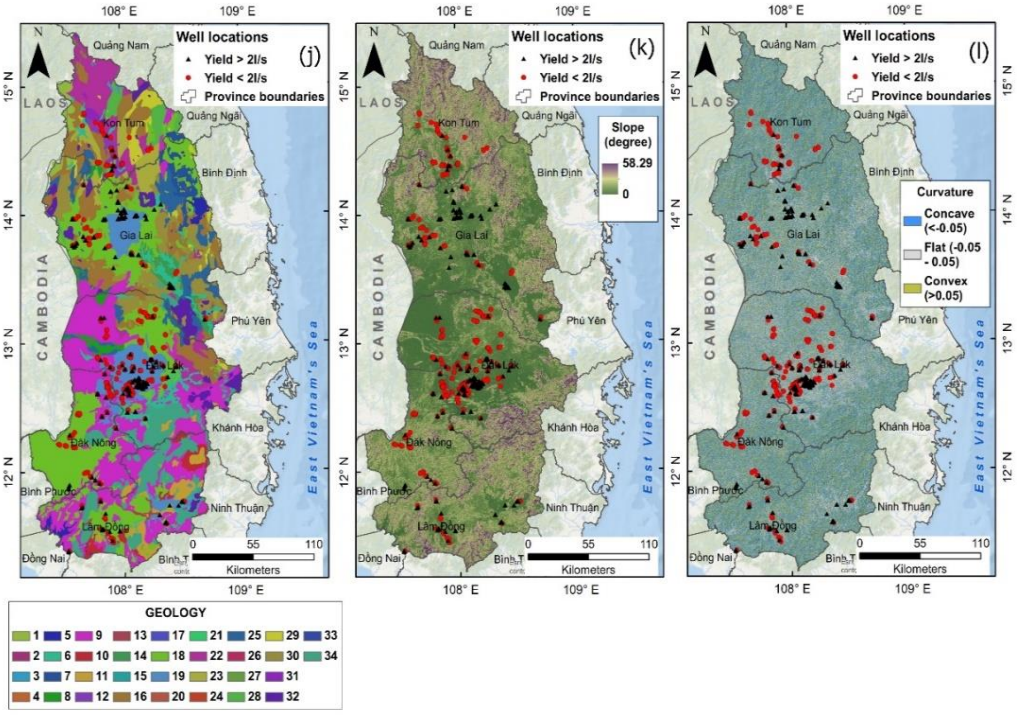

Figure 2. Cont.

2.2. Methods used

2.2.1. Logistics Model Trees (LMT)

LMT represents an innovative fusion of two powerful machine-learning techniques: decision trees and logistic regression (Landwehr et al., 2005). Each technique, in isolation, has its strengths; however, when combined, they present a unique and synergistic approach to predictive modeling. Logistic regression, at its core, is a statistical method designed to forecast the probability of a categorical outcome based on one or more predictor variables. Its strength lies in its ability to provide a continuous probability score for observations in a dataset based on a linear combination of the input features. In contrast, decision trees segment data into subsets through a series of decisions made at each node based on the values of input features. The ultimate goal of a decision tree is to make the data within each final subset as homogenous as possible regarding the outcome variable.

In the training process of LMT, the tree-building process begins with a single leaf, which represents a logistic regression model built using all available training data. As the tree evolves, it assesses potential binary splits rooted in the predictor variables. When a split substantially enhances the fit typically gauged using metrics like the likelihood-ratio tests it's accepted. This recursive segmentation continues until a predetermined stopping criterion is achieved, after which a logistic regression model is precisely fitted to the data within each of the final regions.

In this study, LMT was utilized as a classifier for predicting and assessing potential mapping and modeling of groundwater. In addition, it was also utilized as a base classifier in the ensemble framework of the hybrid model (CG-LMT), described in the following section.

2.2.2. Cascade Generalization based LMT (CG-LMT)

CG-LMT hybridizes LMT and Cascade Generalization (CG) ensemble techniques. CG is an optimization technique used to optimize the training dataset utilized for training the LMT classifier. The main principle of CG is the idea of efficiency through staged filtering or decision-making (Gama and Brazdil, 2000). Having simpler models or processes that filter out easy-to-classify instances or data points reduces the computational burden on subsequent stages. Each stage is a gatekeeper, ensuring only the most challenging or relevant instances pass through, warranting more intricate analysis or processing. In essence, CG is a strategy of hierarchical decision-making. By generalizing and making decisions at multiple levels, it's possible to balance speed and accuracy. This multi-level approach is particularly beneficial in real-world scenarios where time or computational resources are constrained, yet high accuracy is essential.

