sustainability

MDPI

Article

The Predictive Capability of a Novel Ensemble Tree-Based Algorithm for Assessing Groundwater Potential

Soyoung Park [1] and Jinsoo Kim [2,*]

[1] Geomatics Research Institute, Pukyong National University, Busan 48513, Korea; yac100@pknu.ac.kr
[2] Department of Spatial Information Engineering, Pukyong National University, Busan 48513, Korea
* Correspondence: jinsookim@pknu.ac.kr; Tel.: +82-51-629-6658

Abstract: Understanding the potential groundwater resource distribution is critical for sustainable groundwater development, conservation, and management strategies. This study analyzes and maps the groundwater potential in Busan Metropolitan City, South Korea, using random forest (RF), gradient boosting machine (GBM), and extreme gradient boosting (XGB) methods. Fourteen groundwater conditioning factors were evaluated for their contribution to groundwater potential assessment using an elastic net. Curvature, the stream power index, the distance from drainage, lineament density, and fault density were excluded from the subsequent analysis, while nine other factors were used to create groundwater potential maps (GMPs) using the RF, GBM, and XGB models. The accuracy of the resultant GPMs was tested using receiver operating characteristic curves and the seed cell area index, and the results were compared. The analysis showed that the three models used in this study satisfactorily predicted the spatial distribution of groundwater in the study area. In particular, the XGB model showed the highest prediction accuracy (0.818), followed by the GBM (0.802) and the RF models (0.794). The XGB model, which is the most recently developed technique, was found to best contribute to improving the accuracy of the GPMs. These results contribute to the establishment of a sustainable management plan for groundwater resources in the study area.

Keywords: random forest; gradient boosting machine; extreme gradient boosting; groundwater potential assessment; groundwater potential map

Citation: Park, S.; Kim, J. The Predictive Capability of a Novel Ensemble Tree-Based Algorithm for Assessing Groundwater Potential. *Sustainability* **2021**, *13*, 2459. https://doi.org/10.3390/su13052459

Academic Editor: Dino Musmarra

Received: 18 January 2021
Accepted: 17 February 2021
Published: 25 February 2021

1. Introduction

Groundwater quality is typically superior to that of surface water, and is less susceptible to weather changes and contamination from the land surface. Therefore, groundwater is widely used as a secondary water source with a reliable and safe supply of water resources. However, in recent decades, ensuring sustainable groundwater resources has been threatened by an increased demand and the excessive exploitation of groundwater resources that result from population growth and economic development.

Because groundwater resources are not infinite, it is imperative to secure sustainable and high-quality water resources through efficient utilization and systematic management. Groundwater potential maps (GPMs) are useful to establish sustainable groundwater development, conservation, and management strategies [1,2]. These maps show where groundwater can be accessed without expensive and labor-intensive endeavors and provide insights into further groundwater resources. For this reason, it is essential to secure an accurate spatial prediction in potential groundwater resource areas.

To this end, GPMs based on geographic information systems and remote sensing are widely used. In earlier studies, various statistical techniques, such as the frequency ratio, index of entropy, weights of evidence, evidential belief function, and logistic regression, have been applied [3–9]. These techniques can easily be applied and interpret the correlations between groundwater and groundwater conditioning factors.

Recently, machine learning techniques, a branch of artificial intelligence, have been widely adopted in various fields. These techniques rely on the concept that systems

can learn from data, recognize patterns, and make choices with the least human intervention [10]. Various machine learning techniques, such as artificial neural networks (ANN), support vector machines (SVM), linear discriminant analysis, quadratic discriminant analysis, the k-nearest neighbor algorithm, multivariate adaptive regression splines, and decision trees [11–17], have widely been adopted for the analysis and mapping of groundwater potential.

Among these machine learning techniques, the tree-based algorithm is one of the most commonly used techniques. In particular, an ensemble model that improves the goodness-of-fit and prediction ability of decision trees has widely been adopted [18]. The ensemble model is a meta-algorithm that combines multiple machine learning techniques into a single surrogate model, and utilizes bagging, boosting, and stacking to reduce variance and deviation, thereby enhancing prediction accuracy [19].

Bagging (bootstrap aggregating) is an ensemble technique that predicts results by generating multiple decision trees independently of each other [20]. In boosting, multiple decision trees are grown sequentially using information from existing trees [21]. The random forest (RF) and gradient boosting machine (GBM), which are representative algorithms for bagging and boosting, respectively, have been extensively applied for groundwater potential analysis [11–17,22–24].

In previous studies, various techniques have been applied that focus on the generation of a robust model. The key to a robust model is to improve model accuracy and secure generalization performance. Machine learning, especially tree-based ensemble models, have been demonstrated to have excellent prediction performance through comparative studies with the above-mentioned machine learning techniques [11–14]. Although ensemble models have provided reliable and accurate results, most studies showed that these models had high accuracy for training data, but decreased accuracy for validation data.

This problem, called overfitting, can be defined as the model having been trained excessively well to fit training data. As a result, the model reproduces noise and peculiarities of the training data instead of finding a general predictive rule [25,26]. Avoiding overfitting is a crucial part of securing generalization performance. However, it is not easily done because it can occur through various causes. Some common strategies to prevent overfitting problems include regularization, early stopping, and abundant datasets [27,28].

An extreme gradient boosting (XGB) algorithm, which is a more regularized form of GBM, could be a good alternative. The XGB algorithm can utilize regularization and early stopping to reduce overfitting further, and improve prediction accuracy and generalization capability [29]. This algorithm was developed by Chen and Guestrin [30] and improves upon the GBM model. XGB has drawn attention in some major competitions in the big data field, such as the Kaggle and DataCastl competitions, owing to its speed and accuracy [31]. XGB has been applied in various research domains, such as landslide susceptibility analysis [32,33], particulate matter prediction [34,35], land cover classification [36,37], undrained shear strength [38], and soil liquefaction assessment [39]. However, only a few studies applied the algorithm in groundwater potential analysis.

Therefore, this study aimed to assess groundwater potential using RF, GBM, and XGB techniques, and to produce GPMs based on these groundwater potential models. In addition, the performances of the three GPMs were evaluated and compared based on receiver operating characteristic (ROC) curves and the seed cell area index (SCAI). Throughout this process, the ultimate aim was to determine whether the XGB model is robust enough to create accurate GPMs.

2. Study Area

Busan Metropolitan City served as the study area. It is located at the southeastern end of the Korean Peninsula at 128°45′54″–129°18′13″ E and 34°53′12″–35°23′36″ N. In this study, the groundwater potential was analyzed for inland areas, excluding some island areas located at the bottom of Busan Metropolitan City; the study area covers a total area

of approximately 747 km^2 (Figure 1). The eastern part of the study area is hilly, with an elevation of 400–800 m, whereas the western portion consists of low-elevation plains.

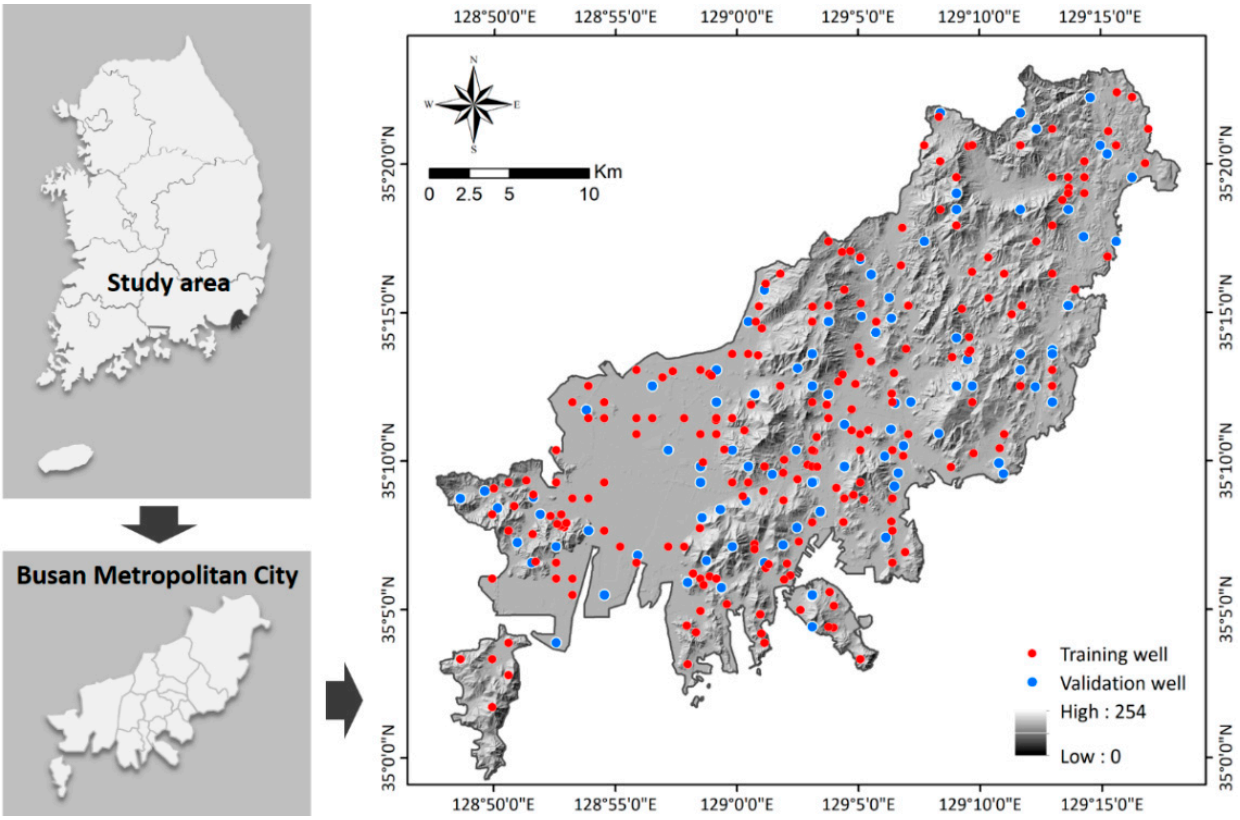


Figure 1. Location of the study area and well locations with hill-shading produced using a 1:5000 topographic digital map obtained from the National Geographic Information Institute.

