Spatial Mapping of Groundwater Potential (GWP) Identification Using Random Forest Machine Learning in Kedah, Malaysia

Zabidi, S. R. S.,[1] Bohari, S. N.,[1*] Narashid, R. H.,[1] Saian, R.,[2] Ismail, M. A. M.,[3] Nasron, N.,[1] Latif, Z. A.[4] and Pa'suya, M. F.[1]

[1]Faculty of Built Environment, Surveying Science and Geomatic Studies, Universiti Teknologi MARA, Perlis Branch, Arau Campus, 02600 Arau, Perlis, Malaysia, E-mail: syarifahraihana10@gmail.com, ashikin10@uitm.edu.my,* rohayuharon@uitm.edu.my, nursy6864@uitm.edu.my, faiz524@uitm.edu.my

[2]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perlis Branch, Arau Campus, 02600 Arau, Perlis, Malaysia, E-mail: rizauddin@uitm.edu.my

[3]School of Civil Engineering, Engineering Campus, Universiti Sains Malaysia, 14300 Nibong Tebal, Penang, Malaysia, E-mail: ceashraf@usm.my

[4]Faculty of Built Environment, Surveying Science and Geomatic Studies, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia, E-mail: zulki721@uitm.edu.my

**Corresponding Author*

DOI: https://doi.org/10.52939/ijg.v22i3.4865

Abstract

Groundwater is a critical source of freshwater in Malaysia, where rapid urban growth, agricultural expansion, and recurrent shortages of surface water have increased relianced on aquifers. Effective groundwater potential (GWP) mapping is therefore essential to identify suitable zones for sustainable extraction and water security planning. This study aims to determine the GWP area in Kedah by using random forest (RF) machine learning techniques involving 15 groundwater conditioning parameters covering topography, hydrogeology, and environmental factors. A total of 350 tube well locations was partitioned into 70:30 for training and testing dataset. The identified GWP area was classified into five different classes: very high, high, medium, low, and very low. It was found about 30.97% covering an area of 2,798.18 km^2 is considered as the highest groundwater area in the western and central part of Kedah. Meanwhile, the lowest groundwater area which is about 25.25% is in the northeast part of Kedah covering an area of 2,281.18 km^2. The performance of the RF model was validated using several evaluation metrics for both training and testing dataset: accuracy, precision, sensitivity, specificity, F1-score and kappa. The validation using receiver operating characteristics (ROC) demonstrated strong discriminative ability with the area under curve (AUC) values of 0.95 (training) and 0.90 (testing). Feature importance analysis revealed elevation was found to be the highest influencing contributors, while lithology was found to be the least influencing contributors for the model's performances. Overall, the findings highlight the effectiveness of integrating geospatial and machine learning techniques in GWP studies that provides a robust framework which contributes significantly towards sustainable groundwater management strategies.

Keywords: Groundwater Potential Identification, Kedah, Machine Learning, Random Forest

1. Introduction

Groundwater is the water that is found below the Earth's surface, seeping through the cracks and spaces in soil, sand and rocks in a different layer of aquifer and has become important to support the daily needs of people, animals and plants. Statistically, according to UNESCO (2022), there are 69% of groundwater used for agriculture, 22% for domestic, 9% for industrial and nearly 50% of groundwater has been used by the global urban population. This proved groundwater plays one of the most important roles in the economy as it is the primary source of water worldwide. In Malaysia, several states such as Kelantan, Perlis and Negeri Sembilan have utilized groundwater to combat the issues of the water shortage due to the rapid growth of economy, urbanization, land development and climate changes [1] and [2]. As practice in Kelantan, 70% of their freshwater sources comes from

groundwater resources. Meanwhile, at the certain area, situation became severe in small isolated islands like Kapas and Manukan, which depends solely on groundwater resources [2].

According to [3] and [4], agricultural sectors is one of the largest consumer of water in Malaysia, accounting for about 76% of the total water supply. Kedah, which is known as the rice bowl of Malaysia have experienced water stress due to the extensive paddy cultivation [5]. This situation is further worsened especially during the wet and dry seasons, where rising temperatures and fluctuating rainfall threatens the crop stability and reduce water availability accross the region [5] and [6]. Additionally, the increasing population density continues to put pressures for water resources between domestic and agriculture uses [4]. Therefore, an alternative solution is required to address these challenges effectively. The utilization of groundwater as supplementary resources will strengthen and enhance the water security in Kedah. This approach aligns with the targets of SDG 6 which aims to ensure the sustainable access to clean water for all.

Traditionally, the estimation of groundwater potential (GWP) area has been approached by using the drilling and strata analysis [7][8] and [9]. The drilling method works by drilling a hole through the aquifer and pumping the water to be brought to the surface. This method is time-consuming as it does not cover a large area and is highly cost because it requires a high skilled labour and high costed instrument to be used [8][9] and [10]. Then, the technology evolved to be more sophisticated techniques, with the utilizing of Geographic Information System (GIS) and Remote Sensing (RS) techniques such as analytical hierarchy process (AHP), weight of evidence (WoE) and logistic regression (LR) offers less uncertainty in identifying the important aspect and being effective in the decision-making process [7][8][11] and [12]. Nowadays, the new method has been employed by using various machine learning algorithm for GWP mapping, including random forest (RF), support vector machine (SVM), boosted regression tree (BRT), k-nearest neighbourhood (KNN) and artificial neural network (ANN) [12] and [13]. These methods delivered higher prediction rate, higher accuracy and consistent reliability in processing complex data compared to the other methods [11][12][14] and [15].

Furthermore, limited studies especially for the advanced data-driven approaches used to predict GWP in this area compared to the other states such as Kelantan, Selangor and Sabah which comparatively few evaluations addressing to the topographic and geology characteristics in the northern areas. Thus, this study aims to determine GWP area in Kedah by using RF machine learning methods that give a high prediction rate and good accuracy compared to conventional methods. The integration of 15 groundwater conditioning parameters from various factors provides a more complete and data-rich representation of the factors influencing groundwater occurrence. The modeling framework is enhanced by using several evaluation metrics and receiver operating characteristcis (ROC) area under curve (AUC) to measure and enhance the accuracy and reliability of the predictive performances. Lastly, the final GWP map is classified into 5 different classes of very high, high, medium, low and very low which to ensure authorities and planners making a right decision to explore, protect and sustain groundwater resources.

2. Study Area and Datasets

2.1 Study Area

The selection of study area is in Kedah, located at the northern part of the west coast of Malaysia (Figure 1). It consists of 12 administrative districts such as Kubang Pasu, Sik, Yan, Kota Setar, Penang, Kuala Muda, Baling, Kulim and Bandar Baharu. Kedah which position at 6˚ 9′ 20.4192″ N and 100˚ 34′ 10.7364″ E covers around 9500 sq km with a population of slightly over 2 million people. This area exhibits diverse topographical characteristics that significantly impacts the groundwater occurrence. The western and central part of the region are dominated by extensive alluvial plains with the elevation typically between -171 m to 107 m. These areas represent the agricultural zone of Kedah which is particularly for paddy cultivation. On the other hand, majority of the eastern part represents mountainous and hilly areas that ranges between 500 m to 1,856 m. The characteristics of these areas act as major groundwater recharge that contribute significantly towards groundwater movement. Climatically, Kedah experiences tropical monsoon climate and receives an average annual rainfall between 2,000 mm to 2,500 mm with the temperature of 27˚ to 32˚ C. The increasing demand of water gives significant pressures on agricultural activities which is one of the major contributors of water consumption in Kedah. The combination of several topography and hydrogeology characteristics in Kedah have significantly influenced the distribution, storage and movement of groundwater, necessity a detailed groundwater potential assessment to support water management planning and enhance long-term water security.