In this study, CG-LMT was used to predict and assess potential mapping and modeling of groundwater. The model was built utilizing Weka software with default hyperparameter values.

2.2.3. Validation metrics

In this study, we have used famous validation metrics, namely accuracy (ACC), negative predictive value (NPV), positive predictive value (PPV), root mean square error (RMSE), specificity (SPF), sensitivity (SST), Kappa statistic, the area under the ROC curve (AUC) for validation of the models. These metrics are detailed in the published works (Bien et al., 2023; Costache et al., 2022; Nguyen et al., 2023; Nhu et al., 2022). In general, smaller RMSE values indicate better landslide models' performance and vice versa. In contrast, more excellent ACC, NPV, PPV, SPF, SST, Kappa, and AUC values show better performance of the landslide models and vice versa (Chen et al., 2022; Doan et al., 2024; Hai et al., 2022; Kumar et al., 2021).

3. Results and discussion

3.1. Evaluation of the models

The models were evaluated on both training and testing datasets, as shown in Table 1 and Fig. 3. With the training dataset, the PPV value of CG-LMT (89%) is higher than those of the LMT model (87.5%), the NPV value of the CG-LMT model (82.31%) is higher than those of LMT (81.63%), the SST value of CG-LMT (87.25%) is higher than those of LMT (86.63%), the SPF value of CG-LMT (84.62%) is higher than those of LMT (86.63%), the ACC value of CG-LMT (86.17%) is higher than those of LMT (85.01%), the K value of CG-LMT (0.720) is higher than those of LMT (0.690), and the RMSE value of CG-LMT (0.320) is slightly lower than those of LMT (0.330). In testing dataset, similarly, the PPV value of CG-LMT (79.07%) is higher than those of LMT model (76.74%), the NPV value of CG-LMT model (57.81%) is higher than those of LMT (56.25%), the SST value of CG-LMT (71.58%) is higher than those of LMT (70.21%), the SPF value of CG-LMT (67.27%) is higher than those of LMT (64.29%), the ACC value of CG-LMT (70.00%) is higher than those of LMT (68.00%), the K value of CG-LMT (0.37) is higher than those of LMT (0.33), and the RMSE value of CG-LMT (0.47) is equal to those of LMT (0.47). Regarding AUC values, it can be observed that the AUC values of CG-LMT for training (0.92) and testing (0.742) datasets are higher than those of LMT for training (0.91) and testing (0.727) datasets.

Table 1. Values of validation metrics used to validate the models

No	Parameters	Models			
		Training dataset		Validation dataset	
		CG-LMT	LMT	CG-LMT	LMT
1	PPV (%)	89.00	87.50	79.07	76.74
2	NPV (%)	82.31	81.63	57.81	56.25
3	SST (%)	87.25	86.63	71.58	70.21
4	SPF (%)	84.62	82.76	67.27	64.29
5	ACC (%)	86.17	85.01	70.00	68.00
6	K	0.720	0.690	0.37	0.33
7	RMSE	0.320	0.330	0.47	0.47

The above validation analysis shows that the hybrid model CG-LMT has better performance than the single LMT model. It means that CG optimization techniques effectively improved the base classifier LMT performance for groundwater potential mapping and modeling. It is because CG has several advantages in enhancing the performance of the single ML models (Gama and Brazdil, 2000; Kotsiantis, 2011) such as (i) it can train sequentially the base models, which can lead to higher accuracy compared to individual models, (ii) it can handle effectively noisy or outlier data points, (iii) it helps in learning more informative feature representations as it focuses on different aspects of the data, (iv) it can adapt to the errors or challenges encountered during the training process, and (v) it can deal with the overfitting problem. Compared with the previously published work in the same area (Bien et al., 2023), it can be observed that the CG-LMT model outperforms MLP (ROC-AUC = 0.69), FURIA (AUC = 0.7), and PART (AUC = 0.72) models; however, its performance is lower than FPA (AUC = 0.76) and DFPA (AUC = 0.77) models.