Busan Metropolitan City is in the temperate monsoon climate zone, and touches the Straits of Korea; thus, it is characterized by an oceanic climate. From 1981 to 2010, the annual average temperature was 14.7 °C, and the average maximum and minimum temperatures were 18.9 °C and 11.3 °C, respectively, showing little difference between summer and winter temperatures. The annual average precipitation was 1519.1 mm, which is considerably higher than the national average (1279.5 mm), and approximately 62% of the annual precipitation is concentrated in May–August [40]. Most of the study area is forested (approximately 45%), while urban and agricultural areas account for approximately 25% and 16%, respectively.

At of the end of 2018, the amount of groundwater used in Busan Metropolitan City was 28,558,796 m^3/year, which comprises approximately 29% of the exploitable groundwater resources (97,553 thousand m^3/year). Approximately 76% of the total groundwater use is for municipal water, followed by agricultural use (fisheries) (approximately 14%), other uses (approximately 6%, including bottled water and hot springs water), and industrial use (approximately 4%). The amount of groundwater use per unit area of the city is 37,092 m^3/year/km^2, which is higher than that of the nationwide groundwater use per unit area (29,074 m^3/year/km^2) [40].

In the study area as a whole, there is much room for further groundwater development; however, nine out of 16 cities/counties/districts used groundwater in excess of the exploitable groundwater resource capacity. In addition, charges for groundwater use were imposed on the entire Busan Metropolitan City area. Therefore, Busan Metropolitan City is a region with relatively high groundwater dependence, and a management plan should be developed to ensure that groundwater resources are used more efficiently and effectively.

3. Materials and Methods

The present study was conducted by applying the following steps (Figure 2): (1) designing a geospatial database that included a groundwater wells feature class and conditioning factors; (2) selecting conditioning factors by use of the feature selection function; (3) generating training and validation datasets for modeling groundwater potential; (4) generating GPMs using RF, GBM, and XGB models; and (5) validation and performance comparison among the produced GPMs. The dataset was prepared with a spatial resolution of 10 m, and the maps were generated using ArcGIS version 10.5 (ESRI, Inc., Redlands, CA, USA). The statistical computations of all three models were conducted using R version 3.5.2 (Foundation for Statistical Computing, Vienna, Austria).

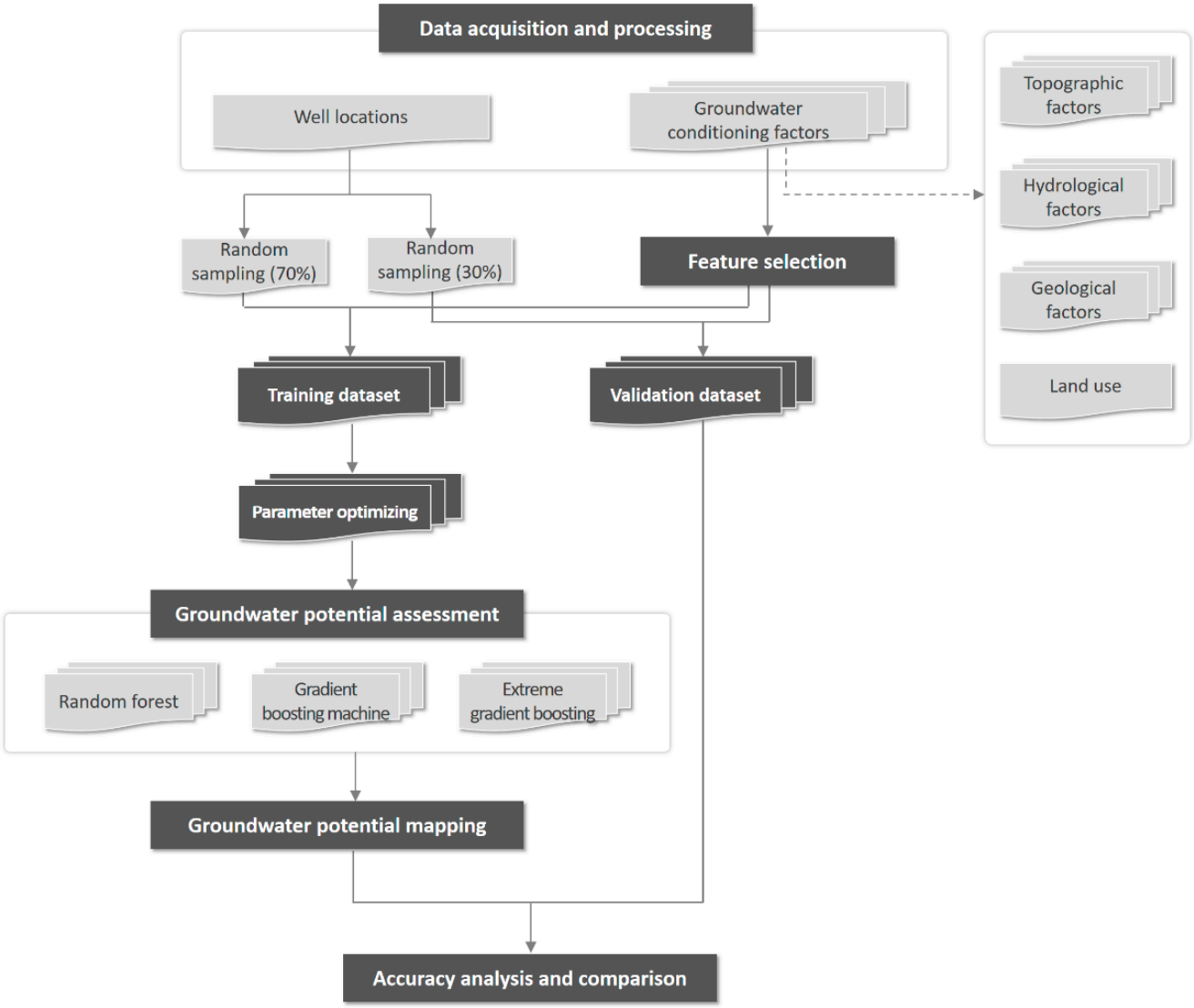


Figure 2. Flow chart of work process.

3.1. Construction of the Spatial Database

3.1.1. Well Data

The groundwater well data used in this study were collected from data obtained through field observations and measurements by Busan Metropolitan City, the groundwater basic survey report, National Groundwater Information Center (gims.go.kr, accessed on 28 January 2020), and Korea Rural Community Corporation. The collected groundwater well data were randomly classified into 153 total datasets, 70% of which were used as the model

training datasets (107); the remaining 30% were used as the model validation datasets (46). Figure 1 shows the locations of the groundwater well data used in this study.

To apply machine learning, data in areas without any groundwater wells were needed to analyze the groundwater potential. The geographical data of the area without wells were extracted with an interval of 20 px (100 m). Then, the same number of areas with wells were determined by a random selection method. The value of 0 was assigned to these areas for analysis by using the GBM, RF, and XGB models.

Finally, the training and validation datasets were generated by assigning new values to all groundwater conditioning factors and applying them to the data, including 214 points and 92 points. The training dataset was used to obtain models using RF, XGB, and GBM. The validation dataset was used to verify the modeled outputs.

3.1.2. Groundwater Conditioning Factors

A total of fourteen groundwater conditioning factors were selected based on a literature review of previous studies. These factors were constructed using thematic maps such as topographic digital maps, geological maps, and land cover maps that were obtained from the government and related organizations (Table 1).

Table 1. Data sources used in this study.

Category	Factor	Source	Scale (Resolution)	GIS and Data Type
Well location	-	National research paper Local research paper Field survey	-	Point
Topographical factors	Elevation	Topographic digital map [1]	1:5000	Polyline, point
	Slope degree Slope aspect Curvature	Digital elevation model	10 × 10 m	Raster
Hydrological factors	Topographic wetness index Stream power index Distance from drainage Drainage density	Digital elevation model	10 × 10 m	Raster
Geological factors	Lithology	Geology map [2]	1:50,000	Polygon
	Distance from lineament Lineament density	Hill-shaded map	10 × 10 m	Raster
	Distance from fault Fault density	Geology map	1:50,000	Polyline
Land cover	Land cover	Land cover map [3]	1:25,000	Polygon

[1] Topographic digital maps were obtained from the National Geographic Information Institute, Korea; [2] Geological maps were obtained from the Korea Institute of Geoscience and Mineral Resources, [3] and cover maps were obtained from the Ministry of Environment, Korea.

Topographic factors such as elevation, slope degree, slope aspect, and curvature were extracted from a digital elevation model (DEM) created using a 1:5000 topographic digital map from the National Geographic Information Institute (Figure 3). The topography of a region is subject to erosion and sedimentation, which in turn affect the physicochemical properties of the soil, and the concentration and movement of surface water and groundwater; thus, topography is one of the key factors that influences a wide range of environmental variables [41].

