

Article

Land Subsidence Susceptibility Mapping Using Bayesian, Functional, and Meta-Ensemble Machine Learning Models

Hyun-Joo Oh ¹, Mutiara Syifa ², Chang-Wook Lee ^{2,*}  and Saro Lee ^{3,4,*} 

¹ Geo-Environmental Hazard Research Center, Korea Institute of Geoscience and Mineral Resources (KIGAM), 124, Gwahak-ro Yuseong-gu, Daejeon 34132, Korea; ohj@kigam.re.kr

² Division of Science Education, Kangwon National University, Chuncheon Campus, 1 Gangwondaehakgil, Chuncheon-si, Gangwon-do 24341, Korea; mutiarsyifa@gmail.com

³ Geoscience Platform Research Division, Korea Institute of Geoscience and Mineral Resources (KIGAM), 124, Gwahak-ro Yuseong-gu, Daejeon 34132, Korea

⁴ Department of Geophysical Exploration, Korea University of Science and Technology, 217 Gajeong-ro Yuseong-gu, Daejeon 34113, Korea

* Correspondence: cwlee@kangwon.ac.kr (C.-W.L.); leesaro@kigam.re.kr (S.L.);
Tel.: +82-33-250-6731 (C.-W.L.); +82-42-868-3057 (S.L.)

Received: 25 January 2019; Accepted: 22 March 2019; Published: 25 March 2019



Abstract: To effectively prevent land subsidence over abandoned coal mines, it is necessary to quantitatively identify vulnerable areas. In this study, we evaluated the performance of predictive Bayesian, functional, and meta-ensemble machine learning models in generating land subsidence susceptibility (LSS) maps. All models were trained using half of a land subsidence inventory, and validated using the other half of the dataset. The model performance was evaluated by comparing the area under the receiver operating characteristic (ROC) curve of the resulting LSS map for each model. Among all models tested, the logit boost, which is a meta-ensemble machine learning model, generated LSS maps with the highest accuracy (91.44%), i.e., higher than that of the other Bayesian and functional machine learning models, including the Bayes net (86.42%), naïve Bayes (85.39%), logistic (88.92%), and multilayer perceptron models (86.76%). The LSS maps produced in this study can be used to mitigate subsidence risk for people and important facilities within the study area, and as a foundation for further studies in other regions.

Keywords: land subsidence; Bayes net; naïve Bayes; logistic; multilayer perceptron; logit boost

1. Introduction

Coal mining was once the driving force of the national industry and economic development in Korea, but this situation changed as demand for coal decreased. Gangwon Province was once Korea's largest coal mining area but most of its mines were closed in the early 1990s. Among the environmental problems that follow mine closures, land subsidence events can threaten human life and damage property and infrastructure, including buildings, houses, railroads, and roads [1–4]. Recovery of surface structures following land subsidence is difficult and costly; therefore, it is necessary to predict land subsidence susceptibility (LSS) zones before subsidence occurs, and to implement management strategies in these zones [3].

Generally, prediction of subsidence susceptibility zones requires the input of several environmental factors and the application of prediction models [5]. Several previous studies have developed quantitative and qualitative models that have been successfully applied in various hazard susceptibility zones worldwide [3–11]. These include logistic regression (LR) [3], frequency ratio

(FR) [3,6], weight of evidence (WOE) [3], evidential belief function (EBF) [4], artificial neural network (ANN) [3,5,7,8], support vector machine (SVM) [9], random forest (RF) [10], and fuzzy logic (FL) [8,11] models. Single LSS mapping models can be combined to form ensemble models, which provide more precise and meaningful results [9]. Ensemble models based on machine learning have recently improved the prediction accuracy and performance of single classifiers [12]. The main advantages of this approach are the ability to represent complex relationships between influential factors, and to incorporate spatial data of various scales [13].

Based on existing studies, probability and statistical models using geographical information systems (GIS) have been applied extensively to predict the susceptibility of geohazards, such as landslides, floods, subsidence, and rockfalls [3,14–16]. Recently, data mining and machine learning models for addressing nonlinear problems have been developed, which have been applied frequently and had their performances compared in landslide susceptibility mapping [17–20]. In ground subsidence hazard mapping, ground subsidence hazard maps around abandoned underground coal mines (AUCMs) have been constructed by integrating the adaptive neuro-fuzzy inference system and GIS [21]. In addition, a fuzzy operator, decision tree with the CHAID and QUEST algorithms, and the frequency ratio have been applied to construct subsidence susceptibility maps at AUCMs in Korea [2,11]. In this study, we investigated the performance of some models that have never been applied to land subsidence prediction. Therefore, in this study, we generated LSS maps for a South Korean district containing abandoned subsurface coal mines using machine learning methods, including a logit boost meta-ensemble model, two Bayesian models (Bayes net and NB models) and two functional models (logistic and multilayer perceptron models). The reliability and accuracy of all models were assessed by comparing their area under the receiver operating characteristic (ROC) curves. Data processing was performed using WEKA 3.9.2 and ArcGIS 10.5 software to produce five machine learning algorithms.

2. Land Subsidence in the Study Area

The study area, Hwajeon, is located in the city of Taebaek, South Korea (Figure 1), at $37^{\circ}11'07''$ – $37^{\circ}11'07''$ N, $128^{\circ}56'40''$ – $128^{\circ}57'43''$ E. Underground coal mining activities were carried out in Taebaek for nearly 20 years. The coal seams in this area were irregularly disturbed and inclined with various widths by reverse and thrust faults [22]. Therefore, the slant-chute block caving method was mainly used. About 10 million tons of coal were mined from the study area between 1953–1991 [22], and coal was transported to other areas by railroad beginning in 1973 (Figure 1). Since 1990, most of the coal mines have been closed due to reduced coal demand. However, the abandoned underground coal mines are currently causing land subsidence in the study area [11,21–23]. Additionally, infrastructure has been damaged by the land subsidence, as shown in photographs in a previous report [11].

Subsidence is caused by a variety of contributing factors, including geological discontinuities, presence of water, mining depth, and weak overburden [24,25]. The two forms of subsidence caused by underground coal mining are trough and sinkhole subsidence [25]. In the study area, a very irregular sinkhole occurred due to many complex underground coal mine pits excavated via slant-chute block caving in combination with the aforementioned factors [22]. After a mine cavity is excavated, roof stability becomes unstable over time due to changes in the strength and stress of the roof strata. Under such conditions, additional contributing factors can lead to the occurrence of sinkholes [25]. The Coal Industry Promotion Board [11,26] has reported 24 land subsidence events within the study area. Figure 1 shows a representative land subsidence from location S1 to location S6 of a subsidence event reported in 1999. Table 1 provides a description of the land subsidence. Locations S1 to S5 of this land subsidence mainly occurred along railways and at elevations above 800 m. Location S6 occurred in residential areas and at a lower elevation than S1–S5. Also, the depth of subsidence of S6 is the deepest (508 mm). Some photographs providing evidence of the land subsidence have been published [11,23].

