



OPEN

## Suggestion for a new deterministic model coupled with machine learning techniques for landslide susceptibility mapping

Dae-Hong Min<sup>1</sup> & Hyung-Koo Yoon<sup>2</sup>✉

Deterministic models have been widely applied in landslide risk assessment (LRA), but they have limitations in obtaining various geotechnical and hydraulic properties. The objective of this study is to suggest a new deterministic method based on machine learning (ML) algorithms. Eight crucial variables of LRA are selected with reference to expert opinions, and the output value is set to the safety factor derived by Mohr–Coulomb failure theory in infinite slope. Linear regression and a neural network based on ML are applied to find the best model between independent and dependent variables. To increase the reliability of linear regression and the neural network, the results of back propagation, including gradient descent, Levenberg–Marquardt (LM), and Bayesian regularization (BR) methods, are compared. An 1800-item dataset is constructed through measured data and artificial data by using a geostatistical technique, which can provide the information of an unknown area based on measured data. The results of linear regression and the neural network show that the special LM and BR back propagation methods demonstrate a high determination of coefficient. The important variables are also investigated though random forest (RF) to overcome the number of various input variables. Only four variables—shear strength, soil thickness, elastic modulus, and fine content—demonstrate a high reliability for LRA. The results show that it is possible to perform LRA with ML, and four variables are enough when it is difficult to obtain various variables.

The probabilistic and deterministic method has been applied to perform landslide risk assessments (LRA)<sup>1–3</sup>. To perform an LRA based on the probabilistic method, the probability of occurrence is determined by historical literature data and is used to define whether the determined value exceeds the reference value<sup>4</sup>. If there is a large amount of accumulated data linked to the criterion data, which is generally the amount of rainfall or the rainfall intensity in the same area, a reliable risk assessment can be easily performed<sup>5</sup>. However, it has a limitation in providing reliable results with a lack of data in the inventory to fully reflect past and present conditions when using the critical criterion of the neighborhood area<sup>6</sup>. The deterministic method uses geotechnical and hydraulic properties of the target area as input parameters for the LRA, and the susceptibility is generally expressed in terms of the quantitative safety factor<sup>7</sup>. A graphic information system (GIS) is a useful method for classifying the target area by grid, so a GIS is applied to obtain risk areas with high resolution<sup>8</sup>. However, the deterministic method also has limitations of time and money in obtaining various input parameters at each grid. If the measurement results are insufficient due to the circumstances, the reliability may decrease. To solve this problem, Jun et al.<sup>9</sup> presented an empirically based model for an LRA, formed with the opinions of experts, and tried to improve the limitations of the deterministic method through the most important parameters of the LRA. However, this method also has limitations in quantitatively estimating LRA due to various input parameters and their different units<sup>10</sup>. Bui et al.<sup>11</sup> constructed relationships between the soil compression coefficient and 12 input variables with different units based on a neural network. The results showed that reliable output values can be derived among various input parameters even if the units of the input values are different. Therefore, this study is focused on finding reliable constants of each parameter and proposing minimum input variables through machine learning to expand the application of the research results of Jun et al.<sup>9</sup>.

Various optimization methods have been used to predict susceptible areas based on the deterministic method. Chang et al.<sup>12</sup> attempted landslide susceptibility mapping using a digital elevation model as a geomorphic factor,

<sup>1</sup>Department of Construction and Disaster Prevention Engineering, Daejeon University, Daejeon 300-716, Korea. <sup>2</sup>Department of Construction and Disaster Prevention Engineering, Daejeon University, Daejeon 300-716, Korea. ✉email: hyungkoo@dju.ac.kr