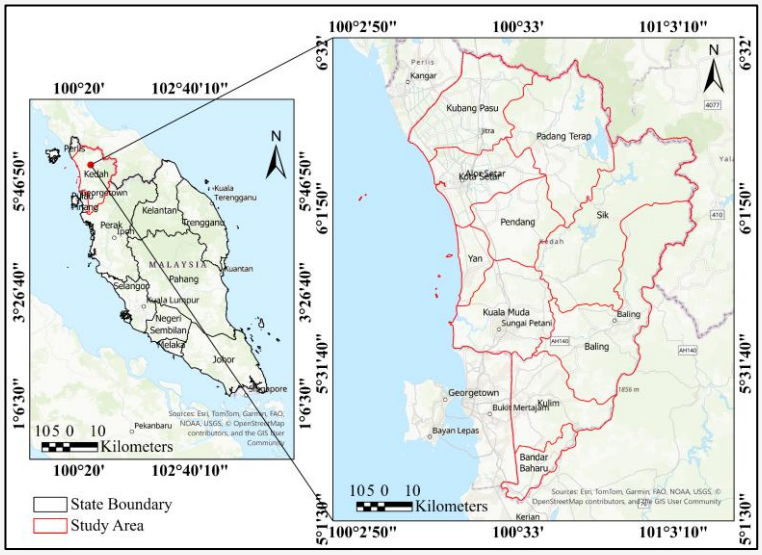

Figure 1: Kedah, the northern part of the west coast of Malaysia

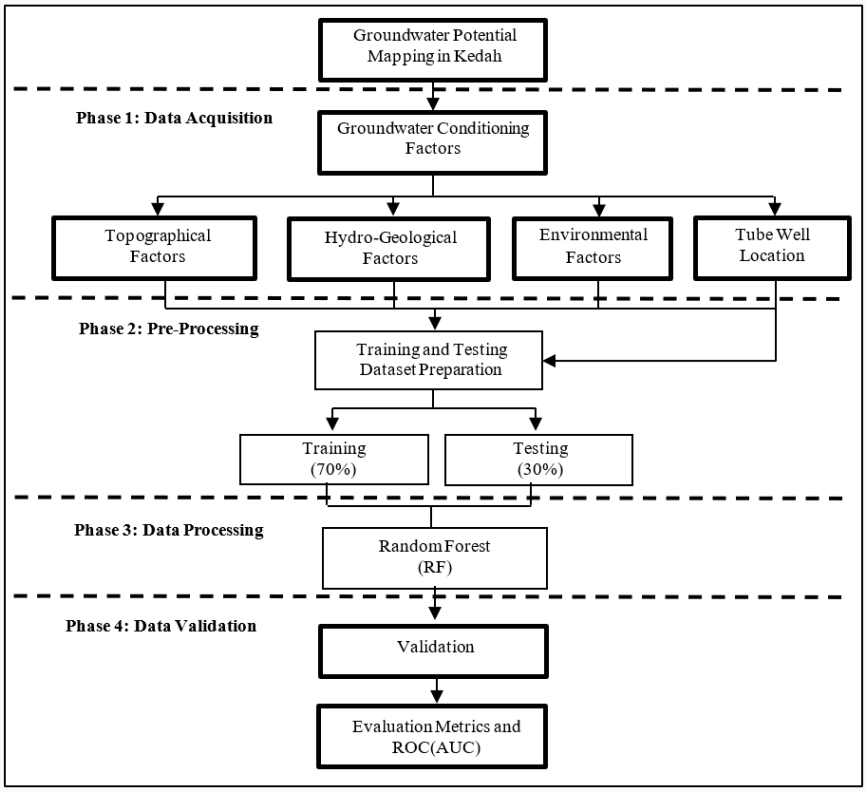

Figure 2: Groundwater potential identification

3. Material and Methods

The methodology applied in the study is presented in Figure 2, which started with the data acquisition by selecting the GWP parameters based on different factors of topography, hydrogeology and environmental factors. Then, the flow continued with the pre-processing which is the preparation of training and testing dataset. The dataset were divided into 70% for training and 30% for testing. The data processing continued with the prediction of GWP using RF Python software. Lastly, the validation of GWP result were validated by using several evaluation metrics.

3.1 Training and Testing Dataset

Groundwater inventory is essential in GWP studies as it provides reliable information based on the actual data of groundwater well. In this study, groundwater tube well points were classified into binary classes, whereas value of 1 denotes as the presence of groundwater, while value of 0 denotes as the absence of groundwater. A total of 350 groundwater tube well points were obtained from Department of Mineral and Geoscience (DMG) Malaysia. Another 350 non-groundwater tube well points were generated through random sampling in ArcGIS Pro. The full dataset was then partitioned into 70% for training and 30% for testing. Figure 3 shows the map of groundwater tube well points and non-groundwater tube well points.

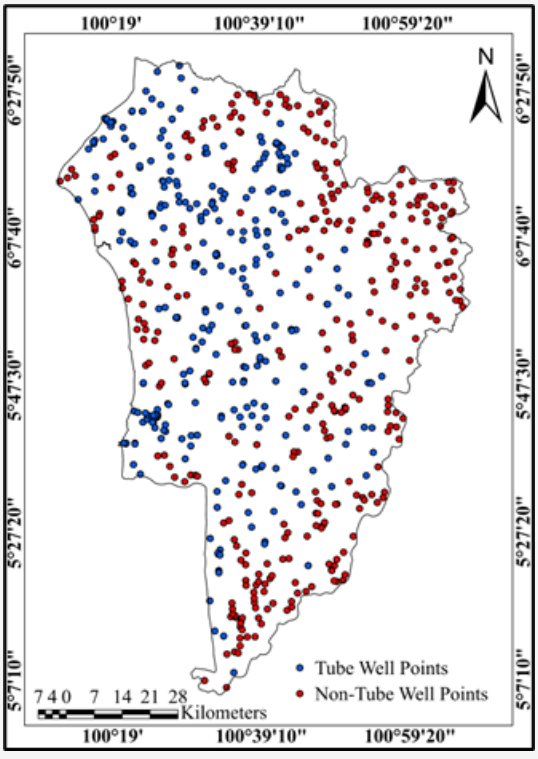

Figure 3: Tube well and non-tube well points

3.2 Groundwater Conditioning Parameters

Groundwater conditioning parameters are important in determining the GWP mapping. In this study, 15 parameters were divided into three categories, which are topographical (elevation, slope, aspect, topographical wetness index (TWI), plan curvature, profile curvature and geomorphology), hydro-geological (drainage density, lineament density, lithology, distance to fault and aquifer) and environmental factors (rainfall, land use and soil types). Table 1 shows the data description used in this study. TanDEM-X with 12m resolution were provided by the German Aerospace Center (DLR) upon a special request and are not publicly available. The digital elevation model (DEM) data has been used to produce topographical factors such as elevation, slope, aspect, TWI, plan curvature, geomorphology, and drainage density. For the hydrogeological factors, lithology, fault and aquifer, have been obtained and from DMG. Meanwhile, rainfall and soil types data has been obtained from Malaysian Meteorological Department (METMalaysia), Department of Drainage and Irrigation (DID) and Department of Agriculture (DOA). Lastly, for land use, the data has been derived from open-source data such as Copernicus Open Access Hub for Sentinel 2A satellite imagery.

3.2.1 Elevation

Elevation is considered as one of the important parametersin determination of groundwater. This is due to its characteristics, which higher elevations indicate lower groundwater potential, while lower elevations suggest a higher potential of groundwater [8][16] and [17]. Figure 4(a) shows the elevation map of Kedah that ranges between -170.99 m to 1856 m.