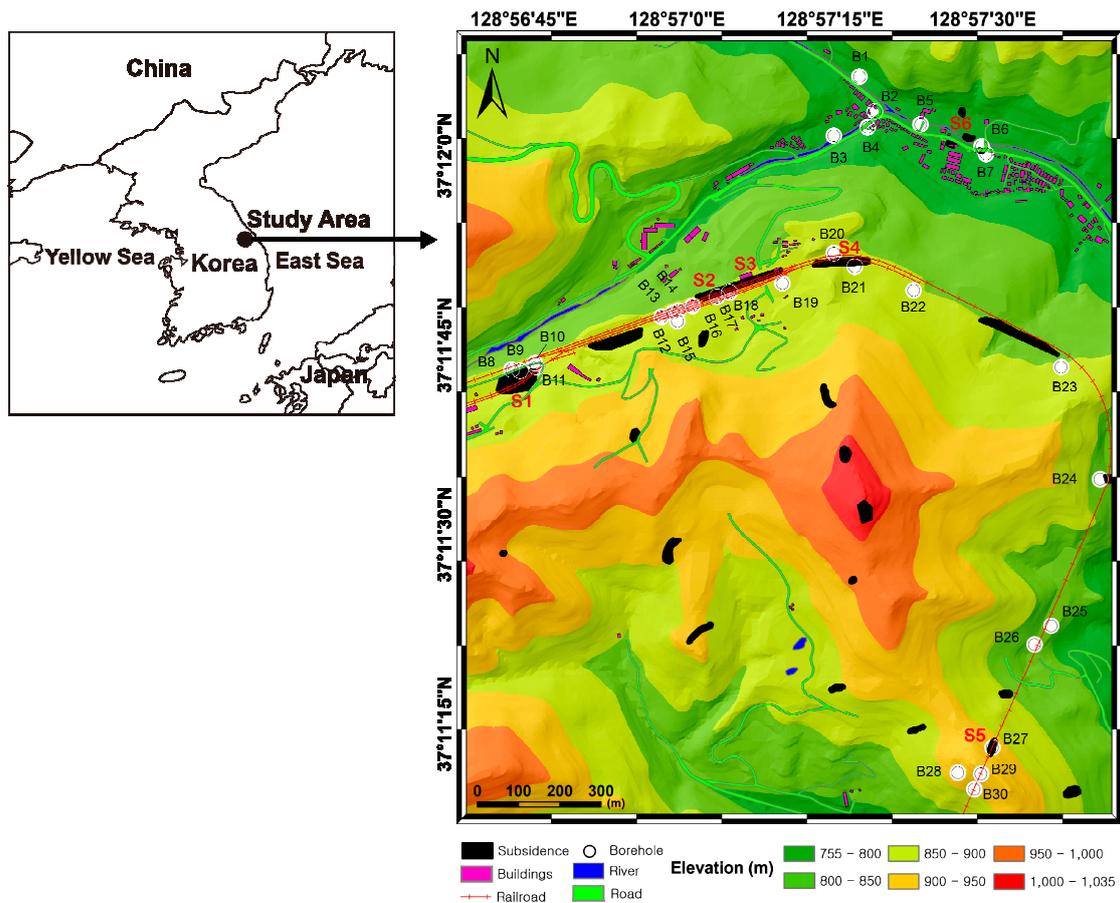


Figure 1. The study area in Taebaek, South Korea.

Table 1. Description of representative land subsidence in the study area [22].

Location	Structure	Elevation (m)	Mining Depth (m)	Thickness (m) and Slope of Coal Seam	Subsidence Depth (mm)	Other
S1	Railway	885	20–30	1–2 40–50°	90	-The coal seam is oblique to the railroad. -Shallow depth of mine -Sinkhole-type subsidence
S2	Railway	885	0	–	72	-Progression of cavity by mining -Subsidence by limestone cavity
S3	Railway	885	30–50	1–2 20°	329	-Subsidence along railway
S4	Railway	885	40–65	2 20°	223	-Shallow depth of mine -Coal bonanza
S5	Tunnel Railway	810	30–260	105 50–70°	65	-The tunnel is located above the mine cavity. -Vertical cracks and leakage in tunnel
S6	Road	765	60–98	3 20°	508	-Residential area and elementary school -Differential subsidence

3. Construction of Spatial Database

It is necessary to determine the factors affecting the land subsidence of a coal mine area. The lithology of the overburden rocks, geological discontinuities, ground slope, scope of the mined cavity, extent and depth of mining, mechanical characteristics of the rock mass rating (RMR), and flow of groundwater are considered the main factor [11,25,27,28]. Spatial data for all of these factors may be difficult to collect and may not be available. The available spatial databases used in this study were constructed using ArcGIS 10.5.

The surface geology with cross section lines was constructed using a digital geological map with 1:50,000 scale [29] published by the Korea Institute of Geoscience and Mineral Resources (KIGAM). The geological formations include the Manhan, Jangseong, Hambaegsan, Dosagog, and Alluvium horizons (Figure 2a, Figure 3). Most of the coal was mined from the Jangseong Formation with a thickness of 80–15 m [22,30]. This formation includes four to five cyclothems consisting of dark-gray sandstone, black shale, and coal seam (Table 2). The land use was constructed from a digital land characteristics map with 1:5,000 scale [31] supplied by the National Geographic Information Institute (NGII). Land use for the study area was classified into 10 categories: wood land, railroad, river, field, plot, road, school, hybrid land, brook and unclassified area (Figure 2b). The rate of land subsidence compared to the area of each category was higher in the railroad and school classes [21]. The surface slope was calculated from a digital elevation model (DEM) constructed from a digital elevation contour line with 1:5000 scale [32] published by NGII (Figure 2c). Surface slope was considered an affecting factor because land subsidence can change surface slope, differential horizontal strain, and vertical displacement [33]. Distance from drift was calculated from a digital drift map provided by the Mine Reclamation Corporation (MIRECO) [26] (Figure 2d). The map is important because it identifies the areas of mining activity in this region. Geological discontinuities are considered to be factors affecting land subsidence, but no geological lineaments appear in the study area on the available 1:50,000 geological map. Therefore, geomorphological lineament was visually extracted from an IKONOS satellite image by a field geologist (Figure 2e). If the location is near a lineament, the value of distance from lineament is low.

The borehole data in the study area, provided by the Mine Reclamation Organization (MIRECO) in 1996 [26], were collected from 29 boreholes (Figure 1 and Table 3). The depths of the boreholes ranged from a minimum of 19.5 m to a maximum of 200 m. The data included hydrologic properties and rock mass information [34]. The depth of groundwater, rock mass rating (RMR), and permeability were obtained from 16, 19, and 6 boreholes, respectively (Table 3 and Figure 2f,g,h). The maximum depth of groundwater was 42.5 m. On the railroad, the upper part of the railroad had a deeper groundwater depth and lower elevation than the lower part of the railway. The RMR was classified as classes 1–5, representing very good, good, fair, poor, and very poor, respectively. In this study, the RMR ranged from 2–4.5. The lowest RMRs appeared in the northwest and southeast portions of the railroad. Permeability was classified as classes 1–6, representing very highly (>1 cm/s), highly ($1-10^{-2}$ cm/s), moderately ($10^{-2}-10^{-3}$ cm/s), slightly ($10^{-3}-10^{-5}$ cm/s), and very slightly ($10^{-5}-10^{-7}$ cm/s) permeable and practically impermeable ($<10^{-7}$ cm/s), respectively. In this study, the permeability grade ranged from 4–4.5 (slightly permeable). The groundwater data were collected from a report published in May 1996 by the Coal Industry Promotion Board. Borehole point data should be converted into raster data for spatial analysis, and the accuracy of a raster map depends on the number of data points. However, the available borehole data were limited in this study. Therefore, raster maps from the limited borehole data were constructed using an inverse distance weighting (IDW) interpolation method, which is useful for predicting values at unmeasured locations where data are insufficient [11].

Eight control factors influencing land subsidence were constructed with $2\text{ m} \times 2\text{ m}$ grid data, resulting in 775 columns and 860 rows, for a total of 666,500 cells within the study area. In total, 24 land subsidence areas as 24 vector-type polygons were converted to $2\text{ m} \times 2\text{ m}$ grid data for a total of 3863 cells with a value of 1. The 3863 cells of land subsidence were randomly classified into training and validation sets, with a 50% (1931 cells) and 50% (1932 cells) distribution, respectively, to evaluate model performance.