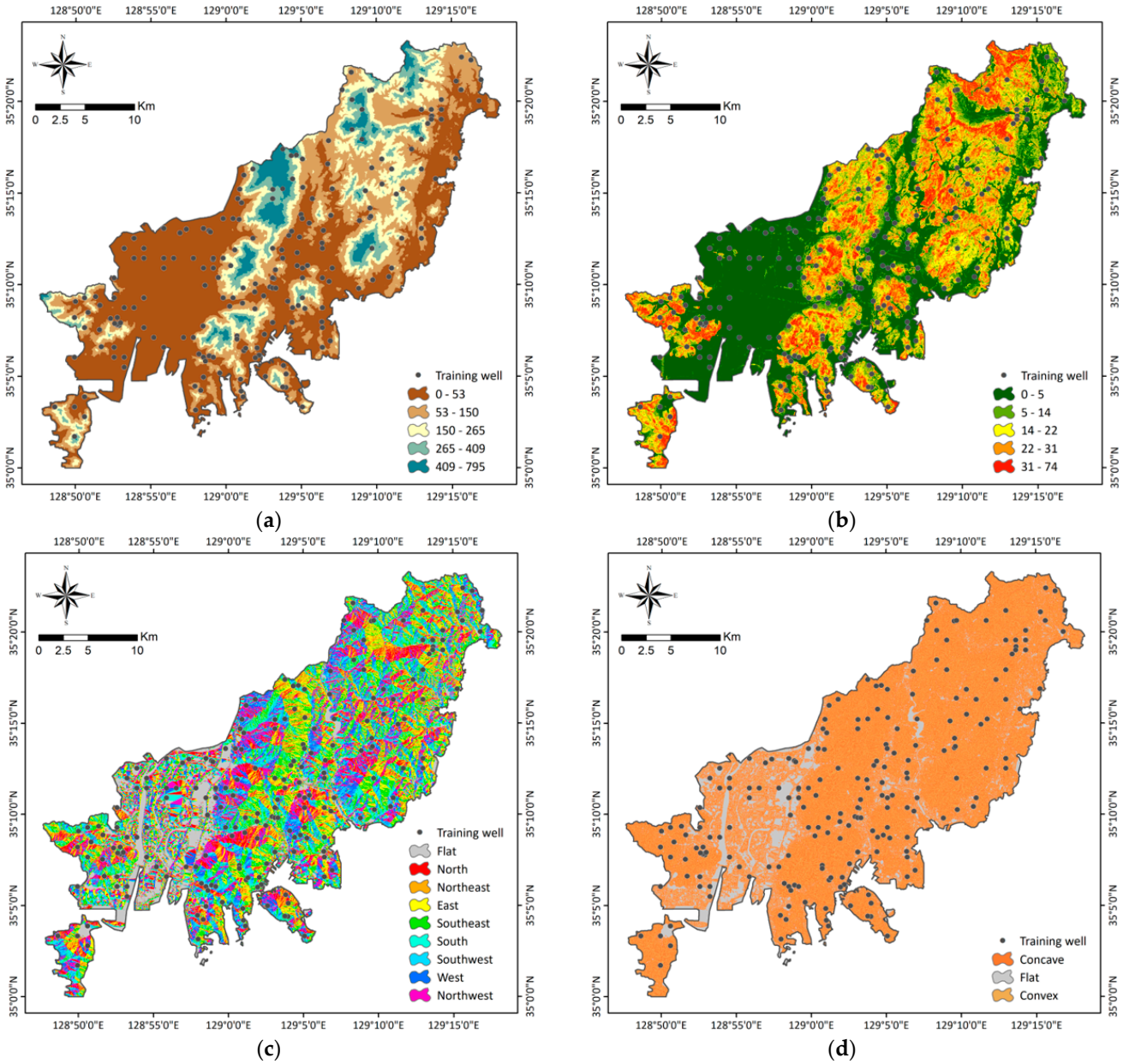


Figure 3. Topographic conditioning factors: (**a**) elevation in meter, (**b**) slope in degree, (**c**) aspect classified according to slope angle rage, and (**d**) curvature classified by negative (concave), zero, and positive (convex) values.

Hydrological factors include the topographic wetness index (TWI), the stream power index (SPI), distance from drainage, and drainage density (Figure 4). Both TWI and SPI are factors used to consider the flow characteristics of surface water and groundwater according to topographic factors and are calculated using the following equations [42]:

$$\mathrm{TWI} = \ln\left(\frac{\alpha}{\tan\beta}\right) \quad (1)$$

$$\mathrm{SPI} = A_s \times \tan\beta \quad (2)$$

where A_s is the specific catchment area, β is the local slope gradient, and α is the local upslope area. In addition, the drainage density is a factor related to permeability and surface runoff. Higher values of drainage density lead to lower values of permeability and higher values of surface runoff. Moreover, drainage density has an indirect effect on the study area's groundwater potential [43]. The drainage density of each cell was calculated

by dividing the surface area (km^2) by the total length of the drainage (km) in the same cell. The drainage density values were calculated using the line density function, whereas the distances from the drainage were determined using the Euclidean distance function using the ArcGIS 10.5 software.

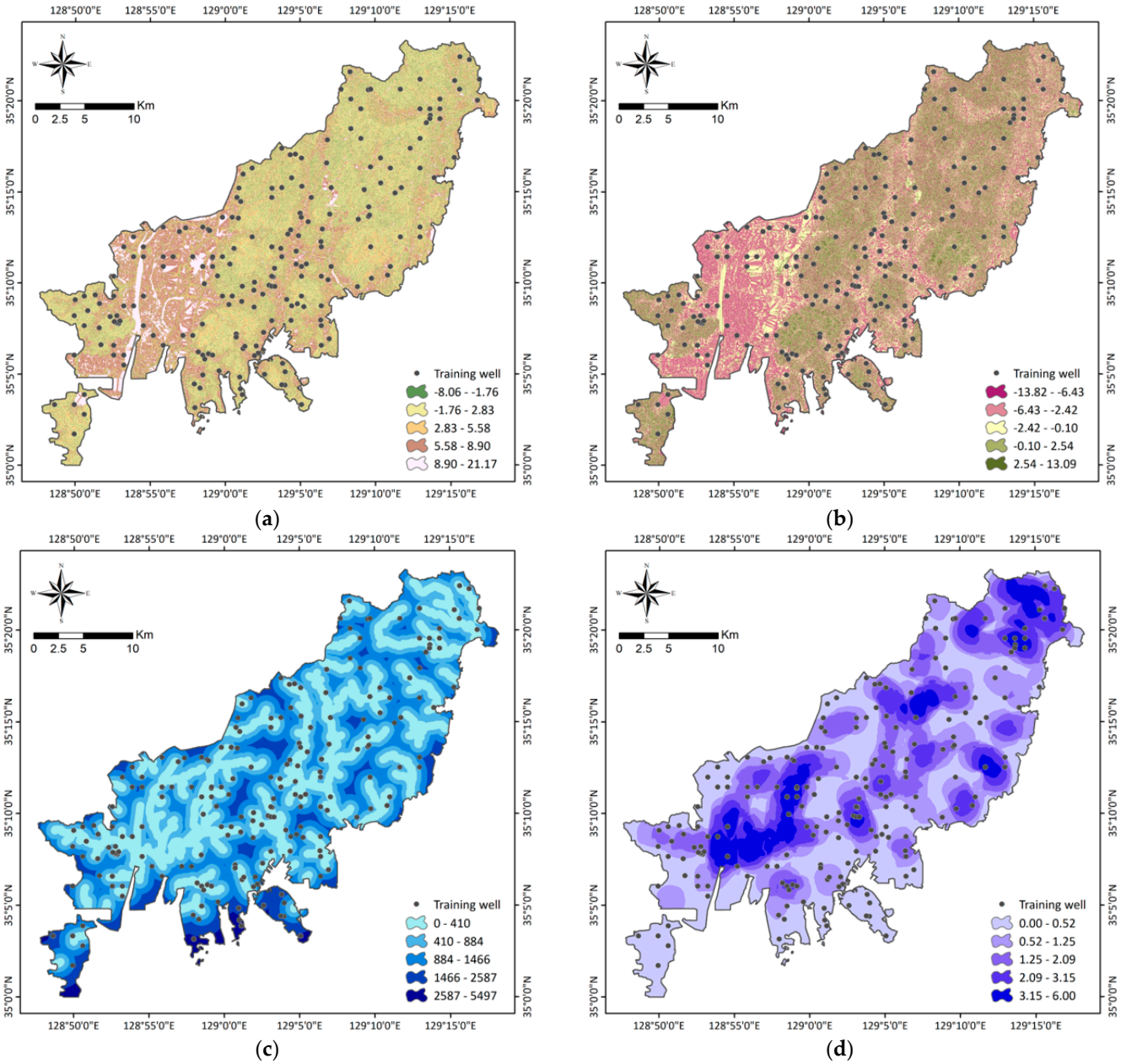


Figure 4. Hydrological conditioning factors: (**a**) topographic wetness index, (**b**) stream power index, (**c**) distance from drainage in meters, and (**d**) drainage density in km^2/km.

For the geological factors of this study, lithology, distance from lineament, lineament density, distance from fault, and fault density were considered (Figure 5). These factors were acquired from the geologic map (1:50,000) of the Korea Institute of Geoscience and Mineral Resources. The lithology of the study area was classified into igneous rock, alluvial rock A, alluvial rock B, and metamorphic rock. In this case, alluvial rocks were further categorized according to permeability; alluvial rock A refers to permeable rocks composed of sandstone and gravel, and alluvial rock B refers to non-permeable rocks composed of shale and clay. The geomatica 2016 software (PCI Geomatics, Markham, ON, Canada) was used to extract the lineament by utilizing a hill-shade map that was generated from

the DEM. The hill-shade map was generated by rendering the images considering a solar zenith angle of 45° and solar azimuth angles of 45°, 90°, and 135°.