3.2.2 Slope

The gentle slope indicates a higher infiltration capacity and potential recharge area which results in higher water capacity [13][18] and [19]. Meanwhile, a steeper slope indicates a low infiltration capacity that limits the water capacity [20] and [21]. Slope map as shown in Figure 4(b) classified into five classes that ranges between 0.01 to 80.3 degrees.

3.2.3 Aspect

Aspect is one of the important factors for groundwater potential area because it determines the direction of the slope [8][12][18] and [22]. Figure 4(c) shows the aspect map and classified the data into 10 classes of direction.

3.2.4 Topographical Wetness Index (TWI)

The TWI map was extracted using Equation 1 [23].

$$TWI = \ln\left(\frac{fa}{\tan\beta}\right) \quad \text{Equation 1}$$

Where:

TWI = Topographical Wetness Index
fa = Flow Accumulation
β = Slope Angle at Point

TWI describes the influence of topography on the location of saturated source zones which contribute to groundwater occurrence [12] and [24]. Figure 4(d) shows the TWI map in Kedah that ranges between –10.68 to 14.89.

Table 1: Data description of groundwater conditioning parameters

Category	Factors	Source of Data	Data Description	Data Resolution/Scale
Topography	Elevation Slope Aspect TWI Plan Curvature Profile Curvature Geomorphology	TanDEM-X (DLR)	DEM	12 m × 12 m
Hydro-geology	Drainage Density	TanDEM-X (DLR)	DEM	12 m × 12 m
	Lineament Density	-	Lineament lines	12 m × 12 m
	Lithology	DMG	Lithological units	1:50,000
	Distance to fault	DMG	Major and minor faults	12 m × 12 m
	Aquifer	DMG	-	1:50,000
Environment	Rainfall	METMalaysia & DID	Average rainfall	-
	Land Use	Copernicus Open Access Hub (Link: https://browser.dataspace.copernicus.eu/)	Sentinel 2A	10 m × 10 m
	Soil Types	DOA	Major soil group	1: 25,000

3.2.4 Plan curvature

Curvature tools are used to derive a plan curvature map. The negative values on plan curvature indicate convergence water flow over the surface. Meanwhile, the positive values indicate divergence in water flow over the surface [12] and [25]. Curvature map as depicted in Figure 4(e) has been classified into three classes of concave (-35.84 – -0.63), linear (-0.62 – 0.48) and convex (0.49 – 34.88).

3.2.5 Profile curvature

Profile curvature represents the acceleration and decelaration of the water flow [24] and [25]. In contrast with the plan curvature, the positive value on profile curvature indicates the surface opening is concave, while the negative value represents the surface is convex [26] and [27]. Figure 4(f) illustrates the profile curvature map that has been classified into three classes: convex (-32.9 – -0.48), linear (-0.47 – 0.7) and concave (0.71 – 42.27).

3.2.6 Geomorphology

Geomorphology influences groundwater by analyzing the structure of the area, whereas the fluvial structure has more permeability rather than the hard rock structure [8] and [18]. The geomorphology map in Figure 5(a) was produced by using slope tools derived from a hill shade map that ranges between 0.01 to 85.08 degree

3.2.7 Drainage Density

Drainage density map has been produced in an ArcGIS environment involving several steps that include hydrology and the conditional tool in the Spatial Analyst Tool. The value of drainage density indicates that a low value of drainage density represents small flow and large infiltration of water, and a high value of drainage density represents large flow and poor infiltration of water [20] and [28]. Figure 5(b) shows the drainage density map in Kedah that has been classified into five classes ranges between 0.05 to 3.74.

3.2.8 Lineament Density

Figure 5(c) illsutrates the lineament density map that ranges between 0 to 1.18. Lineament density map has been produced from the lineament lines on the hill shade surface of DEM in ArcGIS Pro. Lineament increase the permeability of the subsurface to allow water flowing through the faults and fractures [29] and [30].

3.2.9 Lithology

Lithology influence groundwater flow as it affects the permeability of aquifers and the soil porosity [13][21][31] and [32]. The lithology map in Figure 5(d) has been classified into several classes of lithology age such as Cambrian, Carboniferous, Igneous Activities: Triassic, Jurassic-Cretaceous, Ordovician-Silurian, Quaternary, Silurian-Devonian, Tertiary and Triassic. Quaternary deposit has the highest impact on groundwater potential due to the high porosity and permeability, which allow groundwater storage and movement. In contrast, Cambrian and Igneous formations generally have the least groundwater availability.

3.2.10 Aquifer

Aquifer map (Figure 5(e)) has been extracted from the hydro-geological map using georeferencing and

digitizing methods for the aquifer potential area that classified into four different classes, which are very high, high, medium and low.

3.2.11 Distance to fault

Rock formations near fault areas are highly fractured which allowed susbstantial amount of water to infiltrate into the ground [17] and [19]. A shorter distance to fault correlates with a higher potential of water infiltration which suggests as a high probability of groundwater occurrence [17]. The process of producing distance to fault map was by using Euclidean distance in the spatial analyst tools. The input of the data used was the major and minor faulting data that derived from DMG. Then, the map was classified into five different classes as shown in Figure 5(f).

3.2.12 Land Use

Land Use has been generated by using Sentinel 2A satellite images derived from the Copernicus Data Space Ecosystem. The classification process started with assigned image training classification tools (Supervised classification) in ArcGIS software. A total of 200 ground truth points has been validated against the land use classification. The overall accuracy achieved values of 95% and kappa of 0.91, which demonstrated as a strong agreement between predicted and actual classes. Figure 5(g) illsutartes the land use map that has been classified into five classes: barren land, water, agriculture, forest, and urban land.

3.2.13 Rainfall

Rainfall is important in determining GWP area. This is due to the high rainfall affects more infiltration into the ground, resulted high potential of groundwater area [20]. Meanwhile, less rainfall affects the low infiltration into the ground, resulted low groundwater occurrence [25]. In this study, there were total of 36 rainfall stations that were obtained from METMalaysia and DID, which consists of the monthly rainfall amount, coordinates, and elevation from 2013 until 2023. The cross validation result demonstrated that kriging was selected as the interpolation method as it performs better than IDW between predicted and measured rainfall. Figure 5(h) illustrates the rainfall map that ranges between 142.11 mm to 353.09 mm.

3.2.14 Soil

The soil map (Figure 5(i)) in this study was derived from the Department of Agriculture Malaysia that consists of major soil group data. There were six classes of soil map, which are Marine Alluvial Soils, Mined Lands, Reworked Soils, Riverine Alluvial Soils, Sedentary Soils, Urban Lands and also Water Bodies.

3.3 Machine Learning Method

3.3.1 Random forest (RF)

RF is one of the powerful trees learning in machine learning techniques that has been developed by Breiman [33]. RF can be used for both classification and regression problems [33][34] and [35]. The classification task predicts the final output by using majority votes from each decision tree, while the regression task averages the predictions made by the trees [36]. This algorithm performs by creating a number of multiple decision trees and combining their output to model the spatial relationship between dependent variables (tube well points) and independent variables (GWP conditioning parameters) [18] and [37]. Consequently, RF is one of the most commonly applied ML techniques especially in GWP studies. Compared to other widely used method such as SVM and KNN, RF has been achieved higher predictive performance based on the AUC ranges presented in Table 2 (AUC > 0.9) in prior studies [26][38] and [39].