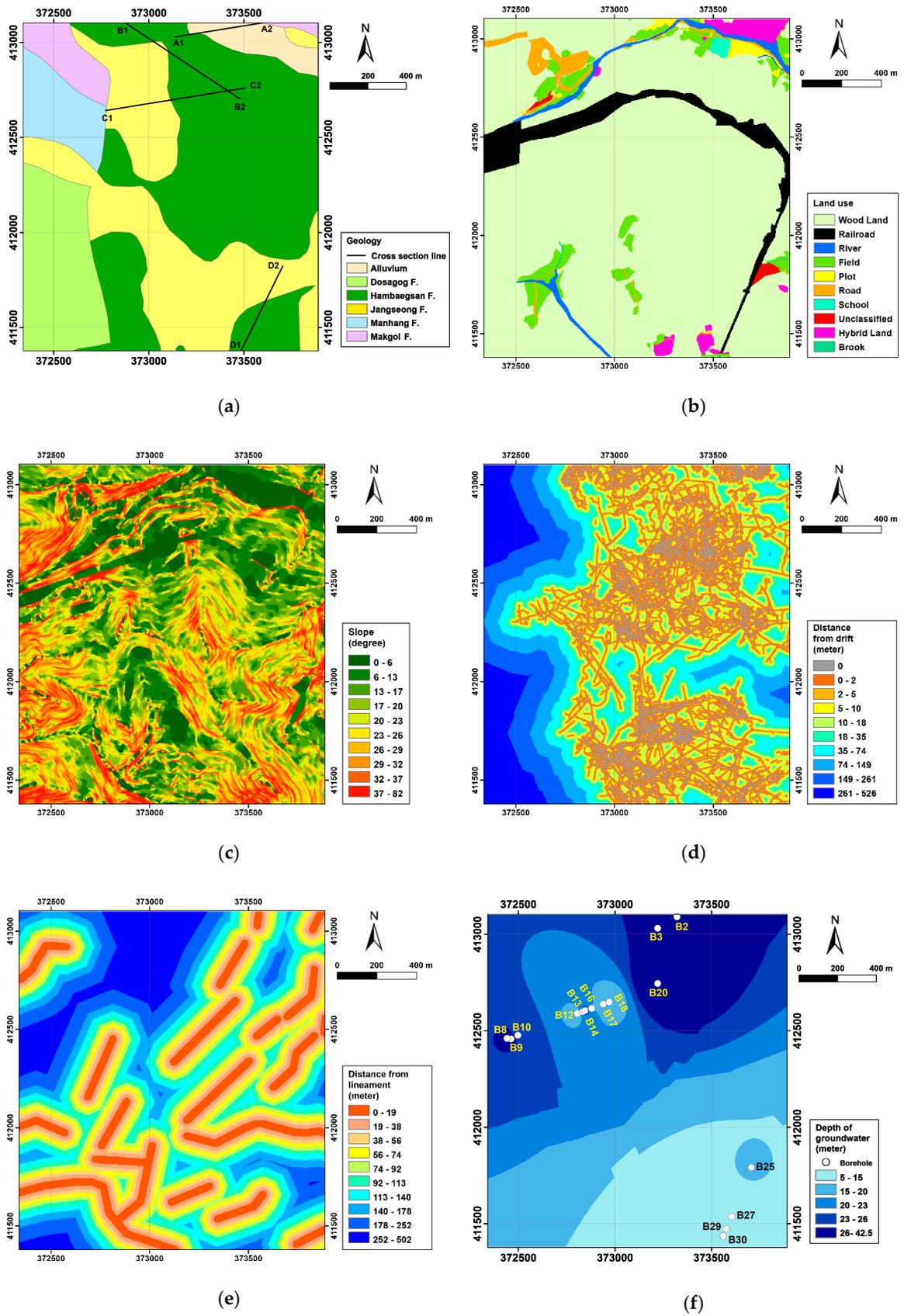


Figure 2. Cont.

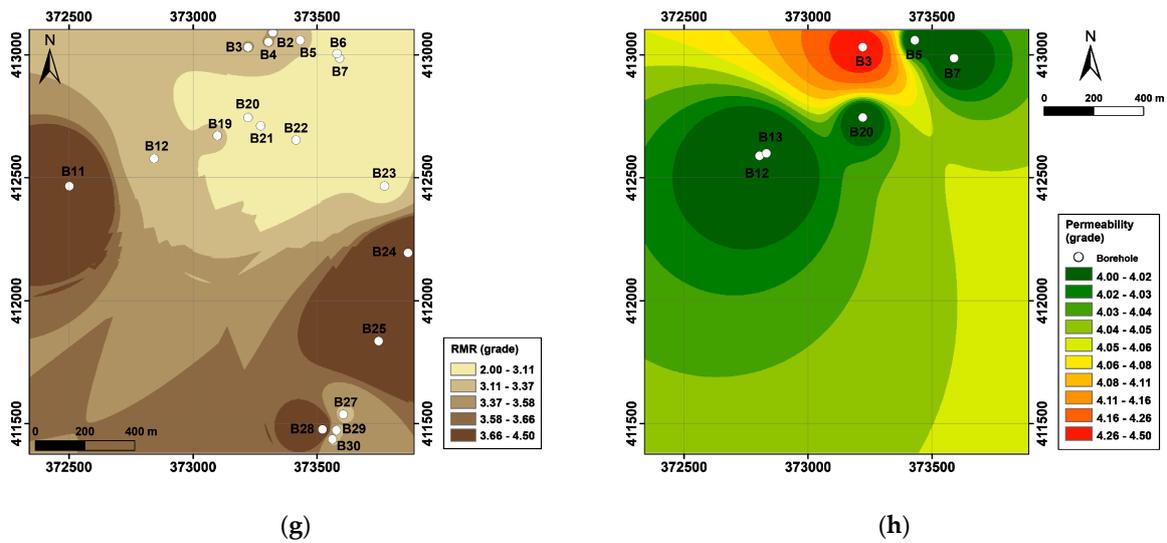


Figure 2. Eight factors influencing coal mine subsidence were used as input data in this study: (a) Geology, (b) land use, (c) slope, (d) distance from drift, (e) distance from lineament, (f) groundwater depth, (g) rock mass rating (RMR), and (h) permeability.

Table 2. Description of geological stratigraphy in Taebaek [30].

Geological Age	Formation	Thickness (m)	Description
Quaternary	Alluvium (Qa)	~20	- Gravel, sand, and clay
			- Mainly milky white–light green coarse–very coarse sandstone with greenish-gray–gray shale interbeds.
	Dosagog (Pd)	250–350	Intercalations of pinkish sandstone, purple shale, and grayish-green sandy shale in the upper part. The sandstone is less compact than that of the Hambaegsan Formation.
Permian	Hambaegsan (Ph)	70–250	- Mainly milky white–light gray coarse sandstone with some interbeds of black shale with thickness of 2–3 m. Some pebbly sandstones occur at the base.
	Jangseong (Pj)	80–150	- Four–five cyclothem consisting of dark-gray sandstone, black shale, and coal seam. Abundant plant fossils occur in the shale above the coal seam, the most valuable anthracite bed, of the 3rd–4th cyclothem from the bottom.
			- Mainly dark-gray–black shale and dark-gray fine sandstone intercalated with dark-gray limestone lenses and two to three thin coal seams
Carboniferous	Geumcheon (Cg)	50–100	
	Manhang (Cm)	250–300	- Mainly purple, greenish-gray, or light-green shale and light-green–green or light-gray medium–very coarse sandstone intercalated with three–four limestone lenses. Conglomerates with a thickness of a few meters occur at the base in some places.
			
Ordovician	Manhang (Om)		- In the upper part, gray–dark gray limestone intercalated with dolomite

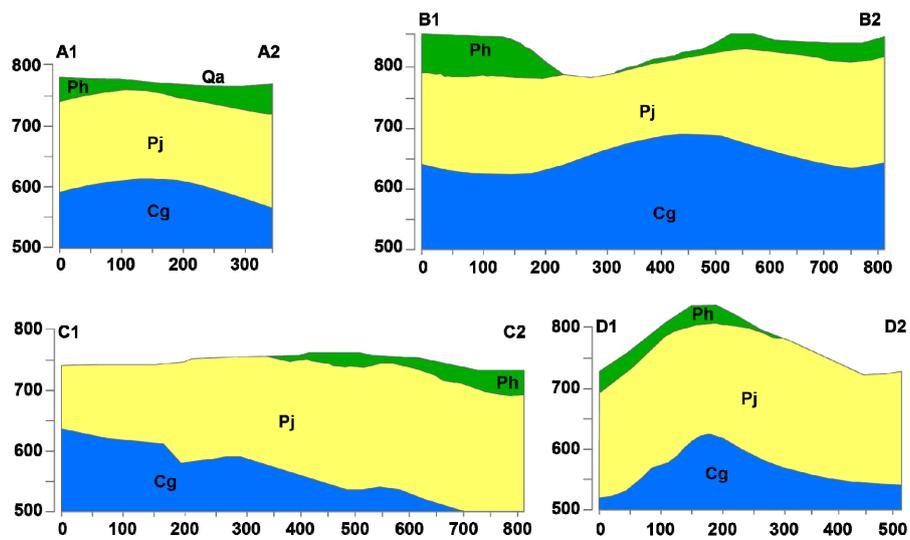


Figure 3. Geological cross sections in the study area.