The land cover was extracted using a 1:25,000 sub-basin land cover map published by the Ministry of Environment (Figure 6). The sub-basin land cover map divides the land cover into 23 categories using the SPOT-5 image. In this study, these items were reclassified into seven categories: urban, agricultural, forest, grassland, wetlands, bare land, and water.

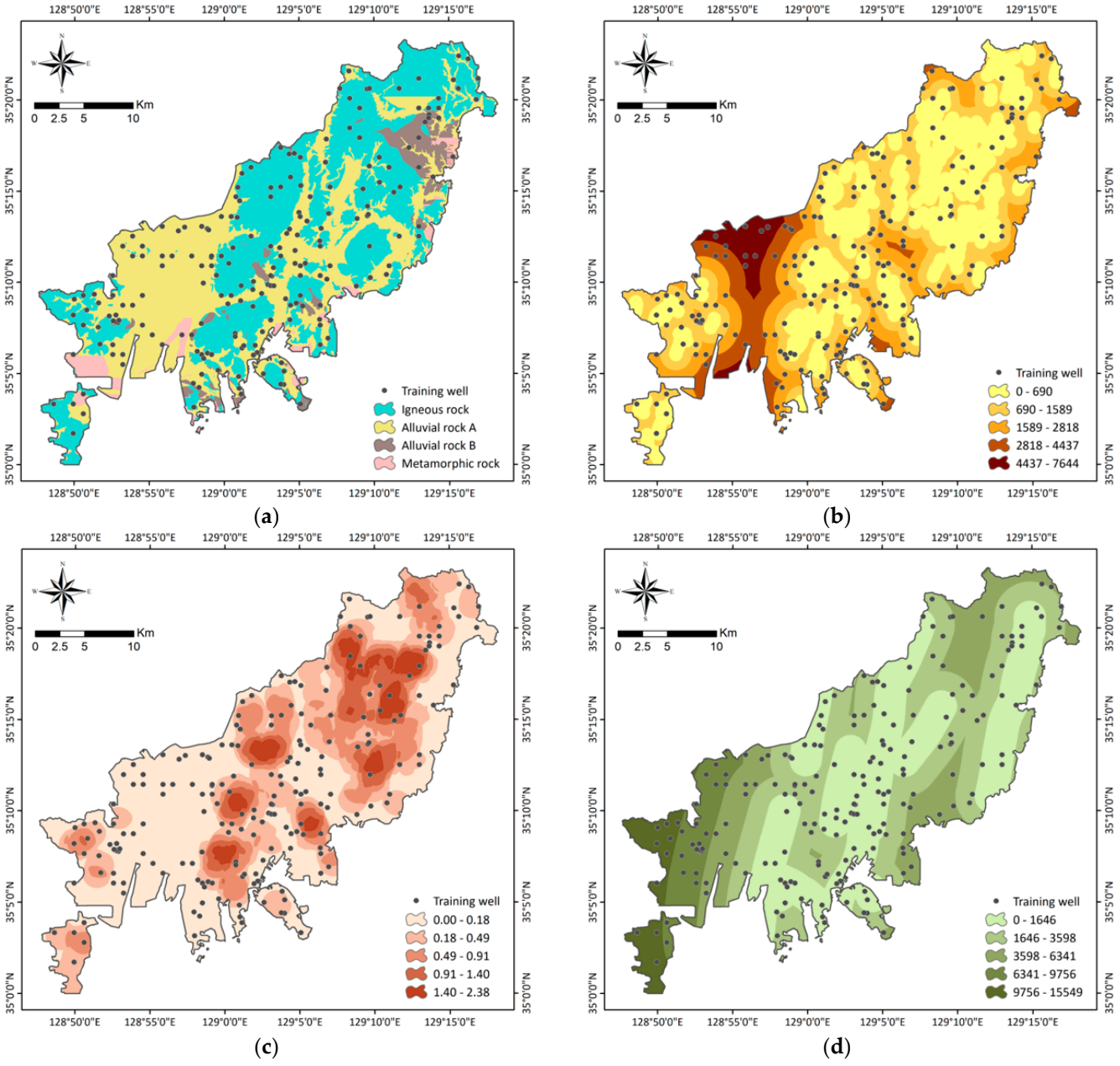
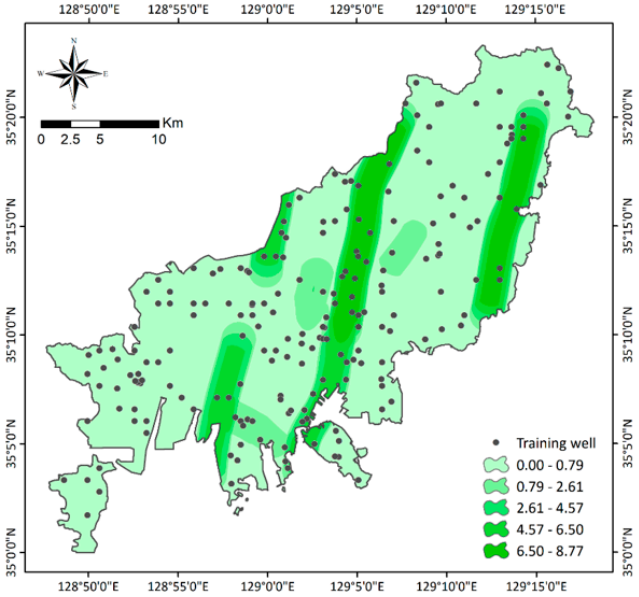


Figure 5. *Cont.*



(**e**)

Figure 5. Geological conditioning factors: (**a**) lithology, (**b**) distance from lineament in meter, (**c**) lineament density in km^2/km, (**d**) distance from faults in meters, and (**e**) fault density in km^2/km.

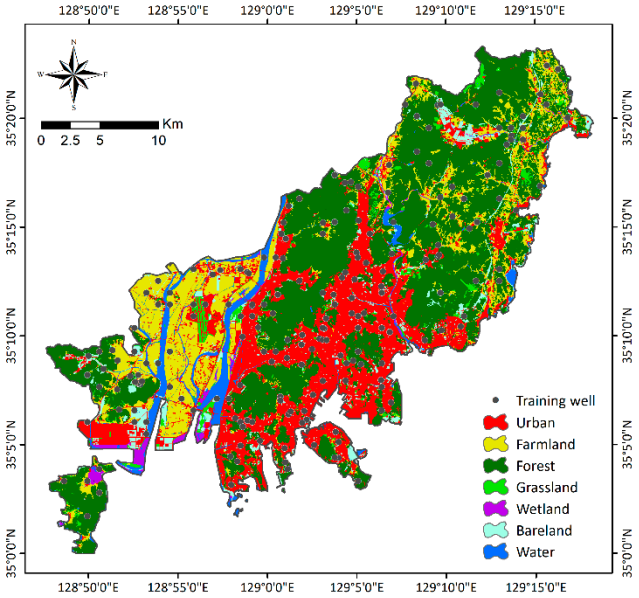


Figure 6. Land cover conditioning factors. The distribution rates of water and urban areas in the study area were 44.65% and 25.44%, respectively.

3.2. Selection of Groundwater Conditioning Factors

As many groundwater conditioning factors were considered in this study, selecting the most appropriate variables by reducing redundancy played a key role in the evaluation of groundwater potential. Removing noise in this way improves the model accuracy, computation speed, and model analysis capability. Therefore, in this study, feature selection was performed prior to model analysis.

In the present study, irrelevant and insignificant variables were selected using an elastic net (Enet). As a regularized regression method, Enet is used to eliminate the ordinary least squares regression method's limitations. In this method, a penalty parameter is used as a regularization parameter, which represents the bias added to the regression coefficient in the equation [44]. The linear regression model that was used in the L1 regularization

method is called the least absolute shrinkage and selection operator (LASSO), while the model that used L2 is called a ridge operator [45].

Enet provides a balance between the two, because it uses a combination of the L1 and L2 regularization methods. The L1 regularization method aims to reduce the number of regression coefficients to zero to generate a sparse model. In comparison, the L2 regularization method has no constraint in the number of selected variables and supports the grouping effect; this stabilizes the L1 regularization path [46,47]. Therefore, Enet regression is prominent in the selection of variables due to its flexibility, variable group selection feature, and variables being selected based on their correlation and predictive capability [48].

3.3. Groundwater Potential Modeling

3.3.1. Random Forest

Ensemble is a method of generating multiple prediction models and combining them to create a final prediction model. RF is a typical example of the ensemble method, which is a tree-based ensemble algorithm that uses the concept of bagging. This unique learning and prediction algorithm is popular as it facilitates the process using random data recursively at each node of the tree and utilizes an error minimization method [49].

RF generates multiple bootstrapped samples using a given training dataset, and then, uses them to create a prediction model by combining them. In this case, bootstrapping is a type of resampling, where a large number of smaller samples of the same size is repeatedly drawn with replacement, thereby ensuring the reliability of the prediction model and contributing to the improvement of prediction performance. Notably, the data not extracted are referred to as "out-of-bag (OOB)" and are used to estimate the prediction error and evaluate the variable importance [50]. The class membership and design of the model (output) are decided by a majority voting process among all trees [51].

3.3.2. Gradient Boost Machine

GBM, also known as gradient-boosted regression tree or gradient tree bosting, is a boosting ensemble algorithm. As in the case of bagging, the boosting algorithm also generates multiple prediction models. In contrast to bagging, the boosting method sequentially generates multiple decision trees using information from previously grown trees. More specifically, in the gradient boosting process applied in this method, a gradient descent algorithm that minimizes the loss for the entire ensemble is used to fit each tree to the residuals of the previous model [26]. The final objective of the method is to minimize the model bias by making the weak learner (each tree) focus on the "harder" samples [52].