and applied logistic regression, random forest, and support vector machine. Di et al.<sup>13</sup> used a gradient boosting machine to predict the susceptibility of debris flow in a watershed, and the results were compared with the outputs deduced by logistic regression, k-nearest neighbor, a support vector machine, and an artificial neural network. In recent years, researches to find a landslide susceptibility area using machine learning have been continuously conducted with selecting different algorithms for the purpose of the study. Van Dao et al.<sup>14</sup>, Kuradusenge et al.<sup>15</sup> and Wang et al.<sup>16</sup> selected the conventional algorithms including logistics regression, support vector machine, random forest, gradient boosting machine, and multilayer perceptron and however, Tien Bui et al.<sup>17</sup> and Ghasemian et al.<sup>18</sup> suggested to the hybrid algorithms of reduced error pruning tree algorithm and combining algorithm both stochastic gradient descent and an AdaBoost meta classifier, respectively, for improving performances. Each study also used different input variables with consideration of classifying the categories of geographical conditions and the obtained values were used for the purpose. In addition, Levenberg–Marquardt and Bayesian regularization algorithms were used as a back propagation method, which has a crucial influence on the outcome of ML<sup>2,3,19</sup> in finding an optimal weight and bias for prediction<sup>20</sup>. Each method is developed to minimize errors when finding the inflection point through a probability technique. Therefore, in this study, ML was applied to perform an LRA based on the safety factor with input variables of various units and ranges, and the back propagation method was also applied to improve reliability.

This paper describes linear regression, a neural network, and random forest among machine learning techniques, and the back propagation methods of gradient descent, Levenberg–Marquardt, and Bayesian regularization are also explained in the background theory. The objectives for obtaining input variables are described, and the experimental results are summarized. In addition, an explanation of the interpolation method to construct the dataset based on geostatistical theory is included, and the LRA is evaluated after determining a reliable hyperparameter. The results were compared with the safety factor calculated by Mohr–Coulomb failure theory<sup>21</sup> to verify the reliability.

## Background theory

**Empirical equation.** Jun et al.<sup>9</sup> suggested parameters closely related to the occurrence of debris flow through the analytic hierarchical process (AHP) technique to overcome the limitations of the existing deterministic method, which assumes various input parameters due to difficulty in obtaining them. The geotechnical properties were categorized into soil structure and particle distribution, the stress and strain of soil, and soil and water, considering the characterizations of solid, structural failure, and water flow. Thus, the fine content (percent passing ratio of a 0.075 mm diameter sieve)<sup>22</sup>, soil thickness (depth from surface to weathered bedrock)<sup>23</sup>, porosity (ratio between volumes of void and solid)<sup>24</sup>, elastic modulus (soil resistance when force is applied)<sup>25</sup>, shear strength (soil strength against yield or failure)<sup>22</sup>, hydraulic conductivity (speed of the moving fluid)<sup>26</sup>, saturation (volumetric water content in soil)<sup>27</sup>, and water content (weighted water content in soil)<sup>28</sup> were selected as main parameters through an expert advisory group. The consistency ratio (CR) was used to estimate the reliability of the determined parameters, and the value was calculated as 0.0022<sup>9</sup>. A CR value less than 0.1 generally indicates an excellent response rate<sup>29</sup>. The weight factor of each parameter was determined by the relative importance and influence on debris flow. However, there is a limitation as a single formula because each factor has a different unit. Even though a scored index based on a range of values in each parameter was proposed, it was still insufficient. To improve the methodology of the previous results through machine learning, the eight selected parameters were used as independent variables, and the dependent variable was fixed to the safety factor. The safety factor was calculated using Eq. (1), derived by Mohr–Coulomb failure theory in infinite slope<sup>21</sup>. This study proposes a formula through the relationships among eight parameters connected to the safety factor as a true value by using machine learning:

$$FS = \frac{C_r + C_s + \cos^2 \theta [\rho_s \cdot g(D - D_w) + (\rho_s \cdot g - \rho_w \cdot g)D_w] \tan \Phi}{D \cdot \rho_s \cdot g \cdot \sin \theta \cdot \cos \theta} \quad (1)$$