Table 3. Borehole data in the study area.

ID	Depth of Borehole	Depth of Groundwater (m)	RMR (grade)	Permeability (grade)	Geology
B1	50.0	32.0	3.4	-	Alluvium-Hambaegsan
B2	50.0	27.2	3.4	4.5	Alluvium-Hambaegsan
B3	30.0	-	3.4	-	Alluvium-Hambaegsan
B4	60.2	-	3.4	4	Alluvium-Hambaegsan
B5	86.3	-	2.0	-	Alluvium-Hambaegsan
B6	80.0	-	2.0	4	Alluvium-Hambaegsan
B7	33.0	27.5	-	-	Jangseong
B8	20.5	27.7	-	-	Jangseong
B9	40.0	26.1	-	-	Jangseong
B10	35.5	-	4.4	-	Jangseong
B11	30.0	15.7	-	4	Jangseong
B12	40.5	21.6	-	4	Jangseong
B13	41.1	29.4	-	-	Jangseong
B14	22.0	-	3.2	-	Jangseong
B15	35.7	20.0	-	-	Jangseong
B16	40.8	20.0	-	-	Jangseong
B17	50.5	14.7	-	-	Jangseong
B18	58.0	-	3.2	-	Jangseong
B19	54.0	42.5	2.5	4	Hambaegsan-Jangseong
B20	60.0	-	3.0	-	Hambaegsan-Jangseong
B21	115.0	-	3.0	-	Hambaegsan-Jangseong
B22	80.0	-	3.0	-	Hambaegsan-Jangseong
B23	80.0	-	4.5	-	Hambaegsan-Jangseong
B24	84.0	-	4.3	-	Jangseong
B25	80.4	18.0	-	-	Jangseong
B26	19.5	5.0	3.3	-	Hambaegsan
B27	200.0	-	4.3	-	Hambaegsan-Jangseong
B28	40.0	5.0	3.3	-	Hambaegsan-Jangseong
B29	35.0	5.5	3.3	-	Hambaegsan-Jangseong

4. Methods

As shown in Figure 4, the mapping process consisted of five steps: (a) Spatial database construction, (b) random categorization of land subsidence locations into training and validation datasets at a ratio of 1:1, (c) selection of land subsidence conditioning factors, (d) application of machine learning methods to map LSS, and (e) validation and comparison of the five models.

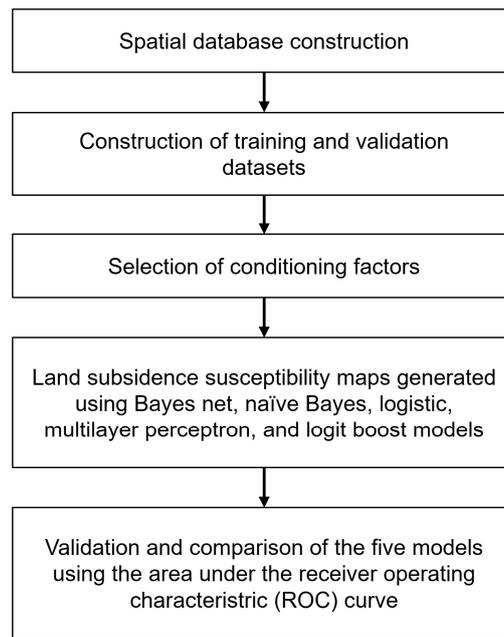


Figure 4. Flowchart for the generation of land subsidence susceptibility (LSS) maps using various machine learning models including Bayes net, naïve Bayes (NB), logistic, multilayer perceptron, and logit boost models.

4.1. Models

4.1.1. Bayes Net (BN)

The BN algorithm applies Bayes’ theorem to produce graphical representations of the probability distribution [35]. BN is commonly used to model complex systems [36]. BN has not yet been used to model land subsidence; however, Pham et al. (2016) [37] applied this algorithm to evaluate landslide risk. The distinct universal probability of a subsidence event for a set of input factors can be estimated as follows:

$$P_B(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P_B(X_i | \Pi_{x_i}) = \prod_{i=1}^n \theta_{x_i | \Pi_{x_i}} \tag{1}$$

where $X = (X_1, X_2, \dots, X_n)$ represents the subsidence input factors, $P_B(X_i | \Pi_{x_i}) = \theta_{x_i | \Pi_{x_i}}$ is a common probability distribution for input factors X_i , and n is the number of subsidence input factors [37].

4.1.2. Naïve Bayes (NB)

The NB algorithm is a classification system that applies Bayes’ theorem under the assumption of conditional independence for all attributes [10,38]. The NB classifier is easy to build, without any need for complicated iterative parameter-estimation schemes [38]. The NB algorithm estimates the probability $P(y_j/x_i)$ for all possible output classes as shown in Equation (2). The class with the largest posterior probability is predicted as follows:

$$y = \operatorname{argmax} P(y_j) \prod_{i=1}^n P(x_i/y_j) \tag{2}$$

{subsidence, no subsidence}

where x_i is the input factor, y_j is the output class, $P(y_j)$ is the prior probability, and $P(y_j/x_i)$ is the conditional probability.

The conditional probability is calculated as

$$P\left(\frac{x_i}{y_j}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/2\sigma^2} \quad (3)$$

where μ is the mean and σ is the standard deviation of x_i .

4.1.3. Logistic Regression (LR)

LR is a statistical technique that allows the predictor to analyze several types of variables [39–41]. LR does not require the normality assumption, which is an advantage over linear and log-linear regression. The inclusion of multiple parameters offers the user the ability to select the best predictors for use in the model [39]. The LR model is formulated as follows [42]:

$$f(x) = \text{logit}(P) = \ln\left[\frac{P}{1-P}\right] = c_0 + c_1x_1 + \dots + c_nx_n \quad (4)$$

$$P = \frac{1}{1 + e^{-f(x)}} = \frac{1}{1 + e^{-(c_0 + c_1x_1 + \dots + c_nx_n)}} \quad (5)$$

where x_1, x_2, \dots, x_n are the input factors, c_0 is the model intercept, and c_1, \dots, c_n are the regression coefficients to be approximated. In this study, P is the probability of subsidence occurrence and $1 - P$ is probability that subsidence will not occur. The function $f(x)$ is represented as $\text{logit}(P)$.

4.1.4. Multilayer Perceptron (MLP)

MLP is an artificial neural network classifier that is widely used in various fields [12,43]. MLP neural nets consist of three structures: Input, hidden, and output layers. In this study, the input layers represent factors that affect land subsidence, and the inputs are processed to become outputs within the hidden layers. The classification results, dividing land subsidence and non-subsidence, are shown in the output layers [12,44]. Two processes are required to train data from MLP neural nets: 1) Forward propagation of the inputs through the hidden layers to obtain output and compare output values to initial values, and 2) adjustment of the connection weights using differences between subsequent values to generate the best results [44,45]. In this study, $t = t_i, i = 1, 2, \dots, 8$ is a vector containing eight land-subsidence conditioning factors, and $\phi = \phi_j, j = 1, 2$ represents the land subsidence and non-subsidence classes. The MLP neural net function is then determined as follows:

$$\phi = f(t) \quad (6)$$

where $f(t)$ is an unknown function that is improved during the training process by adjustable network weights for a given network architecture.

An advantage of MLP is that the user is not required to decide the relative importance of the various input measurements; most inputs can be selected during the training process, based on weight adjustment [46]. Additionally, MLP does not require assumptions about the distribution of the training dataset.

4.1.5. Logit Boost (LB)

LB is a famous machine-learning algorithm introduced by Friedman et al., 2000 [47] that effectively reduces bias and variance; it is a slight modification of the most popular boosting method (AdaBoost) for handling noisy data [48], which reduces training errors and improves classification accuracy [49]. LB has been widely applied in binary classification problems [50], medical science [51], and computer science [52]; however, it has not yet been applied to land-subsidence problems [53].