Table 2: AUC classification

AUC Ranges	Description
0.9 – 1.0	Excellent
0.8 – 0.9	Good
0.7 – 0.8	Fair
0.6 – 0.7	Poor
0.5 – 0.6	Fail

Table 3: RF hyperparameters

Hyperparameters	Ranges
n_estimators	200, 300
max_depth	3, 5, 7
min_samples_split	15, 20, 30
min_samples_leaf	10, 15, 20
max_features	0.2, 0.3, 0.4
bootstrap	True
max_samples	0.8, 0.9

RF is particularly advantageous in giving high performance, handling complex dataset and avoid the issues of overfitting [11][12] and [15]. In this study, RF was implemented using RandomForestClassifier from the scikit-learn (sklearn) library in Python. The optimization of the model performances was made by using RandomizedSearchCV with a 10-fold stratified cross validation to enhance the model performance and prevent overfitting. Table 3 shows the hyperparameters ranges during the optimization process.

The feature importance was evaluated using Gini Importance to measure the variable's contribution based on its total reduction in Gini impurity. This method allows to identify the most influential factors among GWP conditioning parameters. Then, feature selection was employed using SelectFromModel in scikitlearn to remove least importance variables that could reduce the model accuracy and reliability. The utilization of these methods in RF will strengthen the predictive performance of the GWP model.

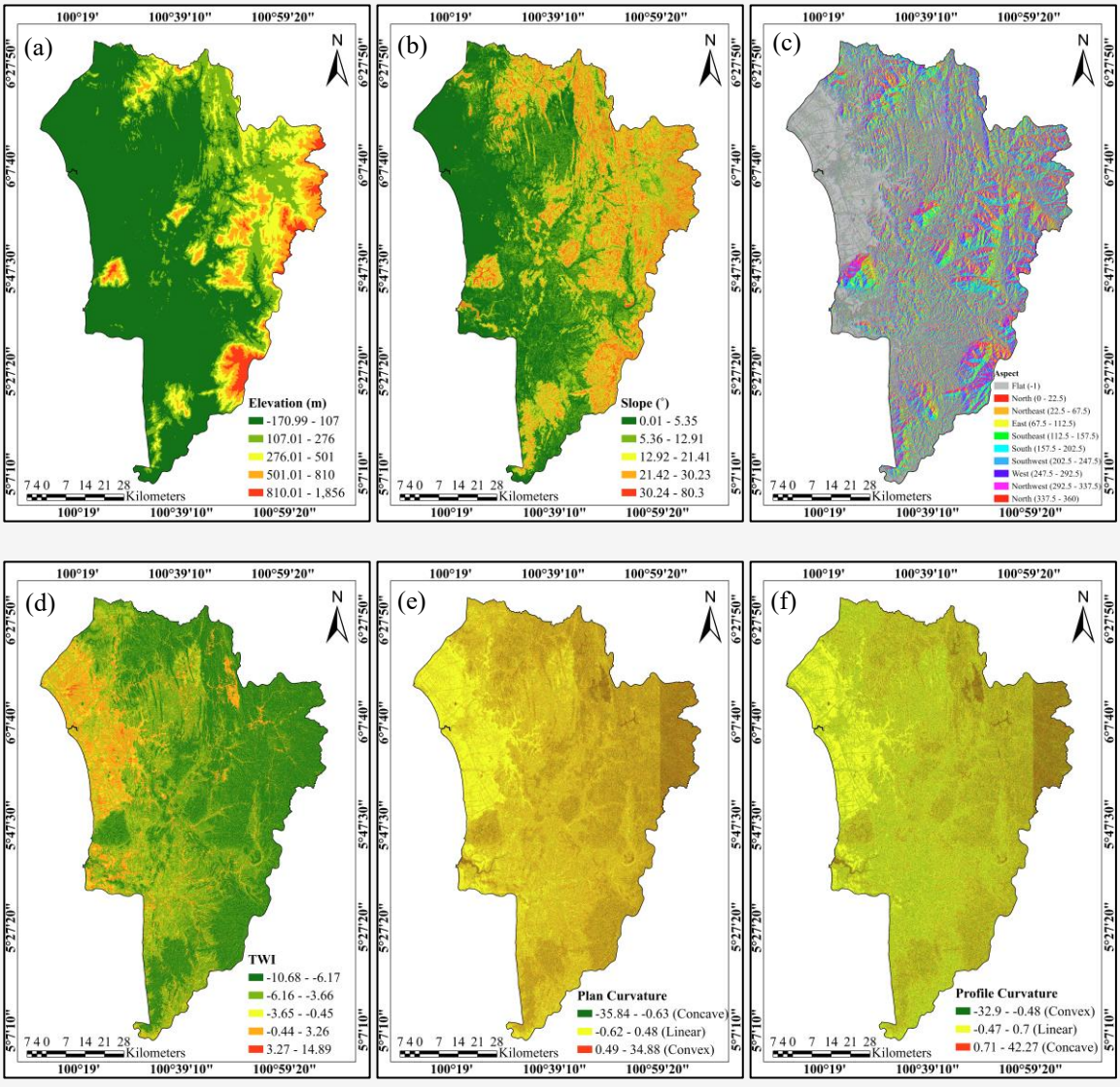

Figure 4: Thematic layers for topography parameters:
(a) elevation, (b) slope, (c) aspect, (d) TWI, (e) plan curvature and (f) profile curvature