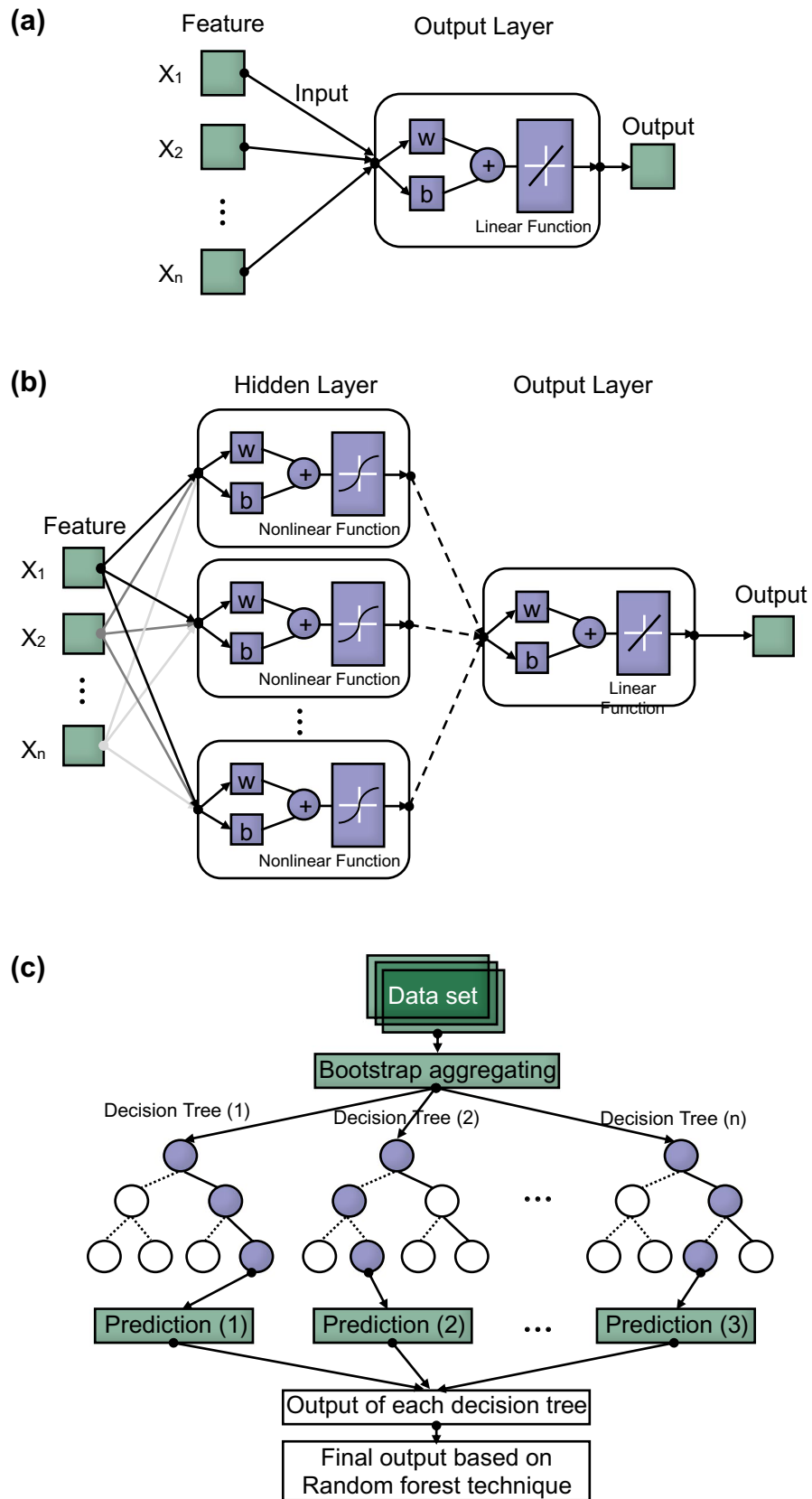
where FS is the safety factor;  $C_r$  and  $C_s$  denote the cohesion ( $\text{N m}^{-2}$ ) of root and soil, respectively;  $\theta$  and  $\phi$  are the slope angle and friction angle ( $^\circ$ );  $\rho_s$  and  $\rho_w$  represent soil density ( $\text{kg m}^{-3}$ ) and water density ( $\text{kg m}^{-3}$ );  $D$  and  $D_w$  represent vertical soil depth (m) and water table height within soil later (m); and  $g$  is gravitational acceleration ( $\text{m s}^{-2}$ ), with a value of 9.81.

**Machine learning.** Linear regression and neural network techniques, based on linear and nonlinear functions of machine learning (ML), were used to investigate the relationships between independent and dependent variables. In addition, the random forest technique based on ML was also applied to find the importance of each factor and characteristics of each technique.