In the current study, we create a vector $x_i = x_1, x_2, \dots, x_n$, where n is the number of input factors; $y = [1, 0]$ represents two output classes (subsidence or non-subsidence). The LB algorithm is trained in the following steps [47]:

1. Assign weights $\omega_i = \frac{1}{n}$, $i = 1, 2, \dots, n$, $f(x) = 0$ and probability estimates $p_e(x_i) = 1/2$.
2. For $m = 1, 2, \dots, m$, repeat the following steps:
 - a. Compute the working response and weights:

$$r_i = \frac{[y_i^* - p_e(x_i)]}{[p_e(x_i)(1 - p(x_i))]}$$

$$\omega_i = p_e(x_i)(1 - p(x_i))$$

- b. Fit the function by weighted least-squares regression of r_i to x_i using weights ω_i .
- c. Update the function as:

$$f(x) \leftarrow f(x) + \frac{1}{2}f_m(x)$$

$$p(x) \leftarrow \frac{e^{f(x)}}{e^{f(x)} + e^{-f(x)}}$$

3. Output the classifier.

$$\text{sign}[f(x)] = \text{sign} \left[\sum_{m=1}^M f_m(x) \right]$$

$$= \begin{cases} 1 \text{ (subsidence) if } f(x) < 0 \\ -1 \text{ (non subsidence) if } f(x) \geq 0 \end{cases}$$

4.2. Model Evaluation and Comparison

During the modeling and validation phases, model efficiency should be evaluated and compared [44]. We quantitatively evaluated and compared the efficiency of the models according to the area under the ROC curve (%). This technique has been applied to assess risk models of various hazards including subsidence [9], landslides [54], and sinkholes [55]; it is a standard method to quantitatively evaluate the quality of probabilistic and statistical models [56]. The x and y axes of the curve are sensitivity and specificity, respectively [56], and the area under the curve ranges from 0.5–1, with higher values indicating higher model accuracy and prediction capability.

5. Results

5.1. LSS Mapping

Figure 5 shows the LSS maps produced by the five algorithms: Bayes net (Figure 5a), NB (Figure 5b), logistic (Figure 5c), multilayer perceptron (Figure 5d), and logit boost (Figure 5e). To generate the LSS maps, we used the LSS index (LSSI) to classify susceptibility events into four classes: Very high (5% of total area), high (5%), moderate (5%), and low (85%). The probability of land subsidence was predicted for each class, and subsidence hazard was predicted for residential areas. The susceptibility indexes from the five algorithms were similar. The region with very high susceptibility appeared from the western part of the region to the eastern part as railroad area, which is marked by the red color. In the Bayes net result, the very high susceptibility area did not appear as often as in the other models. In the middle of the region, the Bayes net result has a low index, whereas the rest of the models have a very high or high index. Some very high indexes also appear in the northeastern part of the region, as elementary school area, but most of the region has a low susceptibility index rank for subsidence.

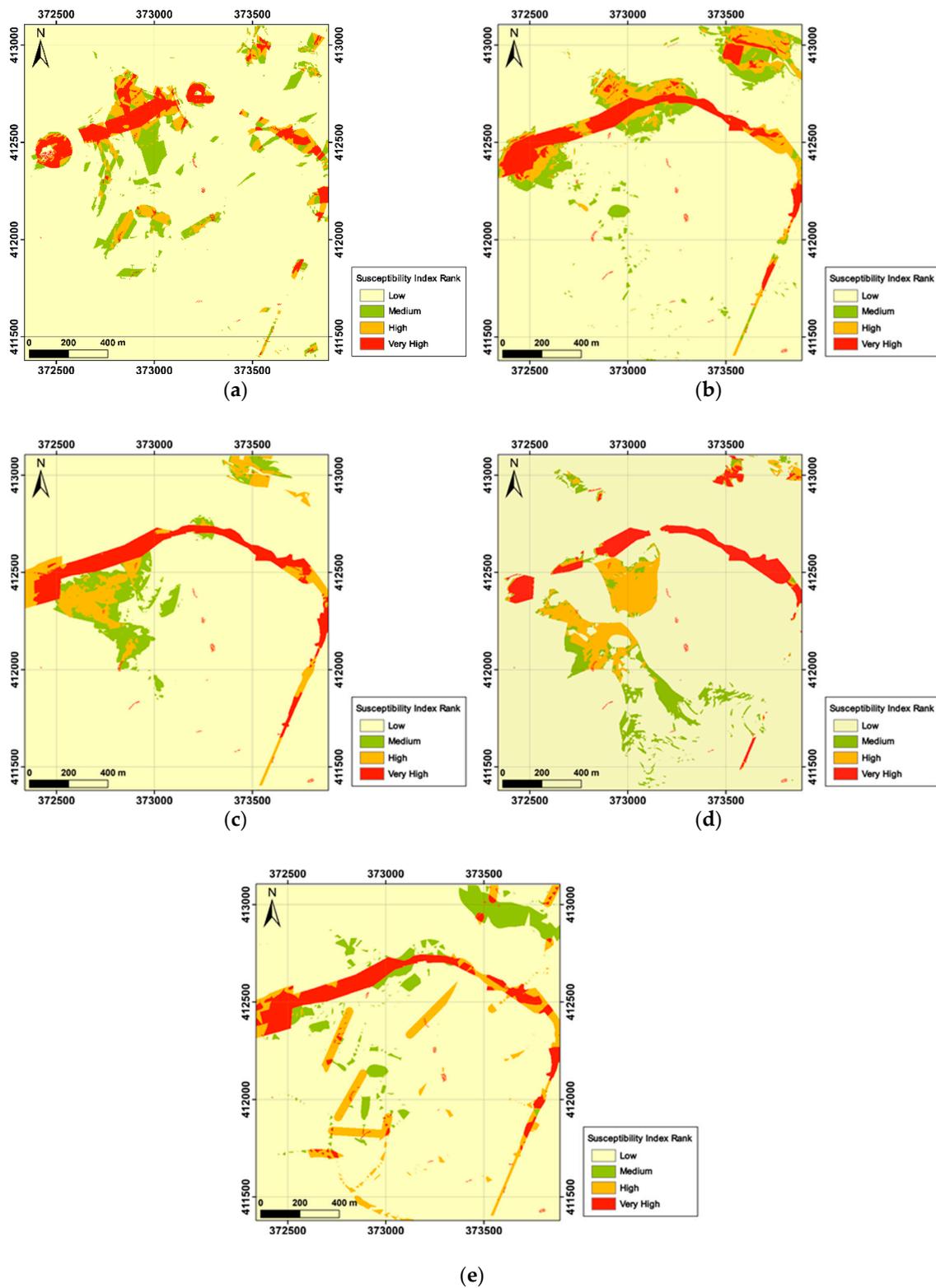


Figure 5. LSS maps generated using the five algorithms: (a) Bayes net, (b) NB, (c) logistic, (d) multilayer perceptron, and (e) logit boost.

However, there are some differences for the medium-susceptibility index rank, marked by the green color. The area with medium susceptibility of land subsidence is spreading and has a different pattern in each model result. For example, the NB and logit boost results show the northern part

of the region is mostly covered by the medium susceptibility index. In contrast, the multilayer perceptron shows the medium index in the southern part of the region. Meanwhile, in the Bayes net and logistic models, the medium index is diffusely distributed from the northern to the middle part of the study area.

5.2. Validation

The land subsidence susceptibility (LSS) analysis results were validated by comparison with 1932 land subsidence cells (i.e., 50% of the total subsidence data) that had not been used in the analysis. A quantitative comparison among all models of the receiver operating characteristic (ROC) curves for model performance is shown in Figure 6. The land subsidence susceptibility index (LSSI) values of all cells were sorted in descending order, divided into 100 classes [57], and associated with the cumulative number of subsidence events for each class (Figure 6). The model with the highest area under the ROC curve was considered to be the model with the best predictive performance. The area under the curve values for the Bayes net, naïve Bayes (NB), logistic, multilayer perceptron, and logit boost models were 0.8640, 0.8539, 0.8892, 0.8676, and 0.9144, respectively; thus, the respective LSS mapping accuracy rates were 86.42, 85.39, 88.92, 86.76, and 91.44%. Although all models had sufficient performance, the different applied models had different prediction performances using same training data. In particular, the logit boost model had a higher predictive accuracy (by about 2.52, 4.68, 5.02, and 6.05%, respectively) than the logistic, multilayer perceptron, Bayes net, and NB. Therefore, model reliability followed the order logit boost > logistic > multilayer perceptron > Bayes net > NB. The percentage differences of the validation result are discussed in Section 6.