3.3.3. Extreme Gradient Boosting

XGB is a boosting ensemble algorithm that can be regarded as an improved GBM. The XGB model integrates several weak learners (each tree) to develop a strong learner through additive learning [53]. Therefore, both XGB and GBM follow the gradient boosting principle. However, XGB improves the training process and prevents overfitting. To this end, XGB implements second-order derivatives to minimize the loss function and obtain more accurate trees, whereas ordinary GBM uses first-order derivatives [25,30]. In XGB, parallel computation is automatically implemented during training to enhance computational efficiency [53]. In addition, it incorporates various regularization features to prevent overfitting. The final prediction of XGB is the sum of the weighted contributions of all decision trees used [25].

4. Results

4.1. Feature Selection

Table 2 shows the results of applying the Enet algorithm to the fourteen groundwater conditioning factors used in this study using the "glmnet" package. Enet has two tuning parameters, λ and α, which represent the overall strength of the penalty and the balance

between the L1 (LASSO) and L2 (ridge) penalties, respectively [44]. The optimized values of λ and α were 0.05043 and 0.5, respectively, and were obtained using 10-fold cross-validation.

Table 2. Regression coefficients of conditioning factors in this study using Enet.

	Regression Coefficient
(Intercept)	0.81012
Elevation	−0.00009
Slope degree	−0.00631
Slope aspect (North)	–
Slope aspect (Northeast)	0.01545
Slope aspect (East)	–
Slope aspect (Southeast)	−0.01922
Slope aspect (South)	−0.06565
Slope aspect (Southwest)	–
Slope aspect (West)	–
Slope aspect (Northwest)	–
Curvature (Flat)	–
Curvature (Convex)	–
Topographic wetness index	−0.00095
Stream power index	–
Distance from drainage	–
Drainage density	−0.03666
Lithology (Alluvial rock A)	0.06866
Lithology (Alluvial rock B)	–
Lithology (Metamorphic rock)	–
Distance from lineament	−0.00004
Lineament density	–
Distance from fault	−0.00002
Fault density	–
Land cover (Farmland)	−0.26278
Land cover (Forest)	−0.22511
Land cover (Grassland)	−0.03889
Land cover (Wetland)	−0.36148
Land cover (Bare land)	–
Land cover (Water)	−0.47592

The results represent the importance (contribution) of the conditioning factors in predicting groundwater potential. If the conditioning factor does not contribute to the groundwater potential, the regression coefficient is left uncalculated. As a result, five conditioning factors with no values (curvature, SPI, distance from drainage, lineament density, and fault density) were excluded from the further analysis of groundwater potential because these conditioning factors may negatively affect the accuracy of the model.

4.2. Groundwater Potential Mapping

The RF, GBM, and XGB models were trained using training datasets to design groundwater potential models. To improve predictivity, each model should be optimized using several hyperparameters. The parameters of each model were optimized using a grid search algorithm based on the "caret" package in the training process. In the present study, GPMs were obtained using RF, GBM, and XGB models based on these optimized hyperparameters.

Finally, the GPMs were reclassified into five possible groups using the natural breaks classification method with characteristics of natural grouping inherent to the data. All of the GPMs had a different range of groundwater potential index (GPI) values. In addition, there is no specific standard to set a threshold for potential classes. Thus, we used the natural breaks classification method that automatically determines thresholds according to the principle of reducing the variance within the classes and maximize the variance between classes [54].

4.2.1. Random Forest

In the RF model, the "randomForest" package was used to obtain the groundwater potential model. The parameter of *ntree* denotes the number of trees in the forest, whereas the parameter of *mtry* denotes the number of variables tested at each node. The parameters of *ntree* and *mtry* were optimized at the values of 370 and 3, respectively.

According to the results of the RF model, the GPI values ranged between 0.03–0.97. To produce the GPM, these values were reclassified into the following five categories, based on their potential: very low (0.03–0.18), low (0.18–0.36), moderate (0.36–0.56), high (0.56–0.76), and very high (0.76–0.97). The area percentages of the classes were found to be 24.69%, 26.95%, 20.73%, 15.88%, and 11.76%, respectively (Figure 7a, Table 3).

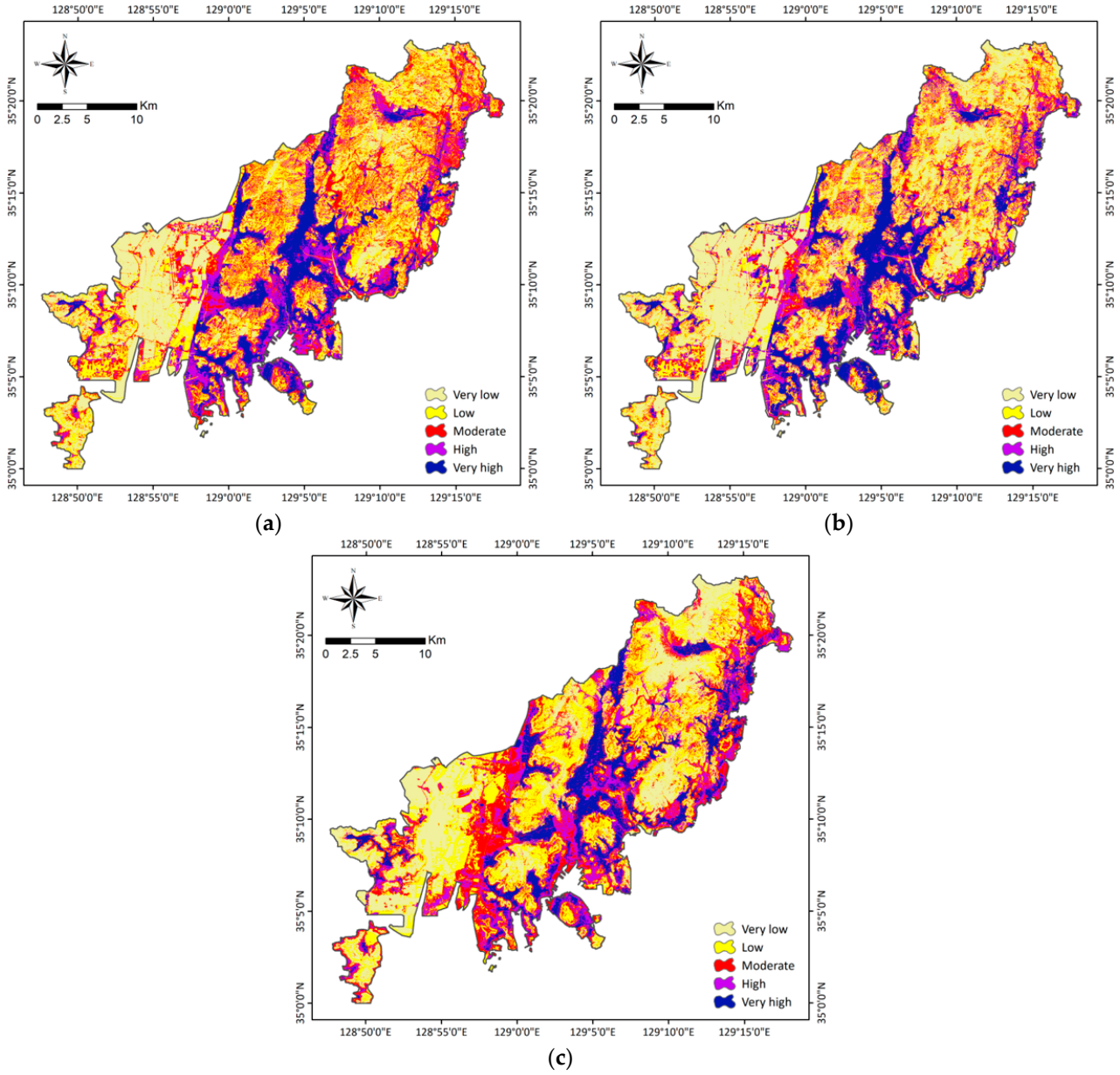


Figure 7. Groundwater potential maps (GPMs) produced from three models: (**a**) random forest, (**b**) gradient boosting machine, and (**c**) extreme gradient boosting.

Table 3. Potential ranges and area distributions of each potential class of the groundwater potential maps (GPMs).

Class/Model	RF		GBM		XGB	
	Range [1]	Area (%) [2]	Range	Area (%)	Range	Area (%)
Very low	0.005–0.215	24.67	0.026–0.182	34.66	0.313–0.398	25.96
Low	0.215–0.371	26.95	0.182–0.359	21.04	0.398–0.456	26.32
Moderate	0.371–0.542	20.73	0.359–0.559	14.76	0.456–0.520	20.75
High	0.542–0.729	15.88	0.559–0.762	14.15	0.520–0.595	14.74
Very high	0.729–0.997	11.76	0.762–0.969	15.39	0.595–0.702	12.23

[1] Class range of each groundwater potential index produced by the respective model; [2] Area ratio of each class to the total area of the study area.

4.2.2. Gradient Boosting Machine

In the GBM model, a priori parameters of *n.minobsinnode*, *interaction.depth*, *bag.fraction*, and *shrinkage* were optimized utilizing the "caret" package. Then, the groundwater model was obtained using the "gbm" package. In this model, *n.minobsinnode* denotes the minimum number of observations in a terminal node, *interaction.depth* denotes the maximum depth of a tree, *bag.fraction* denotes the subsampling fraction, and *shrinkage* denotes the learning rate. The optimization values for the parameters of *n.minobsinnode*, *interaction.depth*, *bag.fraction*, and *shrinkage* were found to be 8, 8, 0.9, and 0.01, respectively.