**Linear regression.** Linear regression (LR) is one of the most widely used techniques in ML to find the best model suitable for the distribution of target data, and it has the advantage of identifying multidimensional linear relationships through iterative operations. The architecture and mathematical relations are addressed in Fig. 1a and Eq. (2), with weight ( $w$ ), bias ( $b$ ), input variable ( $x$ ), and output variable based on actual outcome ( $H$ ).

$$H(x) = wx + b \quad (2)$$

**Neural network.** Neural networks (NNs) are used to solve classification and a regression of networks composed of various functions, and can understand input variables through repetitive learning, which is back



**Figure 1.** Architecture of machine learning algorithm: (a) linear regression; (b) neural network; (c) random forest.

propagation<sup>30</sup>. Back propagation updates the weight of each layer in the reverse direction by propagating the error of the output deduced by feedforward, and iterative learning is performed until the error is minimized<sup>31</sup>. The function to perform iterative learning is defined as an activation function, and various nonlinear functions are applied according to the characterization of the dataset. NNs are composed of input and output layers in the same way as LR, but they include a hidden layer that can perform deep learning step by step between layers, as shown in Fig. 1b. The mathematical relationship between the input and output variables is the same as Eq. (2) of LR, and the repeating process through hidden layers is different in the method of finding the weight and bias. Note that the process for finding weight and bias of LR and NN is similar based on the same optimization technique and however, they can be updated in the only NN through the hidden layer.

**Random forest.** Random Forest (RF) is an ML technique used to infer correlations between input and output variables. Thus, the technique was applied to examine the importance of eight input parameters with the safety factor. The RF architecture is composed of multiple decision trees based on an ensemble model, as shown in Fig. 1c, so the accumulation of errors occurring in each tree can be prevented. Each input parameter is randomly selected and allowed to duplicate through bootstrap aggregation, and the validation is estimated between selected data and out-of-bag (OOB) data, which are not selected data<sup>32</sup>. With the OOB of the original and randomly constructed tree called  $O_{OOB}$  and  $R_{OOB}$ , respectively, the importance of each input factor is expressed as Eq. (3). If the calculated importance score is high, it can be judged as an important factor among each input parameter.

$$\text{Important score}_i = \frac{\frac{1}{t} \sum_{i=1}^t |O_{OOB_i} - R_{OOB_i}|}{\sqrt{\frac{1}{t-1} \sum_{i=1}^t \left[ (|O_{OOB_i} - R_{OOB_i}|) - \left( \frac{1}{t} \sum_{i=1}^t |O_{OOB_i} - R_{OOB_i}| \right) \right]^2}} \quad (3)$$

RF also provides the numerical relationship between two variables in input parameters, called the gray relational grade (GRG), expressed as Eq. (4):

$$\text{GRG} = \frac{1}{n} \sum_{j=1}^n \frac{\min_i \min_j |X_{R,j} - X_{C,j}| + \delta \max_i \max_j |X_{R,j} - X_{C,j}|}{|X_{R,j} - X_{C,j}| + \delta \max_i \max_j |X_{R,j} - X_{C,j}|} \quad (4)$$

where  $X_R$  and  $X_C$  are referencing and comparing variables,  $j$  is a sequence of each variable, and  $\delta$  is the resolving coefficient, generally assumed to be 0.5<sup>33</sup>. If the calculated value is close to +1 or -1, the relationship shows a high correlation.

**Cost function.** The cost function indicates the performance of ML and is calculated as the square of the difference between the predicted value ( $H$ ), calculated by weight ( $w$ ) and bias ( $b$ ), and the actual value ( $Y$ ), as shown in Eq. (5). The cost function shows a quadratic parabolic distribution according to  $w$  and  $b$ . The point at which the slope of the parabola is at a minimum shows the best performance, so the partial derivative is performed at each point of the parabola, as shown in Eq. (6).