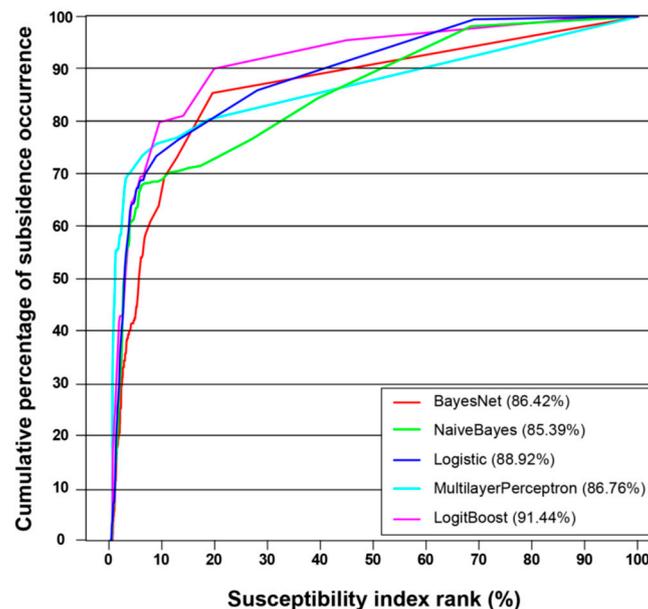


Figure 6. Susceptibility index rank (x-axis) and subsidence occurrence (y-axis) of the five algorithms.

6. Discussion

Recently, there has been great interest within the hazard prediction community toward improving the performance of hazard susceptibility models. In various fields, machine learning techniques have been shown to be effective in terms of performance [58–62]. In particular, ensemble learning has improved machine learning results by combining several models [17,63,64]. The results of different applied models under the same conditions (i.e., study area, input data, ratio of training, and validation datasets) can be compared to the quantitative accuracy values of the area under the ROC to present the predictive power of the model. Models with similar (different) accuracy values can be said to have

similar (differing) performances. Therefore, the reliabilities of the models can be ordered according to the accuracies of the models.

In this study, the logit boost model, based on ensemble machine learning, had a 91.44% accuracy and a predictive accuracy that was higher (by 2.52–6.05%) than those of the logistic, multilayer perceptron, Bayes net, and NB based on machine learning. Similarly, a previous study [2] found that a decision tree model (the CHAID algorithm) produced LSS maps with higher accuracy (94.01%) than the QUEST decision tree (90.37%) and frequency ratio (86.70%). The other algorithms examined in the current study also exhibited high accuracy. Thus, the Bayes net, NB, logistic, and multilayer perceptron models can also be used as alternative models for mapping land subsidence hazard risk. Even though the logit boost model, as an ensemble model, had not been used to predict land subsidence in previous research, the results of the current study indicate that it can achieve high accuracy.

However, some limitations of the models might be a consideration for future studies. For example, the Bayes net model assumes no missing values, and this model also needs to be updated, especially for estimating the conditional probabilities [65]. The benefits and drawbacks of the machine learning models are influenced by several factors, such as the availability of datasets, characteristics of the study area, and condition of the region [18]. The use of Bayesian algorithms, such as the Bayes net and Naïve Bayes, has not been fully verified in natural hazard assessments [18]. According to Mezaal [66], the multilayer perceptron algorithm also has limitations, such as overlearning and high computational complexity.

It has been reported that the sinkhole subsidence attributable to underground mining is caused by shallow depth, weak overburden, geological discontinuities, solution of rocks, rainfall, groundwater, and earthquakes [25]. However, this study used a spatial database obtained from previous studies due to the limitation of available data. No further surveys or new surveys on land subsidence have been conducted in the study area for 14 years. If real-time monitoring data and additional data are obtained in the study area, a 4D underground subsidence model [67] with 3D geological modeling could be constructed to predict land subsidence hazard areas accurately. Thus, continuous monitoring and detailed new surveying for causative factors are essential in the study area. The maps produced in this study can be used as basic data for policymakers and further research. Future studies should develop alternative models and methods to determine the relative influence of factors affecting LSS, so that these methods can be applied in other regions.

7. Conclusions

Land subsidence is a hazardous effect of coal mine abandonment, including that in Korea. To prevent damage and loss of life in the Taebaek region, it is necessary to predict areas with high subsidence risk effectively. In this study, we used Bayesian (i.e., Bayes net and NB), functional (i.e., logistic, multilayer perceptron), and meta-ensemble (i.e., logit boost) machine learning models to perform LSS assessments. Although all models had sufficient performance, the logit boost meta-ensemble machine learning model had the highest accuracy (91.44%) among the five models. The logit boost model also had higher predictive accuracy (by 2.52%, 4.68%, 5.02%, and 6.05%, respectively) than the logistic, multilayer perceptron, Bayes net, and NB models. According to previous studies [11,57] in the same study area, the fuzzy operator with 84.40–88.98% accuracy, frequency ratio with 86.70% accuracy, CHAID decision tree with 94.01% accuracy, and QUEST decision tree with 90.37% accuracy have been applied to the subsidence hazard assessment, but the five models used in this study had been rarely applied. Based on these case studies, the land subsidence hazard rating can be applied to future policy decisions using additional data.

Author Contributions: Conceptualization, S.L. and C.-W.L.; methodology, S.L. and H.-J.O.; software, H.-J.O. and M.S.; validation, H.-J.O. and M.S.; formal analysis, H.-J.O. and M.S.; investigation, S.L. and H.-J.O.; resources, S.L. and H.-J.O.; writing—original draft preparation, H.-J.O. and M.S.; writing—review and editing, S.L. and C.-W.L.; visualization, H.-J.O.; supervision, S.L. and C.-W.L.; project administration, S.L. and C.-W.L.; funding acquisition, S.L. and C.-W.L.

Funding: This research was supported by the Basic Research Project of the Korea Institute of Geoscience and Mineral Resources (KIGAM), funded by the Ministry of Science, ICT, and Future Planning of Korea. This work was supported by a National Research Foundation of Korea (NRF) grant from the Korea government (MSIP) (number NRF-2017R1A2B4003258).