According to the results of this model, the GPM was generated by reclassifying the areas into the same five potential groups used to classify the RF model. The areas of very low potential comprised 34.66%, whereas areas of low, moderate, high, and very high potential comprised 21.04%, 14.76%, 14.15%, and 15.39%, respectively. The area percentage of the very high potential class in the GBM GPM was found to be higher than that in the GPMs generated by the RF and XGB models (Figure 7b, Table 3).

4.2.3. Extreme Gradient Boosting

The XGB model optimizes the greatest number of a priori parameters (six) compared to other models. The XGB model optimized the parameters *eta* (learning rate) = 0.01, *max_depth* (maximum depth of a tree) = 8, *min_child_weight* (minimum sum of instance weight) = 1, *subsample* (subsample ratio of the training instance) = 0.7, *colsample_bytree* (subsample ratio of columns) = 0.6, and *gamma* = 0.3, using the "caret" package; the groundwater model was generated using the "XGBoost" package.

The results of this model indicate that the GPI ranged between 0.31–0.70. As for the other models, these values were reclassified into five groups based on their potential levels: very low (0.31–0.39), low (0.39–0.46), moderate (0.46–0.52), high (0.52–0.60), and very high (0.60–0.70). The distribution of GPI values for each potential class showed similarity with that of the GPM generated using the RF model. While the area percentages of the low (26.32%) and high (14.74%) potential classes were lower than those calculated in the GPM generated by RF, the area percentages of other classes were found to be slightly higher (Figure 7c, Table 3).

4.3. Model Validation and Comparison

The performance and predictive capacity were analyzed using the ROC curve. The ROC curve was plotted using the true-positive rate (sensitivity) and false-positive rate (1–specificity). The area under the ROC curve (AUROC) was found to range between 0.5 and 1.0. A model's classification accuracy is considered to be high when the AUROC is >0.8. The AUROC values were analyzed for each GPM to determine their success rates; all were found to be > 0.966. According to these results, all GPMs fit the data very well (Table 4, Figure 8a). Moreover, the AUROC values of the predictivity were analyzed using the validation dataset. According to the results, the XGB model provided the highest value (0.818), followed by the GBM model (0.802,) and the RF model (0.794) (Table 4, Figure 8b). These results reveal that the XGB model was the best groundwater prediction model among the three models.

Table 4. Parameters of the receiver operating characteristic (ROC) curve with the calibration and validation dataset.

		AUROC	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
					Lower Bound	Upper Bound
Calibration dataset	RF	1.000	0.000	0.000	1.000	1.000
	GBM	0.998	0.001	0.000	0.995	1.000
	XGB	0.966	0.010	0.000	0.946	0.985
Validation dataset	RF	0.794	0.049	0.000	0.698	0.891
	GBM	0.802	0.048	0.000	0.708	0.896
	XGB	0.818	0.045	0.000	0.730	0.906

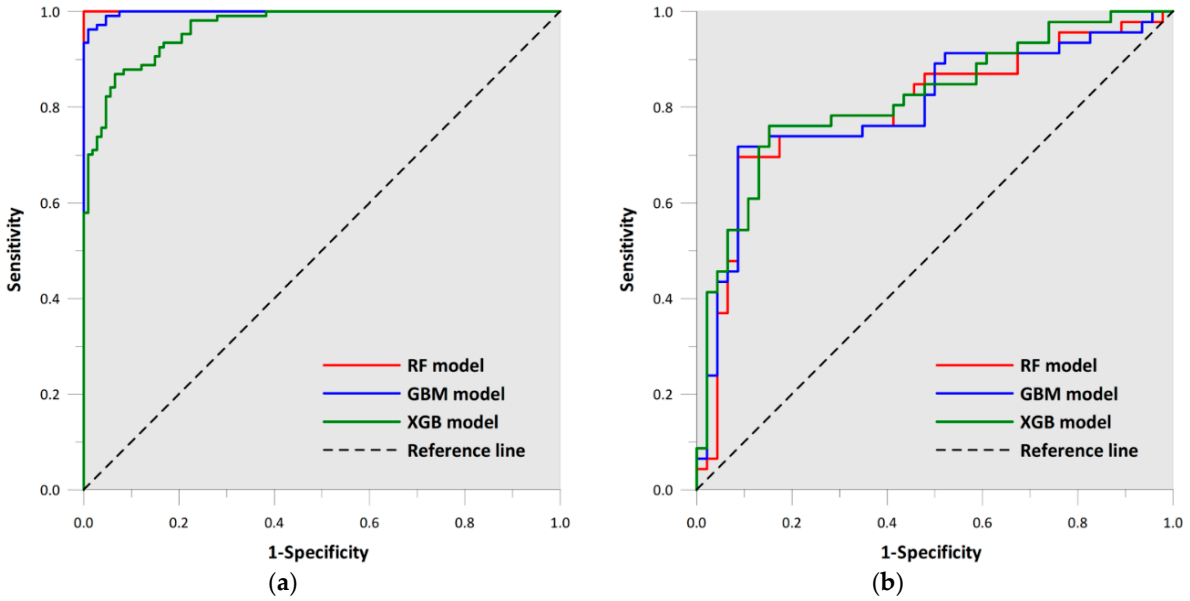


Figure 8. Receiver operating characteristic (ROC) curves for GPMs produced from three models: (**a**) success rate curve using the training dataset, and (**b**) prediction rate curve using the validation dataset.

In addition, the GPMs produced from the three models were validated based on SCAI. SCAI is the ratio of the percentage of groundwater pixels to the percentage of all pixels for each potential class of the GPMs. Generally, SCAI values gradually decrease from very low to very high [55,56], similar to the results of this study (Table 5). At the "very high" class, the SCAI values of the RF, GBM, and XGB models were 0.09, 0.12, and 0.12, respectively. Overall, the models used in this study were suitable for modeling groundwater potential.

Table 5. Seed cell area index values for the RF, GBM, and extreme gradient boosting (XGB) models.

		% of Pixels	Training Datasets		Validation Datasets		Sum	SCAI
			No. of Wells	% of Wells	No. of Wells	% of Wells		
RF	Very low	24.67	0	0.00	4	8.70	8.70	2.84
	Low	26.95	0	0.00	7	15.22	15.22	1.77
	Medium	20.73	0	0.00	3	6.52	6.52	3.18
	High	15.88	8	7.48	15	32.61	40.09	0.40
	Very high	11.76	99	92.52	17	36.96	129.48	0.09
GBM	Very low	34.66	0	0.00	4	8.70	8.70	3.99
	Low	21.04	0	0.00	7	15.22	15.22	1.38
	Medium	14.76	4	3.74	2	4.35	8.09	1.82
	High	14.15	20	18.69	12	26.09	44.78	0.32
	Very high	15.39	83	77.57	21	45.65	123.22	0.12
XGB	Very low	25.96	0	0.00	3	6.52	6.52	3.98
	Low	26.32	2	1.87	6	13.04	14.91	1.76
	Medium	20.75	17	15.89	4	8.70	24.58	0.84
	High	14.74	21	19.63	15	32.61	52.23	0.28
	Very high	12.23	67	62.62	18	39.13	101.75	0.12

5. Discussion

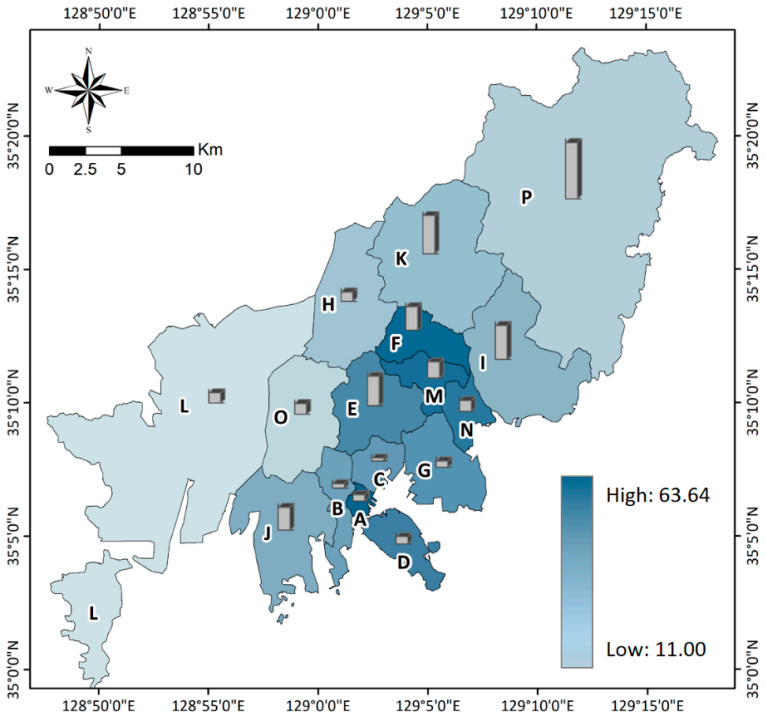
This study was aimed to produce GPMs using the RF, GBM, and XGB models and compare prediction performances of groundwater potential. The main concern as to test the robustness und accuracy of the XGB model. The AUROC values of the all success rate curves were relatively high, almost reaching a value of 1. Contrastingly, the AUROC values of the prediction rate curves decreased by approximately 20%. In particular, the prediction performance of the XGB model was superior to that of the RF and GBM models. In addition, the difference between the success rate and prediction rate curves of the GPM generated by the XGB model was found to be 0.148. This was the lowest value calculated among the three models; the GBM model performed second best (0.196), followed by the RF model (0.206).

These results present the same trend as the findings of previous studies. As mentioned in introduction, the XGB model has been applied to various fields, but hardly to groundwater research. Recent studies used the XGB model to predict groundwater level [57,58], analyze groundwater salinity [59], and assess groundwater quality [60]. In these studies, the XGB model had more accurate results compared to ANN, SVM, and multiple linear regression methods. Especially, Naghibi et al. [61] assessed the groundwater spring potential using RF, parallel random forest, and XGB models. All models yielded AUROC values of approximately 86%, with the XGB model showing the highest values.