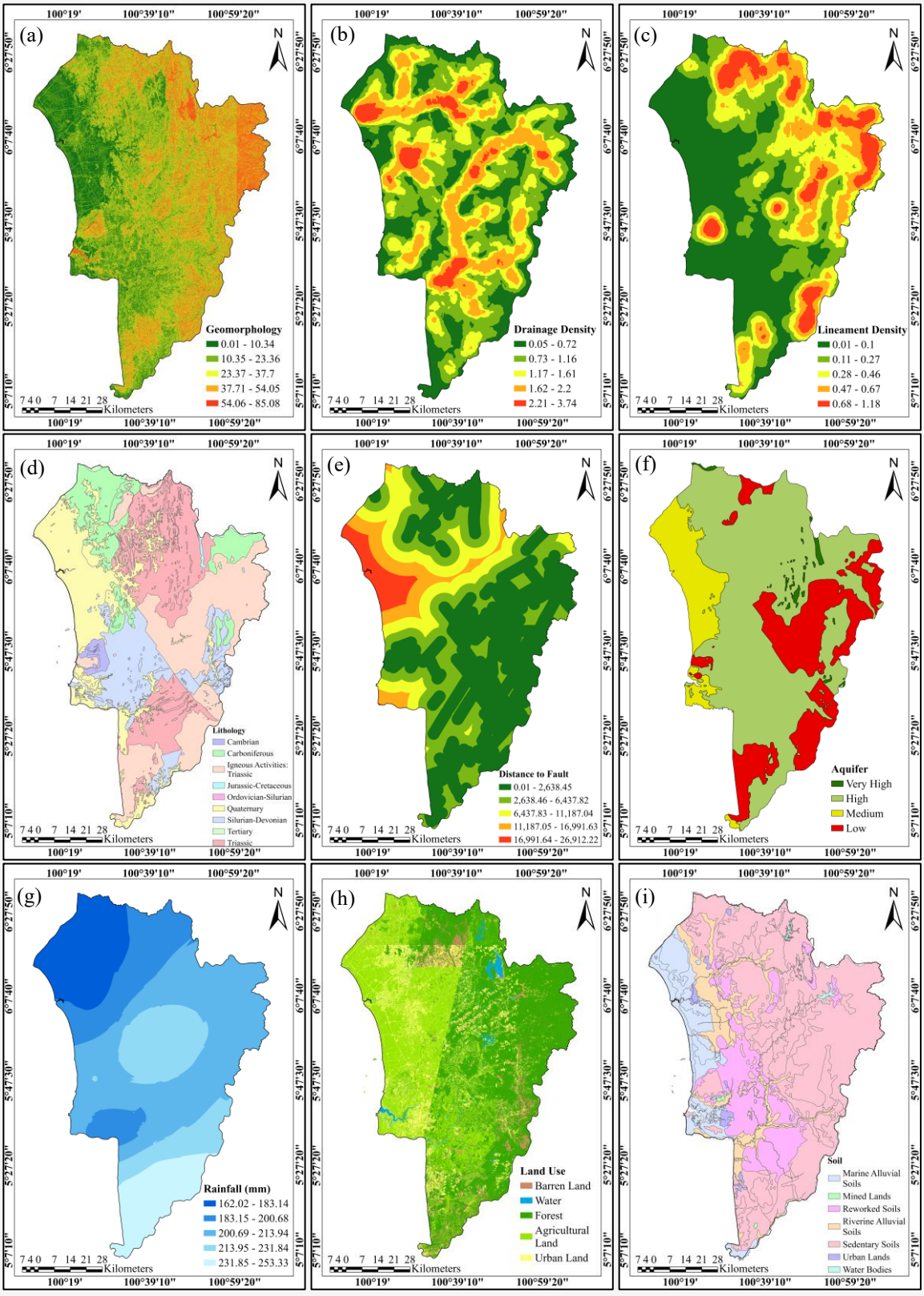

Figure 5: Thematic layers for hydro-geology and environment parameters: (a) geomorphology, (b) drainage density, (c) lineament density, (d) lithology, (e) distance to fault, (f) aquifer, (g) rainfall, (h) land use and (i) soil

3.4 Validation of GWP model

Validation is the crucial step in evaluating the model performances. In this study, several evaluations metrics such as accuracy, precision, sensitivty, specificity, F1-score, kappa and ROC(AUC) were used to measure the effectiveness of the GWP model using RF as shown in Equations 2 to 9.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{Equation 2}$$

$$Precision = \frac{TP}{TP + FP} \quad \text{Equation 3}$$

$$Sensitivity = \frac{TN}{TP + FN} \quad \text{Equation 4}$$

$$Specificity = \frac{TN}{TN + FP} \quad \text{Equation 5}$$

$$Recall = \frac{TP}{TP + FN} \quad \text{Equation 6}$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \text{Equation 7}$$

$$Kappa = \frac{P_0 - P_e}{1 + P_e} \quad \text{Equation 8}$$

$$AUC = \frac{\sum TP + \sum TN}{P + N} \quad \text{Equation 9}$$

Where P and N represent as the actual number of groundwater potential and non-groundwater potential locations. TP (true positive) and TN (true negative) denote the number of correctly predicted groundwater potential and non-groundwater potential areas. Meanwhile, FP (false positive) and FN (false negative) refer to non-groundwater potential and groundwater potential areas that are wrongly predicted. P_0 represents the observed agreement and P_e represents the expected agreement by chance. Based on Equation 2, accuracy represents the ratio of accurately categorized samples compared to its total prediction [27] and [40]. Precision (Equation 3) predicted the proportion of groundwater locations that were correctly identified. Sensitivity (Equation 4) indicates the effectiveness of the model in accurately identifying groundwater locations, while specificity (Equation 5) assesses the effectiveness of the model capacity to correctly identify non-groundwater locations [27] and [40]. Moreover, F1 score (Equation 7) represents the harmonic average of precision and sensitivity [41]. Equation 8 indicates the kappa coefficient, which evaluates the agreement between predicted and actual groundwater potential classification while adjusting for random chance [40]. According to [39] and [42], a kappa coefficient closer to 1 indicates a strong agreement between predicted and actual classification, while a value close to 0 indicates no better than random chance. Lastly, AUC (Equation 9) is used to evaluates the predictive ability of the model performances [25] and [39]. As shown in Table 2, AUC value ranges from 0 to 1, whereas a value closer to 1 describes as excellent performances, while a value below than 0.5 indicates as low performances [12] and [15].

4. Result and Discussion

4.1 GWP Map Using RF

Based on Figure 6, it can be seen that most of the high and very high potential areas cover the western part of Kedah. As shown in Table 4, the very high class occupies the largest area, covering 2,798.18 km^2 (30.97%) and 1,157.56 km^2 (13.01%).

Table 4: Area and percentages of GWP

Classes	Area GWPM (km^2)	Percentage (%)
Very High	2,798.18	30.97
High	1,175.56	13.01
Medium	1,167.53	12.92
Low	1,612.74	17.85
Very Low	2,281.18	25.25
Total	9,035.18	

These areas were largely influenced by several factors, including lower elevation, gentle slope, lower drainage and lineament densities, reworked soil types and extensive agricultural land use, which contribute significantly towards groundwater occurrence [17][40] and [43]. This distribution corresponds with the socioeconomic conditions of Kedah, where the western districts serve as primary agricultural areas, particularly for rice cultivation, providing groundwater a crucial supplementary water source for irrigation and rural populations. Furthermore, most of the low and very low potential areas are mainly concentrated in the northeast part of Kedah.

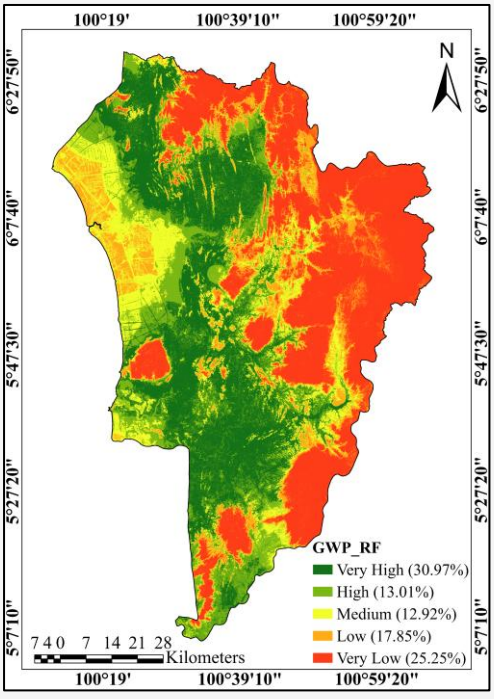

Figure 6: GWP map using RF

According to Table 4, the low class covers an area of 1,612.74 km^2 (17.85%) and very low class accounts for 2,281.18 km^2 (25.25%), representing areas with higher elevations, steeper slopes, sedentary soil types and forest or barren land cover. The restricted groundwater in this region corresponds with the less intensive agricultural activities and lower rural population density which indicates as less dependence on groundwater socioeconomic pressures compared to the western district. Lastly, the medium class encompasses approximately 1,167.53 km^2 (12.92%) of the study area. In general, the spatial distribution of GWP in Kedah highlights the significance influence of several factors such as topography, hydrogeology and land cover towards groundwater availability to provide an essential guideline for groundwater resource management and planning.

4.1.1 Variable importance using RF

Figure 7 depicts the feature importance of the RF model in evaluating the GWP conditioning parameters. As illustrated in the figure, x-axis represents the importance in percentages (0 – 100%) where higher value indicates as greater influence on the model performances. Figure7 shown the highest feature importance of the GWP model were elevation (47.83%), indicating the difference in terrain strongly contributed to the GWP in Kedah. Higher elevation serves as a primary recharge area, while the lowland areas act as discharge zones that accumulate groundwater [18] and [44]. Meanwhile, the lowest feature importance were lithology with the percentages of 0.39%, indicating that geological variation across study areas contributes less to the spatial differentiation of the GWP model compared to the topographical controls. This finding is consistent with the study from [15][18] and [43], indicating that elevation plays a significant role as a primary controlling factor in GWP modelling, while lithology is the least contributing factors in the model performances.