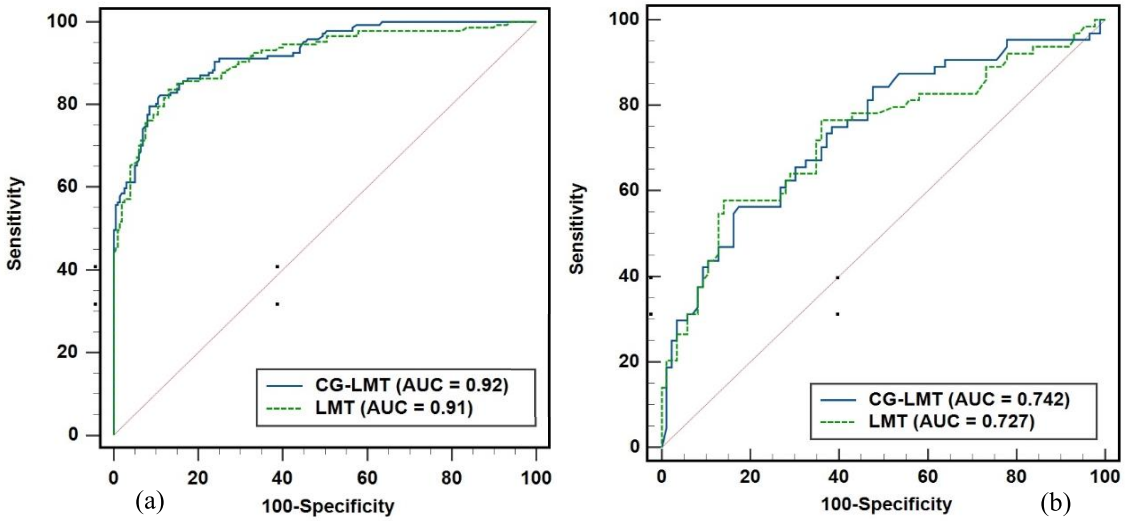

Figure 3. ROC curve analysis of the models using (a) training dataset and (b) testing dataset

3.2. Potential maps

Using two models, CG-LMT and LMT, groundwater potential maps of the study area were generated, as shown in Fig. 4. These maps were classified into five categories: high, high, moderate, low, and low potential. These categories were obtained by classifying groundwater potential indexes, which were created from training the models for all pixels of the study area. The natural breaks classification method was utilized for the classification of the maps. Table 2 shows the percentage of the groundwater potential areas of the study area. In the case of the map generated from the LMT model, it can be

observed that the shallow class obtained the highest percentage of the area (35.51%), followed by very high (23.14%), low (14.68%), high (14.45%), and moderate (12.21%) classes, respectively. With the map generated from CG-LMT, it can be seen that shallow class obtained the highest percentage of the area (37.54%), followed by very high (22.99%), low (14.50%), moderate (12.52%), and high (12.45%) classes, respectively.