The superior results are likely due to the advantages of the XGB model. The XGB model, which improves the loss function by Taylor expansion, provides gradient convergence quickly and more accurately as compared to other methods. This model can handle missing/sparse data and contribute to a higher speed/accuracy using cross-validation, early stop, and parallel processes [31,39,62]. In addition, this model effectively avoids overfitting using many strategies, such as the normalization of the objective function, shrinkage, and column subsampling [39]. Through these advantages, the XGB model ensures excellent prediction performance, applying not only to classification, but also to linear regressions.

The GPMs proposed in this study can be used to establish a sustainable groundwater usage and management policy for the study area. For example, we compared the groundwater potential areas and annual groundwater usage in each administrative district (Figure 9). The groundwater potential areas were calculated via the ratio of the high and very high potential classes from the GPM generated by the XGB model to each administrative district's total area. The annual groundwater usage, as of June 2020, was obtained from the Public Data Portal in Korea [63] and calculated as the percentage of the entire study area.

Among the administrative districts, the annual groundwater usage in Haeundae, Geumjeong, and Gijang accounted for more than 10% of the total groundwater usage in the study area. However, the potential area ratio was relatively low in these districts as compared to others. Therefore, these regions need efficient policy for a sustainable groundwater use. The GPM, as reference data, could help to establish such policy and support decision-making processes. In the future, the accuracy of the GPM can be further improved by examining a wider range of groundwater conditioning factors for the study area, such as hydraulic and hydrological factors, topography, pedological factors, and precipitation.



(**a**)

Code	Name	Area ratio	Usage rate
A	Jung	63.64	2.13
B	Seo	34.29	1.47
C	Dong	37.79	1.00
D	Yeongdo	47.62	2.52
E	Busanjin	42.88	10.25
F	Dongnae	61.39	8.13
G	Nam	41.50	2.22
H	Buk	30.82	3.09
I	**Haeundae**	**33.25**	**11.84**
J	Saha	34.25	7.95
K	**Geumjeong**	**32.31**	**13.31**
L	Gangseo	11.00	3.48
M	Yeonje	60.32	5.54
N	Suyeong	54.51	3.78
O	Sasang	21.79	3.64
P	**Gijang**	**23.71**	**19.64**

(**b**)

Figure 9. Area ratio and usage rate of each administrative district: (**a**) distribution map (graduated colors indicate the area ratio, and bar charts indicate the usage rate), and (**b**) specific data (area ratio indicates the area ratio of the high and very high potential classes to the total area of each administrative district; usage rate indicates the percentage of annual groundwater usage (m^3/year) to that in the total study area.

6. Conclusions

In this study, GPMs for Busan Metropolitan City, South Korea, were created using RF, GBM, and EGB models; the results were comparatively analyzed using AUROC and SCAI, for which a spatial database of groundwater wells and groundwater conditioning factors was constructed. Fourteen groundwater conditioning factors were obtained from thematic maps obtained from the government and related organizations. These factors were evaluated for their contributions to groundwater potential using Enet. Finally, nine groundwater conditioning factors were selected, including altitude, slope degree, slope aspect, TWI, drainage density, lithology, distance from lineament, distance from fault, and land cover. Groundwater potential analyses and mapping were performed with these nine groundwater conditioning factors using the RF, GBM, and EGB models. The GPMs produced by RF, GBM, and EGB were evaluated using AUROC and SCAI. Overall, the three GPMs exhibited good performance with the used training and validation datasets. These results indicate that all the studied models were successful in creating GPMs for the study area. Since the EGB model outperformed the RF and GBM models, it can improve predictive performance with its powerful capability. It is suggested that the GPM generated by this model in the present study can be used efficiently and cost-effectively to investigate groundwater resources in Busan Metropolitan City. It can further be used to prepare groundwater management plans for a sustainable use of groundwater resources.

Author Contributions: S.P. wrote the paper and analyzed the data; J.K. suggested the idea for the study. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education and the Ministry of Science and ICT (2019M3E7A1113103, 2020R1A2B5B02002198, 2020R1I1A1A01075106).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education and the Ministry of Science and ICT (2019M3E7A1113103, 2020R1A2B5B02002198, 2020R1I1A1A01075106).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Panahi, M.; Sadhasivam, N.; Pourghasemi, H.R.; Rezaie, F.; Lee, S. Spatial prediction of groundwater potential mapping based on convolutional neural network (CNN) and support vector regression (SVR). *J. Hydrol.* **2020**, *588*, 125033. [CrossRef]
2. Chen, W.; Li, H.; Hou, E.; Wang, S.; Wang, G.; Panahi, M.; Li, T.; Peng, T.; Guo, C.; Niu, C.; et al. GIS-based groundwater potential analysis using novel ensemble weights-of-evidence with logistic regression and functional tree models. *Sci. Total Environ.* **2018**, *634*, 853–867. [CrossRef] [PubMed]
3. Ozdemir, A. 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.* **2011**, *411*, 290–308. [CrossRef]
4. Pourtaghi, Z.S.; Pourghasemi, H.R. GIS-based groundwater spring potential assessment and mapping in the Birjand Township, southern Khorasan Province, Iran. *Hydrogeol. J.* **2014**, *22*, 643–662. [CrossRef]
5. Mohammady, M.; Pourghasemi, H.R.; Pradhan, B. Landslide susceptibility mapping at Golestan Province, Iran: A comparison between frequency ratio, Dempster-Shafer, and weights-of-evidence models. *J. Asian Earth Sci.* **2012**, *61*, 221–236. [CrossRef]
6. Pradhan, B.; Lee, S. Delineation of landslide hazard areas on Penang Island, Malaysia, by using frequency ratio, logistic regression, and artificial neural network models. *Environ. Earth Sci.* **2010**, *60*, 1037–1054. [CrossRef]
7. Al-Abadi, A.M. Modeling of groundwater productivity in northeastern Wasit Governorate, Iraq using frequency ratio and Shannon's entropy Models. *Appl. Water Sci.* **2017**, *7*, 699–716. [CrossRef]
8. Liu, J.; Duan, Z. Quantitative assessment of landslide susceptibility comparing statistical index, index of entropy, and weights of evidence in the Shangnan area, China. *Entropy* **2018**, *20*, 868. [CrossRef]
9. Nampak, H.; Pradhan, B.; Abd Manap, M. Application of GIS based data driven evidential belief function model to predict groundwater potential zonation. *J. Hydro.* **2014**, *513*, 283–300. [CrossRef]
10. Rahmati, O.; Tahmasebipour, N.; Haghizadeh, A.; Haghizadeh, H.; Pourghasemi, H.R.; Feizizadeh, B. Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion. *Geomorphology* **2017**, *298*, 118–137. [CrossRef]
11. Chen, W.; Tsangaratos, P.; Ilia, I.; Duan, Z.; Chen, X. Groundwater spring potential mapping using population-based evolutionary algorithms and data mining methods. *Sci. Total Environ.* **2019**, *684*, 31–49. [CrossRef] [PubMed]
12. Kalantar, B.; Al-Najjar, H.A.; Pradhan, B.; Saeidi, V.; Halin, A.A.; Ueda, N.; Naghibi, S.A. Optimized conditioning factors using machine learning techniques for groundwater potential mapping. *Water* **2019**, *11*, 1909. [CrossRef]
13. Naghibi, S.A.; Pourghasemi, H.R.; Abbaspour, K. A comparison between ten advanced and soft computing models for groundwater qanat potential assessment in Iran using R and GIS. *Theor. Appl. Climatol.* **2018**, *131*, 967–984. [CrossRef]
14. Rahmati, O.; Moghaddam, D.D.; Moosavi, V.; Kalantari, Z.; Samadi, M.; Lee, S.; Tien Bui, D. An automated python language-based tool for creating absence samples in groundwater potential mapping. *Remote Sens.* **2019**, *11*, 1375. [CrossRef]
15. Al-Fugara, A.; Pourghasemi, H.R.; Al-Shabeeb, A.R.; Habib, M.; Al-Adamat, R.; Al-Amoush, H.; Collins, A.L. A comparison of machine learning models for the mapping of groundwater spring potential. *Environ. Earth Sci.* **2020**, *79*, 1–19. [CrossRef]
16. Golkarian, A.; Naghibi, S.A.; Kalantar, B.; Pradhan, B. Groundwater potential mapping using C5. 0, random forest, and multivariate adaptive regression spline models in GIS. *Environ. Monit. Assess* **2018**, *190*, 149. [CrossRef]
17. Chen, W.; Li, Y.; Tsangaratos, P.; Shahabi, H.; Ilia, I.; Xue, W.; Bian, H. Groundwater spring potential mapping using artificial intelligence approach based on kernel logistic regression, random forest, and alternating decision tree models. *Appl. Sci. Basel* **2020**, *10*, 425. [CrossRef]
18. Park, S.; Hamm, S.Y.; Kim, J. Performance evaluation of the GIS-based data-mining techniques decision tree, random forest, and rotation forest for landslide susceptibility modeling. *Sustainability* **2019**, *11*, 5659. [CrossRef]
19. Zhang, W.; Zhang, R.; Wu, C.; Goh, A.T.C.; Lacasse, S.; Liu, Z.; Liu, H. State-of-the-art review of soft computing applications in underground excavations. *Geosci. Front.* **2020**, *11*, 1095–1106. [CrossRef]
20. Brédy, J.; Gallichand, J.; Celicourt, P.; Gumiere, S.J. Water table depth forecasting in cranberry fields using two decision-tree-modeling approaches. *Agric. Water Manag.* **2020**, *233*, 106090. [CrossRef]
21. Tziachris, P.; Aschonitis, V.; Chatzistathis, T.; Papadopoulou, M. Assessment of spatial hybrid methods for predicting soil organic matter using DEM derivatives and soil parameters. *Catena* **2019**, *174*, 206–216. [CrossRef]
22. Naghibi, S.A.; Pourghasemi, H.R.; Dixon, B. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environ. Monit. Assess* **2016**, *188*, 44. [CrossRef] [PubMed]
23. Mosavi, A.; Hosseini, F.S.; Choubin, B.; Goodarzi, M.; Dineva, A.A.; Sardooi, E.R. Ensemble Boosting and Bagging Based Machine Learning Models for Groundwater Potential Prediction. *Water Resour. Manag.* **2020**, *35*, 23–37. [CrossRef]