4.2 Validation of GWP Model

The feature selection result indicates eight parameters were retained in GWP model which is elevation, slope, drainage density, lineament density, distance to faults, lulc, rainfall and soil. These parametes gives strong contributions towards groundwater occurrence with their influences on infiltration, recharge, permeability and geological structure. Then, the GWP model was optimized using hyperparameters tuning with the best parameters were: n_estimators: 300, min_samples_split: 15, min_samples_leaf: 15, max_samples: 0.9, max_features: 0.4, max_depth: 7, and bootstrap = True.

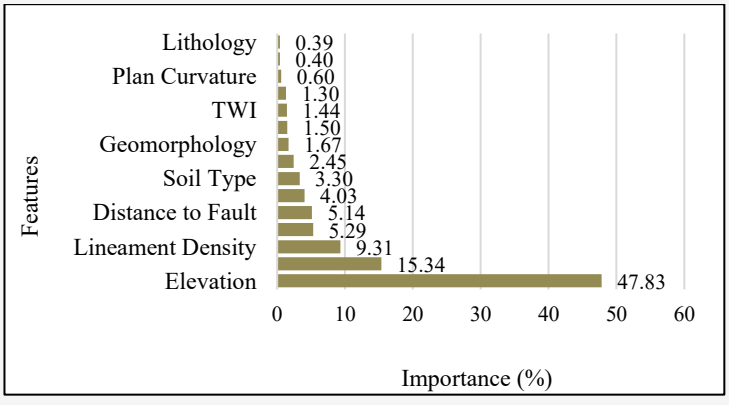

Figure 7: Feature importance

Metric evaluations such as accuracy, precision, sensitivity, specificity, F1-score, kappa and ROC(AUC) have been used to assess the performance of the GWP model (Table 5).

Table 5: Evaluation metrics for training and testing

Evaluation Metrics	Training	Testing
Accuracy	0.85	0.80
Precision	0.81	0.78
Sensitivity	0.91	0.83
Specificity	0.80	0.76
F1-Score	0.86	0.80
Kappa	0.70	0.59
ROC(AUC)	0.95	0.90

The validation was evaluated by using both training and testing dataset. For the training dataset, the model achieved an accuracy of 0.85, precision of 0.81, sensitivity of 0.91, specificity of 0.80 and an F1-score of 0.86. The kappa coefficient of 0.70 demonstrates strong agreement between predicted and observed classes, indicating that the model's performs well beyond random chances and falls within substantial agreement [45]. Meanwhile, the testing dataset gives consistent and dependable results with an accuracy of 0.80, precision of 0.78, sensitivity of 0.83, specificity of 0.76 and F1-score of 0.80. The kappa coefficient score 0.59 which represents as moderate agreement, indicating that the model generalizes well to the unseen data [45].

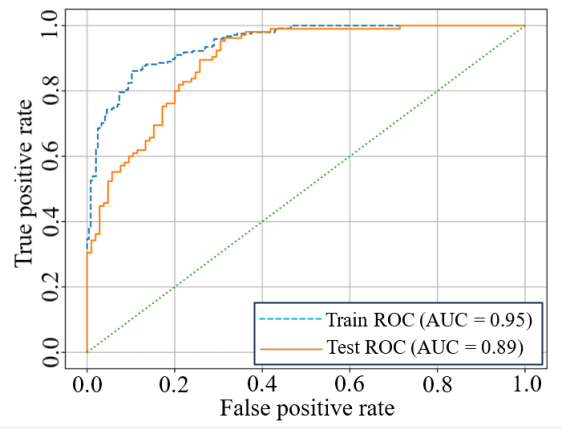

Figure 8: ROC curve

Figure 8 illustrates the ROC curve for GWP model. The AUC for training dataset indicates a value of 0.95, indicating as an excellent model performance with a very high true positive rate at different threshold. This shows that the model is highly effective in correctly identifying groundwater potential zones during training. Meanwhile, the AUC for testing data indicates the value of 0.89 indicates as strong predictive performance on unseen data. Although the testing data is slightly lower than the training, the model continues to generalize well and exhibits no sign of overfitting. These results are consistent with previous GWP studies in Asia, such as [15], which reported AUC values of 0.84 and [41], who obtained AUC values of 0.99. Similarly, [35] and [39] demonstrated RF model consistently exhibits a strong and excellent performances with AUC values exceeding 0.80 which performs better than other ML models. These findings highlights RF is one of the most commonly applied ML models in other studies that gives excellent performance to predict GWP areas.

5. Conclusion

GWP study is essential in determining the water source in any specific area. As the demand for groundwater continues to arise, it becomes essential to evaluate the potential of groundwater resources to prevent any water shortages from happening. This study demonstrates the effectiveness of RF in predicting the GWP areas by integrating 15 conditioning parameters from various factors including topography, hydrogeology and environmental factors. The model demonstrated high prediction ability and accuracy, which confirming its reliability for the GWP mapping. The western and central of Kedah identified as the most promising GWP areas, while the northeast part described as the lowest GWP areas. These insights can be used for authorities and policy makers in identifying the most suitable area for groundwater extraction, while optimizing resource location and avoiding unnecessary expenditure of time, cost and labor. Furthermore, out of all 15 parameters, elevation recognized as the most influencing parameters towards GWP model performances which making it one of the parameters that are suggested to be included in future GWP studies. Subsequently, in Kedah, this study is one of the first study that integrates machine learning with geospatial data and validation against established tube well distribution compared to prior GWP approaches. The findings provide practical insights for groundwater resource management as a guide in establishing the new wells, supporting regional water security strategies, and supporting sustainable agricultural practices in water-scarce regions. Additionally, this study develops a comprehensive framework for GWP mapping that may be applied to other locations which contribute to scientific progress and sustainable resource management in Malaysia.

Acknowledgments

The Ministry of Higher Education supported the research through the Fundamental Research Grant Scheme (Grant no. FRGS/1/2023/WAB07/UITM/02 /3). The researchers would like to express our sincere gratitude to the German Aerospace Agency (DLR), Department of Mineral and Geoscience (DMG), Malaysian Meteorological Department (METMalaysia), Department of Drainage and Irrigation (DID) and Department of Agriculture (DOA) for providing DEM, hydrogeological, rainfall and soil data, which have significantly contributed to this research. We also extend our appreciation to organization that provide open-source data, which is the European Space Agency (ESA), for their invaluable support in facilitating this study. Their contributions have been instrumental in advancing our research. Lastly, we sincerely thanks Universiti Teknologi MARA for providing us with the opportunity and support to advance our research. Their encouragement and resources have been invaluable in facilitating our academic and scientific endeavours.

References

[1] Khan, M. M. A., Raj, K., Rak, A. A. E., Mansor, H. E., Mostapa, R., Samuding, K. and Shah, Z. A., (2021). Stable Isotope Evidence on Mechanisms and Sources of Groundwater Recharge in Quaternary Aquifers of Kelantan, Malaysia. *Arabian Journal of Geosciences,* Vol. 14(16). https://doi.org/10.1007/s12517-021-07646-7.

[2] Kura, N. U., Ramli, M. F., Sulaiman, W. N. A., Ibrahim, S. and Aris, A. Z., (2018). An Overview of Groundwater Chemistry Studies in Malaysia. *Environmental Science and Pollution Research,* Vol. 25(8); 7231–7249. https://doi.org/10.1007/s11356-015-5957-6.