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

References

1. Ghorbanzadeh, O.; Rostamzadeh, H.; Blaschke, T.; Gholaminia, K.; Aryal, J. A New Gis-Based Data Mining Technique Using an Adaptive Neuro-Fuzzy Inference System (Anfis) and K-Fold Cross-Validation Approach for Land Subsidence Susceptibility Mapping. *Nat. Hazards* **2018**, *94*, 497–517. [[CrossRef](#)]
2. Lee, S.; Park, I. Application of decision tree model for the ground subsidence hazard mapping near abandoned underground coal mines. *J. Environ. Manag.* **2013**, *127*, 166–176. [[CrossRef](#)] [[PubMed](#)]
3. Oh, H.-J.; Lee, S. Integration of ground subsidence hazard maps of abandoned coal mines in Samcheok, Korea. *Int. J. Coal Geol.* **2011**, *86*, 58–72. [[CrossRef](#)]
4. Pradan, B.; Abokharima, M.H.; Jebur, M.N.; Tehrani, M.S. Land Subsidence Susceptibility Mapping at Kinta Valley (Malaysia) Using the Evidential Belief Function Model in Gis. *Nat. Hazards* **2014**, *73*, 1019–1042. [[CrossRef](#)]
5. Lee, S.; Park, I.; Choi, J.K. Spatial prediction of ground subsidence susceptibility using an artificial neural network. *Environ. Manag.* **2012**, *49*, 347–358. [[CrossRef](#)] [[PubMed](#)]
6. Oh, H.-J.; Ahn, S.-C.; Choi, J.-K.; Lee, S. Sensitivity analysis for the GIS-based mapping of the ground subsidence hazard near abandoned underground coal mines. *Environ. Earth Sci.* **2011**, *64*, 347–358. [[CrossRef](#)]
7. Pishro, M.; Khosravi, S.; Tehrani, S.M.; Mousavi, S.R. Modeling and zoning of land subsidence in the southwest of Tehran using artificial neural networks. *Int. J. Hum. Cap. Urban Manag.* **2016**, *1*, 159–168.
8. Rafie, M.; Samimi Namin, F. Prediction of subsidence risk by FMEA using artificial neural network and fuzzy inference system. *Int. J. Min. Sci. Technol.* **2015**, *25*, 655–663. [[CrossRef](#)]
9. Tien Bui, D.; Shahabi, H.; Shirzadi, A.; Chapi, K.; Pradhan, B.; Chen, W.; Khosravi, K.; Panahi, M.; Bin Ahmad, B.; Saro, L. Land Subsidence Susceptibility Mapping in South Korea Using Machine Learning Algorithms. *Sensors* **2018**, *18*, 2464. [[CrossRef](#)]
10. Ilija, I.; Loupasakis, C.; Tsangaratos, P. Land subsidence phenomena investigated by spatiotemporal analysis of groundwater resources, remote sensing techniques, and random forest method: The case of Western Thessaly, Greece. *Environ. Monit. Assess.* **2018**, *190*, 623. [[CrossRef](#)]
11. Choi, J.-K.; Kim, K.-D.; Lee, S.; Won, J.-S. Application of a fuzzy operator to susceptibility estimations of coal mine subsidence in Taebaek City, Korea. *Environ. Earth Sci.* **2010**, *59*, 1009–1022. [[CrossRef](#)]
12. Pham, B.T.; Tien Bui, D.; Prakash, I.; Dholakia, M.B. Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. *CATENA* **2017**, *149*, 52–63. [[CrossRef](#)]
13. Kanevski, M.; Parkin, R.; Pozdnukhov, A.; Timonin, V.; Maignan, M.; Demyanov, V.; Canu, S. Environmental data mining and modeling based on machine learning algorithms and geostatistics. *Environ. Model. Softw.* **2004**, *19*, 845–855. [[CrossRef](#)]
14. Baillifard, F.; Jaboyedoff, M.; Sartori, M. Rockfall hazard mapping along a mountainous road in Switzerland using a GIS-based parameter rating approach. *Nat. Hazards Earth Syst. Sci.* **2003**, *3*, 435–442. [[CrossRef](#)]
15. Samanta, S.; Pal, D.K.; Palsamanta, B. Flood susceptibility analysis through remote sensing, GIS and frequency ratio model. *Appl. Water Sci.* **2018**, *8*, 66. [[CrossRef](#)]
16. Lee, S.; Pradhan, B. Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models. *Landslides* **2007**, *4*, 33–41. [[CrossRef](#)]
17. Kadavi, P.R.; Lee, C.-W.; Lee, S. Application of Ensemble-Based Machine Learning Models to Landslide Susceptibility Mapping. *Remote Sens.* **2018**, *10*, 1252. [[CrossRef](#)]
18. Pham, B.T.; Prakash, I.; Khosravi, K.; Chapi, K.; Trinh, P.T.; Ngo, T.Q.; Hosseini, S.V.; Bui, D.T. A comparison of Support Vector Machines and Bayesian algorithms for landslide susceptibility modelling. *Geocarto Int.* **2018**, 1–23. [[CrossRef](#)]

19. Darabi, H.; Choubin, B.; Rahmati, O.; Torabi Haghighi, A.; Pradhan, B.; Kløve, B. Urban flood risk mapping using the GARP and QUEST models: A comparative study of machine learning techniques. *J. Hydrol.* **2019**, *569*, 142–154. [[CrossRef](#)]
20. Truong, X.L.; Mitamura, M.; Kono, Y.; Raghavan, V.; Yonezawa, G.; Truong, X.Q.; Do, T.H.; Tien Bui, D.; Lee, S. Enhancing Prediction Performance of Landslide Susceptibility Model Using Hybrid Machine Learning Approach of Bagging Ensemble and Logistic Model Tree. *Appl. Sci.* **2018**, *8*, 1046. [[CrossRef](#)]
21. Park, I.; Choi, J.; Jin Lee, M.; Lee, S. Application of an adaptive neuro-fuzzy inference system to ground subsidence hazard mapping. *Comput. Geosci.* **2012**, *48*, 228–238. [[CrossRef](#)]
22. Bang, G.M.; Choi, S.S.; Oh, S.H.; Sin, J.S.; Jeon, M.G.; Woo, S.U.; Heo, J.E.; Cheon, M.N.; Jo, N.S.; Kim, B.C.; et al. *A Detailed Survey of Monitoring in Whajeon Region*; Coal Industry Promotion Board: Seoul, Korea, 1999; p. 681.
23. Kim, J.N. *A Case Study on Stability Analysis of Ground Subsidence in Abandoned Mine Area—Focused on the Vicinity of the Chunjeon Station*; Seoul National University of Science and Technology: Seoul, Korea, 2011.
24. Canbulat, I.; Zhang, C.; Black, K.; Johnston, J.; McDonald, S. Assessment of Sinkhole Risk in Shallow Coal Mining. In Proceedings of the 10th Triennial Conference of Mine Subsidence: Adaptive Innovation for Managing Challenges, Pokolbin, Australia, 5–7 November 2017.
25. Singh, K.B.; Dhar, B.B. Sinkhole subsidence due to mining. *Geotech. Geol. Eng.* **1997**, *15*, 327–341. [[CrossRef](#)]
26. Board, C.I.P. *Ground Stability Investigation for Hwajeon (Korean Edn)*; Coal Industry Promotion Board: Seoul, Korea, 1996; pp. 9–84.
27. Sahu, P.; Lokhande, R.D. An Investigation of Sinkhole Subsidence and its Preventive Measures in Underground Coal Mining. *Procedia Earth Planet. Sci.* **2015**, *11*, 63–75. [[CrossRef](#)]
28. Lee, F.T.; Abel, J.F. *Subsidence from Underground Mining: Environmental Analysis and Planning Considerations*; USGS: Reston, VA, USA, 1983. [[CrossRef](#)]
29. The Geological Society of Korea. *Geologic Atlas of Taebaegsan Region: Homyeong Geological Sheet*; The Geological Society of Korea: Seoul, Korea, 1962.
30. Seo, H.K.; Kim, D.S.; Park, S.H.; Park, J.S.; Bae, D.J.; Yu, Y.S.; Lee, D.Y.; Im, S.B.; Jang, Y.H.; Jo, M.J. *Geologic Atlas of the Samcheog Coalfield*; Korea Institute of Geoscience and Mineral Resources: Daejeon, Korea, 1979; p. 22.
31. National Geographic Information Institute. *Digital Land Characteristics Map with 1:5000 Scale: No. of index: 37816018, 37816019, 37816028, 37816029*; National Geographic Information Institute: Suwon, Korea, 1996.
32. National Geographic Information Institute. *Digital Topographical Map with 1:5000 Scale: No. of index: 37816018, 37816019, 37816028, 37816029*; National Geographic Information Institute: Suwon, Korea, 1996.
33. Tripathi, N.; Singh, R.S. Underground Coal Mine Subsidence Impacts Forest Ecosystem. 2010. Available online: <https://www.researchgate.net/publication/275207695> (accessed on 25 January 2019).
34. Öge, İ.F.; Çırak, M. Relating rock mass properties with Lugeon value using multiple regression and nonlinear tools in an underground mine site. *Bull. Eng. Geol. Environ.* **2017**. [[CrossRef](#)]
35. Marcot, B.; Steventon, J.; Sutherland, G.; McCann, R. Guidelines for Developing and Updating Bayesian Belief Networks Applied to Ecological Modeling and Conservation. *Can. J. Forest Res.* **2006**, *36*, 3063–3074. [[CrossRef](#)]
36. Song, Y.; Gong, J.; Gao, S.; Wang, D.; Cui, T.; Li, Y.; Wei, B. Susceptibility assessment of earthquake-induced landslides using Bayesian network: A case study in Beichuan, China. *Comput. Geosci.* **2012**, *42*, 189–199. [[CrossRef](#)]
37. Pham, B.T.; Pradhan, B.; Tien Bui, D.; Prakash, I.; Dholakia, M.B. A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India). *Environ. Model. Softw.* **2016**, *84*, 240–250. [[CrossRef](#)]
38. Tien Bui, D.; Pradhan, B.; Lofman, O.; Revhaug, I. Landslide susceptibility assessment in vietnam using support vector machines, decision tree, and naive bayes models. *Math. Probl. Eng.* **2012**, *2012*. [[CrossRef](#)]
39. Erener, A.; Mutlu, A.; Sebnem Düzgün, H. A comparative study for landslide susceptibility mapping using GIS-based multi-criteria decision analysis (MCDA), logistic regression (LR) and association rule mining (ARM). *Eng. Geol.* **2016**, *203*, 45–55. [[CrossRef](#)]
40. Ozdemir, A.; Altural, T. A comparative study of frequency ratio, weights of evidence and logistic regression methods for landslide susceptibility mapping: Sultan Mountains, SW Turkey. *J. Asian Earth Sci.* **2013**, *64*, 180–197. [[CrossRef](#)]