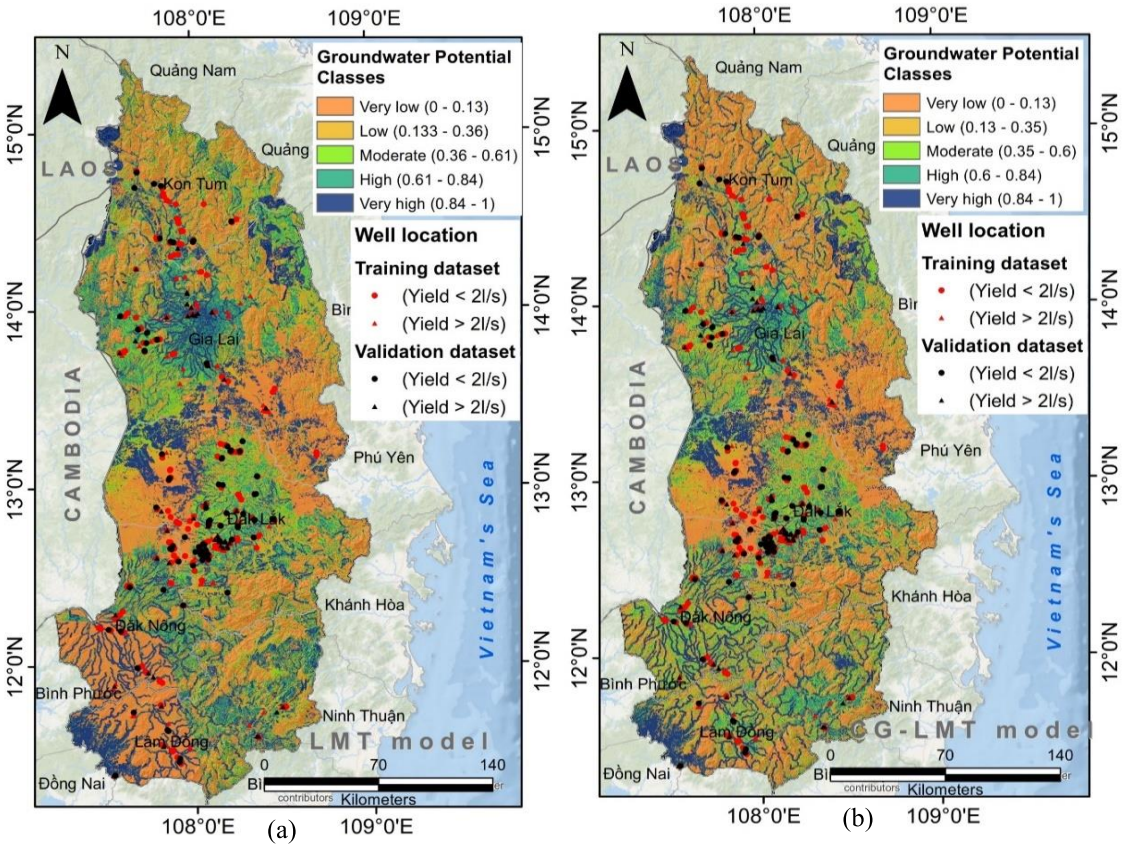

Figure 4. Groundwater potential maps using the models: (a) CG-LMT and (b) LMT

Table 2. Performance of the groundwater potential maps

Models	Potential classes	Percentage of the area
CG-LMT	Very low	35.51
	Low	14.68
	Moderate	12.21
	High	14.45
	Very high	23.14
LMT	Very low	37.54
	Low	14.50
	Moderate	12.52
	High	12.45
	Very high	22.99

In general, the potential map of groundwater generated from this study might be a helpful tool in water resource management, providing valuable insights into the availability and accessibility of groundwater in the study area (Chatterjee and Dutta, 2022). In addition, the mas informs decision-making processes, guiding sustainable management practices to ensure the responsible utilization of groundwater resources for various societal needs while safeguarding long-term environmental

integrity in the study area (Miraki et al., 2019).

4. Conclusions

In the present work, two ML models, including CG-LMT and LMT, were developed and applied for potentially mapping groundwater in the Central Highlands (Vietnam). Out of these, CG-LMT is a novel hybrid model that is a hybridization of the CG optimization technique and the base classifier LMT. Various validation metrics, including AUC, were utilized to validate and compare the models. Results of this study showed that both ML models performed well for potential modeling and mapping of groundwater, but the hybrid model CG-LMT (AUC = 0.742) outperforms the single model LMT (AUC = 0.727). Thus, it can be concluded that CG optimization is an excellent tool for optimization of the LMT for improving the performance of groundwater potential mapping. In addition, potential groundwater maps generated from this study might be utilized for better water resource management.

References

Anh D.T., Pandey M., Mishra V.N., Singh K.K., Ahmadi K., Janizadeh S., Tran T.T., Linh N.T.T., Dang N.M., 2023. Assessment of groundwater potential modeling using support vector machine optimization based on Bayesian multi-objective hyperparameter algorithm. Applied Soft Computing, 132, 109848.

Arabameri A., Pal S.C., Rezaie F., Nalivan O.A., Chowdhuri I., Saha A., Lee S., Moayedi H., 2021. Modeling groundwater potential using novel GIS-based machine-learning ensemble techniques. Journal of Hydrology: Regional Studies, 36, 100848.

Bien T.X., Jaafari A., Van Phong T., Trinh P.T., Pham B.T., 2023. Groundwater potential mapping in the Central Highlands of Vietnam using spatially explicit machine learning. Earth Science Informatics, 16, 131–146.

Chatterjee S., Dutta S., 2022. Assessment of groundwater potential zone for sustainable water resource management in south-western part of Birbhum District, West Bengal. Applied Water Science, 12, 40.

Chen Y., Chen W., Chandra Pal S., Saha A., Chowdhuri I., Adeli B., Janizadeh S., Dineva A.A., Wang X., Mosavi A., 2022. Evaluation efficiency of hybrid deep learning algorithms with neural network decision tree and boosting methods for predicting groundwater potential. Geocarto International, 37, 5564–5584.

Costache R., Ali S.A., Parvin F., Pham Q.B., Arabameri A., Nguyen H., Crăciun A., Anh D.T., 2022. Detection of areas prone to flood-induced landslides risk using certainty factor and its hybridization with FAHP, XGBoost and deep learning neural network. Geocarto International, 37, 7303–7338.