[3] Ab Rashid, M. F., Abd Rahman, A. and Abdul Rashid, S. M. R., (2021). Analyzing the Factors and Effects of Water Supply Disruption in Penang Island, Malaysia. *Malaysian Journal of Society and Space,* Vol. 17(3). https://doi.org/10.17576/geo-2021-1703-05.

[4] Anang, Z., Padli, J., Abdul Rashid, N. K., Alipiah, R. M. and Musa, H., (2019). Factors Affecting Water Demand: Macro Evidence in Malaysia. *Jurnal Ekonomi Malaysia,* Vol. 53(1); 17–25. https://doi.org/10.17576/JEM-2019-5301-2.

[5] Wan Ahmad, W. A. A., Nik Sulaiman, N. M. and Mahmood, N. Z., (2025). Unveiling the Future Water Footprint of Paddy Cultivation: Evidence from a Humid Tropical Country. Environmental Research Communications, Vol. 7(9). https://doi.org/10.1088/2515-7620/ae0982

[6] Lee, W. K., Faba, S. K. and Muhamad, N. S., (2025). Water Availability and Demand Analysis for Kedah River Basin Using Water Evaluation and Planning (WEAP) Model. *In Sustainable Green Infrastructure: Materials and Technologies,*139–154. Springer Nature. https://doi.org/10.1007/978-981-96-1486-8_8.

[7] Ghosh, A. and Bera, B., (2024). Potentialities and Development of Groundwater Resources Applying Machine Learning Models in the Extended Section of Manbhum-Singhbhum Plateau, India. *HydroResearch,* Vol. 7; 1–14. https://doi.org/10.1016/j.hydres.2023.11.002.

[8] Masroor, M., Rehman, S., Sajjad, H., Rahaman, M. H., Sahana, M., Ahmed, R. and Singh, R., (2021). Assessing the Impact of Drought Conditions on Groundwater Potential in Godavari Middle Sub-Basin, India using Analytical Hierarchy Process and Random Forest Machine Learning Algorithm. *Groundwater for Sustainable Development,* Vol. 13. https://doi.org/10.1016/j.gsd.2021.100554.

[9] Ponnusamy, D. and Elumalai, V., (2022). Determination of Potential Recharge Zones and Its Validation Against Groundwater Quality Parameters through the Application of GIS and Remote Sensing Techniques in Mhlathuze Catchment, KwaZulu-Natal, South Africa. *Chemosphere,* Vol. 307. https://doi.org/10.1016/j.chemosphere.2022.136121.

[10] Azma, A., Narreie, E., Shojaaddini, A., Kianfar, N., Kiyanfar, R., Alizadeh, S. M. S. and Davarpanah, A., (2021). Statistical Modeling for Spatial Groundwater Potential Map Based on GIS Technique. *Sustainability (Switzerland),* Vol. 13(7). https://doi.org/10.3390/su13073788.

[11] Das, R. and Saha, S., (2022). Spatial Mapping of Groundwater Potentiality Applying Ensemble of Computational Intelligence and Machine Learning Approaches. *Groundwater for Sustainable Development,* Vol. 18. https://doi.org/10.1016/j.gsd.2022.100778.

[12] Prasad, P., Loveson, V. J., Kotha, M. and Yadav, R., (2020). Application of Machine Learning Techniques in Groundwater Potential Mapping Along the West Coast of India. *GIScience and Remote Sensing,* 735–752. https://doi.org/10.1080/15481603.2020.1794104.

[13] Maskooni, E. K., Naghibi, S. A., Hashemi, H. and Berndtsson, R., (2020). Application of Advanced Machine Learning Algorithms to Assess Groundwater Potential Using Remote Sensing-Derived Data. *Remote Sensing,* Vol. 12(17). https://doi.org/10.3390/RS12172742.

[14] Dahal, K., Sharma, S., Shakya, A., Talchabhadel, R., Adhikari, S., Pokharel, A., Sheng, Z., Pradhan, A. M. S. and Kumar, S., (2023). Identification of Groundwater Potential Zones in Data-Scarce Mountainous Region Using Explainable Machine Learning. *Journal of Hydrology,* Vol. 627. https://doi.org/10.1016/j.jhydrol.2023.130417.

[15] Roy, S. K., Hasan, M. M., Mondal, I., Akhter, J., Roy, S. K., Talukder, S., Islam, A. K. M. S., Rahman, A. and Karuppannan, S., (2024). Empowered Machine Learning Algorithm to Identify Sustainable Groundwater Potential Zone Map in Jashore District, Bangladesh. *Groundwater for Sustainable Development,* Vol. 25. https://doi.org/10.1016/j.gsd.2024.101168.

[16] Moghaddam, D. D., Rahmati, O., Panahi, M., Tiefenbacher, J., Darabi, H., Haghizadeh, A., Haghighi, A. T., Nalivan, O. A. and Tien Bui, D., (2020). The Effect of Sample Size on Different Machine Learning Models for Groundwater Potential Mapping in Mountain Bedrock Aquifers. *Catena,* Vol. 187. https://doi.org/10.1016/j.catena.2019.104421.

[17] Wei, A., Li, D., Bai, X., Wang, R., Fu, X. and Yu, J., (2022). Application of Machine Learning to Groundwater Spring Potential Mapping Using Averaging, Bagging, and Boosting Techniques. *Water Supply,* Vol. 22(8), 6882–6894. https://doi.org/10.2166/ws.2022.283.

[18] Fatah, K. K., Mustafa, Y. T. and Hassan, I. O., (2024). Groundwater Potential Mapping in Arid and Semi-Arid Regions of Kurdistan Region of Iraq: A Geoinformatics-Based Machine Learning Approach. *Groundwater for Sustainable Development,* Vol. 27. https://doi.org/10.1016/j.gsd.2024.101337.

[19] Kumar, P., Singh, P., Asthana, H., Yadav, B. and Mukherjee, S., (2024). Groundwater Potential Zone Mapping of Middle Andaman Using Multi-Criteria Decision-Making and Support Vector Machine. *Groundwater for Sustainable Development,* Vol. 26. https://doi.org/10.1016/j.gsd.2024.101191.

[20] Al-Kindi, K. M. and Janizadeh, S., (2022). Machine Learning and Hyperparameters Algorithms for Identifying Groundwater Aflaj Potential Mapping in Semi-Arid Ecosystems Using LiDAR, Sentinel-2, GIS Data, and Analysis. *Remote Sensing,* Vol. 14(21). https://doi.org/10.3390/rs14215425.

[21] Islam, F., Tariq, A., Guluzade, R., Zhao, N., Shah, S. U., Ullah, M., Hussain, M. L., Ahmad, M. N., Alasmari, A., Alzuaibr, F. M., Askary, A. E. and Aslam, M., (2023). Comparative Analysis of GIS And RS Based Models for Delineation of Groundwater Potential Zone Mapping. *Geomatics, Natural Hazards and Risk,* Vol. 14(1). https://doi.org/10.1080/19475705.2023.2216852.

[22] Sharma, Y., Ahmed, R., Saha, T. K., Bhuyan, N., Kumari, G., Roshani, Pal, S. and Sajjad, H., (2024). Assessment of Groundwater Potential and Determination of Influencing Factors Using Remote Sensing and Machine Learning Algorithms: A study of Nainital district of Uttarakhand State, India. *Groundwater for Sustainable Development,* Vol. 25. https://doi.org/10.1016/j.gsd.2024.101094.

[23] Moore, I. D., Grayson, R. B. and Ladson, A. R., (1991). Digital Terrain Modelling: A Review of Hydrological, Geomorphological, and Biological Applications. *Hydrological Processes,* Vol. 5(1); 3–30. https://doi.org/10.1002/hyp.3360050103.