41. Mertler, C.A.; Reinhart, R.V. *Advanced and Multivariate Statistical Methods: Practical Application and Interpretation: Sixth Edition*; Routledge: New York, NY, USA, 2016; pp. 1–374. [[CrossRef](#)]
42. Hosmer, D.W., Jr.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*; John Wiley & Sons: Hoboken, NJ, USA, 2013; Volume 398.
43. Haykin, S.S. *Neural Networks and Learning Machines/Simon Haykin*; Prentice Hall: New York, NY, USA, 2009.
44. Pham, B.T.; Tien Bui, D.; Pourghasemi, H.R.; Indra, P.; Dholakia, M.B. Landslide susceptibility assessment in the Uttarakhand area (India) using GIS: A comparison study of prediction capability of naïve bayes, multilayer perceptron neural networks, and functional trees methods. *Theor. Appl. Climatol.* **2017**, *128*, 255–273. [[CrossRef](#)]
45. Bui, D.T.; Pradhan, B.; Revhaug, I.; Nguyen, D.B.; Pham, H.V.; Bui, Q.N. A novel hybrid evidential belief function-based fuzzy logic model in spatial prediction of rainfall-induced shallow landslides in the Lang Son city area (Vietnam). *Geomat. Nat. Hazards Risk* **2015**, *6*, 243–271. [[CrossRef](#)]
46. Gardner, M.W.; Dorling, S.R. Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences. *Atmos. Environ.* **1998**, *32*, 2627–2636. [[CrossRef](#)]
47. Friedman, J.; Tibshirani, R.; Hastie, T. Additive Logistic Regression: A Statistical View of Boosting (With Discussion and a Rejoinder by the Authors). *Ann. Stat.* **2000**, *28*, 337–407. [[CrossRef](#)]
48. Zhang, G.; Fang, B. LogitBoost classifier for discriminating thermophilic and mesophilic proteins. *J. Biotechnol.* **2007**, *127*, 417–424. [[CrossRef](#)] [[PubMed](#)]
49. Song, J.; Lu, X.; Liu, M.; Wu, X. Stratified Normalization LogitBoost for Two-Class Unbalanced Data Classification. *Commun. Stat. Simul. Comput.* **2011**, *40*, 1587–1593. [[CrossRef](#)]
50. Fraz, M.M.; Remagnino, P.; Hoppe, A.; Uyyanonvara, B.; Rudnicka, A.R.; Owen, C.G.; Barman, S.A. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Trans. Biomed. Eng.* **2012**, *59*, 2538–2548. [[CrossRef](#)]
51. Cai, Y.-D.; Feng, K.-Y.; Lu, W.-C.; Chou, K.-C. Using LogitBoost classifier to predict protein structural classes. *J. Theor. Biol.* **2006**, *238*, 172–176. [[CrossRef](#)]
52. Lutz, R.W. Logitboost with trees applied to the wcci 2006 performance prediction challenge datasets. In Proceedings of the 2006 IEEE International Joint Conference on Neural Network, Vancouver, BC, Canada, 16–21 July 2006; pp. 1657–1660.
53. Pham, B.T.; Tien Bui, D.; Dholakia, M.B.; Prakash, I.; Pham, H.V. A Comparative Study of Least Square Support Vector Machines and Multiclass Alternating Decision Trees for Spatial Prediction of Rainfall-Induced Landslides in a Tropical Cyclones Area. *Geotech. Geol. Eng.* **2016**, *34*, 1807–1824. [[CrossRef](#)]
54. Conforti, M.; Pascale, S.; Robustelli, G.; Sdao, F. Evaluation of prediction capability of the artificial neural networks for mapping landslide susceptibility in the Turbolo River catchment (Northern Calabria, Italy). *CATENA* **2014**, *113*, 236–250. [[CrossRef](#)]
55. Ozdemir, A. Sinkhole susceptibility mapping using logistic regression in Karapınar (Konya, Turkey). *Bull. Eng. Geol. Environ.* **2016**, *75*, 681–707. [[CrossRef](#)]
56. Zweig, M.H.; Campbell, G. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clin. Chem.* **1993**, *39*, 561–577.
57. Park, I.; Lee, S. Spatial prediction of landslide susceptibility using a decision tree approach: A case study of the Pyeongchang area, Korea. *Int. J. Remote Sens.* **2014**, *35*, 6089–6112. [[CrossRef](#)]
58. Korup, O.; Stolle, A. Landslide Prediction from Machine Learning. *Geol. Today* **2014**, *30*, 26–33. [[CrossRef](#)]
59. McGaughey, W.J.; Laflèche, V.; Howlett, C.; Sydor, J.L.; Campos, D.; Purchase, J.; Huynh, S. Automated, Real-Time Geohazard Assessment in Deep Underground Mines. In *Proceedings of the Eighth International Conference on Deep and High Stress Mining*; Wesseloo, J., Ed.; Australian Centre for Geomechanics: Perth, Australia, 2017; pp. 521–528.
60. Tayfur, G.; Singh, V.P.; Moramarco, T.; Barbetta, S. Flood Hydrograph Prediction Using Machine Learning Methods. *Water* **2018**, *10*, 968. [[CrossRef](#)]
61. Karpatne, A.; Ebert-Uphoff, I.; Ravela, S.; Babaie, H.A.; Kumar, V. Machine Learning for the Geosciences: Challenges and Opportunities. *IEEE Trans. Knowl. Data Eng.* **2018**. [[CrossRef](#)]
62. Canli, E.; Mergili, M.; Thiebes, B.; Glade, T. Probabilistic Landslide Ensemble Prediction Systems: Lessons to Be Learned from Hydrology. *Nat. Hazards Earth Syst. Sci.* **2018**, *18*, 2183–2202. [[CrossRef](#)]

63. Mojaddadi, H.; Pradhan, B.; Nampak, H.; Ahmad, N.; Ghazali, A.H.B. Ensemble Machine-Learning-Based Geospatial Approach for Flood Risk Assessment Using Multi-Sensor Remote-Sensing Data and Gis. *Geomat. Nat. Hazards Risk* **2017**, *8*, 1080–1102. [[CrossRef](#)]
64. Chen, W.; Sun, Z.; Han, J. Landslide Susceptibility Modeling Using Integrated Ensemble Weights of Evidence with Logistic Regression and Random Forest Models. *Appl. Sci.* **2019**, *9*, 171. [[CrossRef](#)]
65. Bouckaert, R.R. *Bayesian Network Classifiers in Weka for Version 3-5-7*; The University of Waikato: Waikato, New Zealand, 2008; Volume 11, pp. 369–387.
66. Mezaal, M.; Pradhan, B.; Shafri, H.; Md Yusoff, Z.; Al-Zuhairi, M. Optimized Neural Architecture for Automatic Landslide Detection from High-Resolution Airborne Laser Scanning Data. *Appl. Sci.* **2017**, *7*, 730. [[CrossRef](#)]
67. Mira Geoscience. *Geohazmap Workflow Earth Modelling Solutions for Mining*. Montreal: Mira Geoscience; Mira Geoscience: Montreal, QC, Canada, 2007.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).