Dey B., Abir K.A.M., Ahmed R., Salam M.A., Redowan M., Miah M.D., Iqbal M.A., 2023. Monitoring groundwater potential dynamics of north-eastern Bengal Basin in Bangladesh using AHP-Machine learning approaches. Ecological Indicators, 154, 110886.

Doan V.L., Nguyen C.C., Nguyen C.T., 2024. Effect of time-variant rainfall on landslide susceptibility: A case study in Quang Ngai Province, Vietnam. Vietnam Journal of Earth Sciences. https://doi.org/10.15625/2615-9783/20065.

Gama J., Brazdil P., 2000. Cascade generalization. Machine Learning, 41, 315–343.

Goswami T., Ghosal S., 2022. Understanding the suitability of two MCDM techniques in mapping the groundwater potential zones of semi-arid Bankura District in eastern India. Groundwater for Sustainable Development, 17, 100727.

Hai H.D., Ngo H.T.T., Van P.T., Duc D.N., Avand M., Huu D.N., Amiri M., Van Le H., Prakash I., Thai P.B., 2022. Development and application of hybrid artificial intelligence models for groundwater potential mapping and assessment. Vietnam J. Earth Sci., 44(3), 410–429. https://doi.org/10.15625/2615-9783/17240.

Kotsiantis S.B., 2011. Cascade generalization with reweighting data for handling imbalanced problems. The Computer Journal, 54, 1547–1559.

Kumar R., Dwivedi S.B., Gaur S., 2021. A comparative study of machine learning and Fuzzy-AHP technique to groundwater potential mapping in the data-scarce region. Computers & Geosciences, 155, 104855.

Landwehr N., Hall M., Frank E., 2005. Logistic model trees. Machine Learning, 59, 161-205.

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

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

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

Li P., Karunanidhi D., Subramani T., Srinivasamoorthy K., 2021. Sources and consequences of groundwater contamination. Archives of Environmental Contamination and Toxicology, 80, 1–10.

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 Resources Management, 33, 281–302.

Morgan H., Madani A., Hussien H.M., Nassar T., 2023. Using an ensemble machine learning model to delineate groundwater potential zones in desert fringes of East Esna-Idfu area, Nile valley, Upper Egypt. Geoscience Letters, 10, 9.

Mosavi A., Sajedi Hosseini F., Choubin B., Goodarzi M., Dineva A.A., Rafiei Sardooi E., 2021. Ensemble boosting and bagging based machine learning models for groundwater potential prediction. Water Resources Management, 35, 23–37.

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

Naghibi S.A., Pourghasemi H.R., Abbaspour K., 2018. A comparison between ten advanced and soft computing models for groundwater qanat potential assessment in Iran using R and GIS. Theoretical and Applied Climatology, 131, 967-984.

Nguyen H.D., Nguyen V.H., Du Q.V.V., Nguyen C.T., Dang D.K., Truong Q.H., Dang N.B.T., Tran Q.T., Nguyen Q.-H., Bui Q.-T., 2024. Application of hybrid model-based machine learning for groundwater potential prediction in the north central of Vietnam. Earth Science Informatics, 1–21.

Nhu V.-H., Bui T.T., My L.N., Vuong H., Duc H.N., 2022. A new approach based on integration of random subspace and C4. 5 decision tree learning method for spatial prediction of shallow landslides. Vietnam J. Earth Sci., 44(3), 327–342. https://doi.org/10.15625/2615-9783/16929.

Prasad P., Loveson V.J., Kotha M., Yadav R., 2020. Application of machine learning techniques in groundwater potential mapping along the west coast of India. GIScience & Remote Sensing, 57, 735–752.

Sachdeva S., Kumar B., 2021. Comparison of gradient boosted decision trees and random forest for groundwater potential mapping in Dholpur (Rajasthan), India. Stochastic Environmental Research and Risk Assessment, 35, 287–306.

Van Phong T., Pham B.T., 2023. Performance of Naïve Bayes Tree with ensemble learner techniques for groundwater potential mapping. Physics and Chemistry of the Earth, Parts A/B/C, 132, 103503.

Yariyan P., Avand M., Omidvar E., Pham Q.B., Linh N.T.T., Tiefenbacher J.P., 2022. Optimization of statistical and machine learning hybrid models for groundwater potential mapping. Geocarto International, 37, 3877–3911.