24. Ouedraogo, I.; Defourny, P.; Vanclooster, M. Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale. *Hydrogeol. J.* **2019**, *27*, 1081–1098. [CrossRef]
25. Dietterich, T. Overfitting and undercomputing in machine learning. *ACM Comput. Surv. (CSUR)* **1995**, *27*, 326–327. [CrossRef]
26. Zanotti, C.; Rotiroti, M.; Sterlacchini, S.; Cappellini, G.; Fumagalli, L.; Stefania, G.A.; Nanucci, M.S.; Leoni, B.; Bonomi, T. Choosing between linear and nonlinear models and avoiding overfitting for short and long term groundwater level forecasting in a linear system. *J. Hydrol.* **2019**, *578*, 124015. [CrossRef]
27. Martínez-Santos, P.; Renard, P. Mapping groundwater potential through an ensemble of big data methods. *Groundwater* **2020**, *58*, 583–597. [CrossRef]
28. Jabbar, H.; Khan, R.Z. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). *Comput. Sci. Commun. Instrum. Devices* **2015**, 163–172. [CrossRef]
29. Cai, Z.; Jiang, B.; Lu, Z.; Liu, J.; Ma, P. isAnon: Flow-Based Anonymity Network Traffic Identification Using Extreme Gradient Boosting. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; IEEE: New York City, NY, USA, 2019; pp. 1–8.
30. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794.
31. Chang, Y.C.; Chang, K.H.; Wu, G.J. Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Appl. Soft. Comput.* **2018**, *73*, 914–920. [CrossRef]
32. Sahin, E.K. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using EGBoost, gradient boosting machine, and random forest. *SN Appl. Sci.* **2020**, *2*, 1–17. [CrossRef]
33. Jamali, A. Landslide hazard risk modeling in north-west of Iran using optimized machine learning models. *Model Earth Syst. Environ.* **2021**, *7*(1), 191–208. [CrossRef]
34. Chen, Z.Y.; Zhang, T.H.; Zhang, R.; Zhu, Z.M.; Yang, J.; Chen, P.Y.; Ou, C.Q.; Guo, Y. Extreme gradient boosting model to estimate PM2. 5 concentrations with missing-filled satellite data in China. *Atmos. Environ.* **2019**, *202*, 180–189. [CrossRef]
35. Gui, K.; Che, H.; Zeng, Z.; Wang, Y.; Zhai, S.; Wang, Z.; Luo, M.; Zhang, L.; Liao, T.; Zhao, H.; et al. Construction of a virtual PM2.5 observation network in China based on high-density surface meteorological observations using the Extreme Gradient Boosting model. *Environ. Int.* **2020**, *141*, 105801. [CrossRef] [PubMed]
36. Hamedianfar, A.; Gibril, M.B.A.; Hosseinpoor, M.; Pellikka, P.K. Synergistic use of particle swarm optimization, artificial neural network, and extreme gradient boosting algorithms for urban LULC mapping from WorldView-3 images. *Geocarto Int.* **2020**. [CrossRef]
37. Georganos, S.; Grippa, T.; Vanhuysse, S.; Lennert, M.; Shimoni, M.; Wolff, E. Very high resolution object-based land use-land cover urban classification using extreme gradient boosting. *IEEE Geosci. Remote S.* **2018**, *15*, 607–611. [CrossRef]
38. Zhang, W.; Wu, C.; Zhong, H.; Li, Y.; Wang, L. Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geosci. Front.* **2021**, *12*, 469–477. [CrossRef]
39. Chen, Z.; Li, H.; Goh, A.T.C.; Wu, C.; Zhang, W. Soil liquefaction assessment using soft computing approaches based on capacity energy concept. *Geosciences* **2020**, *10*, 1–19. [CrossRef]
40. KMA Data Open Portal. Available online: https://kma.go.kr (accessed on 10 December 2020).
41. Oh, H.J. Landslide susceptibility analysis and validation using Weight-of-Evidence model. *J. Geol. Soc. Korea* **2010**, *46*, 157–170.
42. Moore, I.D.; Grayson, R.B.; Ladson, A.R. Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. *Hydrol. Process* **1991**, *5*, 3–30. [CrossRef]
43. Pradhan, B. Groundwater potential zonation for basaltic watersheds using satellite remote sensing data and GIS techniques. *Open Geosci.* **2009**, *1*, 120–129. [CrossRef]
44. Acharjee, A.; Finkers, R.; Visser, R.G.; Maliepaard, C.J.M. Comparison of regularized regression methods for ~omics data. *Metabolomics* **2013**, *3*, 9. [CrossRef]
45. Adab, H.; Morbidelli, R.; Saltalippi, C.; Moradian, M.; Ghalhari, G.A.F. Machine learning to estimate surface soil moisture from remote sensing data. *Water* **2020**, *12*, 3223. [CrossRef]
46. Park, H.; Konishi, S. Robust logistic regression modelling via the elastic net-type regularization and tuning parameter selection. *J. Stat. Comput. Simul.* **2015**, *86*, 1–12. [CrossRef]
47. Liu, W.; Li, Q. An efficient elastic net with regression coefficients method for variable selection of spectrum data. *PLoS ONE* **2017**, *12*, e0171122. [CrossRef]
48. Giglio, C.; Brown, S.D. Using elastic net regression to perform spectrally relevant variable selection. *J. Chemom.* **2018**, *32*, e3034. [CrossRef]
49. Moghaddam, D.D.; Pourghasemi, H.R.; Rahmati, O. Assessment of the contribution of geo-environmental factors to flood inundation in a semi-arid region of SW Iran: Comparison of different advanced modeling approaches. In *Natural Hazards GIS Based Spatial Modeling Using Data Mining Techniques*; Springer: Cham, Switzerland, 2019; pp. 59–78.
50. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
51. Micheletti, N.; Foresti, L.; Robert, S.; Leuenberger, M.; Pedrazzini, A.; Jaboyedoff, M.; Kanevski, M. Machine learning feature selection methods for landslide susceptibility mapping. *Math. Geosci.* **2014**, *46*, 33–57. [CrossRef]

52. Xiao, T.; Zhu, J.; Liu, T. Bagging and boosting statistical machine translation systems. *Artif. Intell.* **2013**, *195*, 496–527. [CrossRef]
53. Fan, J.; Zheng, J.; Wu, L.; Zhang, F. Estimation of daily maize transpiration using support vector machines, extreme gradient boosting, artificial and deep neural networks models. *Agric. Water Manag* **2020**, *245*, 106547. [CrossRef]
54. Jenks, G.F. The Data Model Concept in Statistical Mapping. *Int. Yearb. Cartogr.* **1967**, *7*, 186–190.
55. Süzen, M.L.; Doyuran, V. A comparison of the GIS based landslide susceptibility assessment methods: Multivariate versus bivariate. *Environ. Geol.* **2004**, *45*, 665–679. [CrossRef]
56. Arabameri, A.; Rezaei, K.; Cerda, A.; Lombardo, L.; Rodrigo-Comino, J. GIS-based groundwater potential mapping in Shahroud plain, Iran. A comparison among statistical (bivariate and multivariate), data mining and MCDM approaches. *Sci. Total Environ.* **2019**, *658*, 160–177. [CrossRef] [PubMed]
57. Osman, A.I.A.; Ahmed, A.N.; Chow, M.F.; Huang, Y.F.; El-Shafie, A. Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. *Ain. Shams. Eng. J.* **2021**. [CrossRef]
58. Rahman, A.S.; Hosono, T.; Quilty, J.M.; Das, J.; Basak, A. Multiscale groundwater level forecasting: Coupling new machine learning approaches with wavelet transforms. *Adv. Water Resour.* **2020**, *141*, 103595. [CrossRef]
59. Sahour, H.; Gholami, V.; Vazifedan, M. A comparative analysis of statistical and machine learning techniques for mapping the spatial distribution of groundwater salinity in a coastal aquifer. *J. Hydrol.* **2020**, *591*, 125321. [CrossRef]
60. Bedi, S.; Samal, A.; Ray, C.; Snow, D. Comparative evaluation of machine learning models for groundwater quality assessment. *Environ. Monit. Assess* **2020**, *192*, 1–23. [CrossRef] [PubMed]
61. Naghibi, S.A.; Hashemi, H.; Berndtsson, R.; Lee, S. Application of extreme gradient boosting and parallel random forest algorithms for assessing groundwater spring potential using DEM-derived factors. *J. Hydrol.* **2020**, *589*, 125197. [CrossRef]
62. Fan, J.; Yue, W.; Wu, L.; Zhang, F.; Cai, H.; Wang, X.; Lu, X.; Xiang, Y. Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China. *Agric. Forest Meteorol.* **2018**, *263*, 225–241. [CrossRef]
63. Public Data Portal. Available online: https://www.data.go.kr (accessed on 28 January 2020).