[24] Chen, W., Zhao, X., Tsangaratos, P., Shahabi, H., Ilia, I., Xue, W., Wang, X. and Ahmad, B. B., (2020). Evaluating the Usage of Tree-Based Ensemble Methods in Groundwater Spring Potential Mapping. *Journal of Hydrology,* Vol. 583. https://doi.org/10.1016/j.jhydrol.2020.124602.

[25] Kumar, M., Singh, P. and Singh, P., (2023). Machine Learning and GIS-RS-Based Algorithms for Mapping the Groundwater Potentiality in the Bundelkhand region, India. *Ecological Informatics,* Vol. 74. https://doi.org/10.1016/j.ecoinf.2023.101980.

[26] Kumar, R., Dwivedi, S. B. and Gaur, S., (2021). A Comparative Study of Machine Learning and Fuzzy-AHP Technique to Groundwater Potential Mapping in the Data-Scarce Region. *Computers and Geosciences,* Vol. 155. https://doi.org/10.1016/j.cageo.2021.104855.

[27] Liu, R., Li, G., Wei, L., Xu, Y., Gou, X., Luo, S. and Yang, X. (2022). Spatial Prediction of Groundwater Potentiality Using Machine Learning Methods with Grey Wolf and Sparrow Search Algorithms. *Journal of Hydrology,* Vol. 610. https://doi.org/10.1016/j.jhydrol.2022.127977.

[28] Dey, B., Abir, K. A. M., Ahmed, R., Salam, M. A., Redowan, M., Miah, M. D. and Iqbal, M. A., (2023). Monitoring Groundwater Potential Dynamics of North-Eastern Bengal Basin in Bangladesh Using AHP-Machine Learning Approaches. *Ecological Indicators,* Vol. 154. https://doi.org/10.1016/j.ecolind.2023.110886.

[29] Ahmed, R. and Sajjad, H., (2018). Analyzing Factors of Groundwater Potential and Its Relation with Population in the Lower Barpani Watershed, Assam, India. *Natural Resources Research,* Vol. 27(4); 503–515. https://doi.org/10.1007/s11053-017-9367-y.

[30] Masroor, M., Sajjad, H., Kumar, P., Saha, T. K., Rahaman, M. H., Choudhari, P., Kulimushi, L. C., Pal, S. and Saito, O., (2023). Novel Ensemble Machine Learning Modeling Approach for Groundwater Potential Mapping in Parbhani District of Maharashtra, India. *Water (Switzerland),* Vol. 15(3). https://doi.org/10.3390/w15030419.

[31] Abrar, H., Legesse Kura, A., Esayas Dube, E. and Likisa Beyene, D., (2023). AHP Based Analysis of Groundwater Potential in the Western Escarpment of the Ethiopian Rift Valley. *Geology, Ecology, and Landscapes,* Vol. 7(3); 175–188. https://doi.org/10.1080/24749508.2021.1952761.

[32] Jari, A., Bachaoui, E. M., Jellouli, A., Harti, A. El, Khaddari, A. and Jazouli, A. El. (2022). Use of GIS, Remote Sensing and Analytical Hierarchy Process for Groundwater Potential Assessment in an Arid Region – A Case Study. *Ecological Engineering and Environmental Technology,* Vol. 23(5); 234–255. https://doi.org/10.12912/27197050/152141.

[33] Breiman, L., (2001). Random Forests. *Machine Learning,* Vol. 45(1); 5–32. https://doi.org/10.1023/A:1010933404324.

[34] Al-Abadi, A. M., Fryar, A. E., Rasheed, A. A. and Pradhan, B., (2021). Assessment of Groundwater Potential in Terms of the Availability and Quality of the Resource: A Case Study from Iraq. *Environmental Earth Sciences,* Vol. 80(12). https://doi.org/10.1007/s12665-021-09725-0.

[35] Arabameri, A., Pal, S. C., Rezaie, F., Nalivan, O. A., Chowdhuri, I., Saha, A., Lee, S. and Moayedi, H., (2021). Modeling Groundwater Potential Using Novel GIS-Based Machine-Learning Ensemble Techniques. *Journal of Hydrology: Regional Studies,* Vol. 36. https://doi.org/10.1016/j.ejrh.2021.100848.

[36] Khan, Z. A. and Jhamnani, B., (2023). Identification of Groundwater Potential Zones of Idukki District Using Remote Sensing and GIS-Based Machine-Learning Approach. *Water Supply,* Vol. 23(6); 2426–2446. https://doi.org/10.2166/ws.2023.134.

[37] Morgan, H., Madani, A., Hussien, H. M. and 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,* Vol. 10(1). https://doi.org/10.1186/s40562-023-00261-2.

[38] Arabameri, A., Roy, J., Saha, S., Blaschke, T., Ghorbanzadeh, O. and Bui, D. T., (2019). Application of Probabilistic and Machine Learning Models for Groundwater Potentiality Mapping in Damghan Sedimentary Plain, Iran. *Remote Sensing,* Vol. 11(24). https://doi.org/10.3390/rs11243015.

[39] Mussa, M. M., Lohani, T. K. and Eshete, A. A., (2024). Evaluation of Groundwater Potential Zones Using GIS-Based Machine Learning Ensemble Models in the Gidabo Watershed, Ethiopia. *Global Challenges.* https://doi.org/10.1002/gch2.202400137.

[40] Vafadar, S., Rahimzadegan, M. and Asadi, R. (2023). Evaluating the Performance of Machine Learning Methods and Geographic Information System (GIS) in Identifying Groundwater Potential Zones in Tehran-Karaj Plain, Iran. *Journal of Hydrology,* Vol. 624. https://doi.org/10.1016/j.jhydrol.2023.129952.

[41] Halder, K., Srivastava, A. K., Ghosh, A., Nabik, R., Pan, S., Chatterjee, U., Bisai, D., Pal, S. C., Zeng, W., Ewert, F., Gaiser, T., Pande, C. B., Islam, A. R. M. T., Alam, E. and Islam, M. K., (2024). Application of Bagging and Boosting Ensemble Machine Learning Techniques for Groundwater Potential Mapping in A Drought-Prone Agriculture Region of Eastern India. *Environmental Sciences Europe,* Vol. 36(1). https://doi.org/10.1186/s12302-024-00981-y.

[42] Rwanga, S. S. and Ndambuki, J. M., (2017). Accuracy Assessment of Land Use/Land Cover Classification Using Remote Sensing and GIS. *International Journal of Geosciences,* Vol. 08(04); 611–622. https://doi.org/10.4236/ijg .2017.84033.

[43] Anh, D. T., Pandey, M., Mishra, V. N., Singh, K. K., Ahmadi, K., Janizadeh, S., Tran, T. T., Linh, N. T. T. and Dang, N. M., (2023). Assessment of Groundwater Potential Modeling Using Support Vector Machine Optimization Based on Bayesian Multi-Objective Hyperparameter Algorithm. *Applied Soft Computing,* Vol. 132. https://doi.org/10. 1016/j.asoc.2022.109848.

[44] Rizeei, H. M., Pradhan, B., Saharkhiz, M. A. and Lee, S., (2019). Groundwater Aquifer Potential Modeling Using an Ensemble Multi-Adoptive Boosting Logistic Regression Technique. *Journal of Hydrology,* Vol. 579. https://doi.org/10.1016/j.jhydrol.2019.124172.

[45] Landis, J. R. and Koch, G. G., (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics,* Vol. 33(1); 159–174. https://doi.org/10.2307/2529310.