$$\text{cost}(w, b) = \frac{1}{m} \sum_{i=1}^m (H(x_i) - Y_i)^2 \quad (5)$$

$$w', b' = w, b - \eta \frac{\partial}{\partial w, \partial b} \text{cost}(w, b) = w, b - \eta \frac{\partial J_n}{\partial a_s^{(L+1)}} \cdot \frac{\partial a_s^{(L+1)}}{\partial z_s^{(L+1)}} \cdot \frac{\partial z_s^{(L+1)}}{\partial w_{s,r}^{(L)}, \partial b_{s,r}^{(L)}} \quad (6)$$

where  $w'$  and  $b'$  indicate the weight and bias from which the error has been removed; and  $\eta$  is the learning rate, which is the degree of the learning step as a partial differential interval and is generally 0.001<sup>34</sup>. The cost function can be subdivided into weight and bias at corresponding nodes through the chain rule.  $L$  denotes the number of the hidden layer, and  $s$  and  $r$  indicate the node numbers of and before the  $L$  layer.  $Z_s$  is the sum of weights and biases corresponding to each node, and  $a_s$  is the value of the activation function that is applied to  $Z_s$ ,  $J_n$  is the Jacobian matrix of the error computed at the hidden layer. A method of using partial differentiation at a certain learning rate is called gradient descent, and this is a commonly used method to find  $w$  and  $b$  in ML. However, this method has a limitation, the reliability of the cost function depends on the learning rate interval. Methods to overcome this have been proposed. In this study, Levenberg–Marquardt and Bayesian regularization algorithms were additionally considered to improve reliability.

**Levenberg–Marquardt algorithm.** The Levenberg–Marquardt (LM) algorithm is suitable for solving nonlinear least square problems. Marquardt developed the algorithm in 1963 based on a theory suggested by Levenberg in 1944. This method estimates an optimal  $w$  and  $b$  based on the Gauss–Newton method and can provide the solution by locally approximating the nonlinear function as a linear function by generalizing the multivariate vector. LM can flexibly change the learning rate without fixing a value. A small interval of the learning rate is given when the solution is stably converged, and a large interval value is set in the opposite case. Therefore, it is an advantage to compensate for a divergent differential value due to a fixed learning rate in gradient descent. The slope of the cost function based on LM ( $\text{slope}_{LM}$ ) is calculated by the Jacobian matrix ( $J$ ), as shown in Eq. (7).

diag (JrT)r, which is an intrinsic value of the Hessian matrix, represents the curvature of the function and is used to determine the magnitude of the learning rate<sup>35</sup>:

$$\text{slope}_{LM} = (J^T J + \eta \text{diag}(J^T J))^{-1} J^T r(p) \quad (7)$$

where  $r(p)$  is the error of the model, and  $\text{diag}$  is the diagonal matrix.

**Bayesian regularization (BR).** Bayesian regularization (BR) is a method proposed by MacKay<sup>36</sup> that uses a Hessian matrix in the same way as the LM algorithm. However, the weight is determined according to a random variable by adding Bayesian theory<sup>37</sup>. BR is mathematically expressed as a probability density function with weight ( $w$ ), dataset ( $D$ ), function parameters ( $\alpha, \beta$ ), and a particular model ( $M$ ), as shown in Eq. (8):

$$P(w|D, \alpha, \beta, M) = \frac{P(D|w, \beta, M)P(w|\alpha, M)}{P(D|\alpha, \beta, M)} = \frac{\exp(-(\beta E_D + \alpha \sum_{k=1}^l \sum_{i,j=1}^m (w_{ij}^k)^2))}{(\int \exp(-\alpha \sum_{k=1}^l \sum_{i,j=1}^m (w_{ij}^k)^2) dw) (\frac{2\pi}{\beta})^{\frac{N}{2}}} \quad (8)$$

where  $P(D|w, \beta, M)$  is the possibility that the weights, assumed by the likelihood function, can appear, and  $P(w|\alpha, M)$  shows the prior density representing the weight of the previous step.  $P(D|\alpha, \beta, M)$  is a normalization factor; the total probability of occurrence can be finalized to 1. Each probability can be extensively expressed with a sum of squared error ( $E_D$ ) through Gaussian noise and weight based on the maximum posterior probability deduced to the optimal value.

**Performance.** Studies based on the ML technique generally use the root mean square error (RMSE) and mean square error (MSE) to find the performance-comparing quantitative value<sup>38–41</sup>. In this method, RMSE and MSE were also used to verify the reliability of each optimization method. The mathematical expressions of RMSE and MSE are as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Z(s_i) - \hat{Z}(s_i))^2} \quad (9)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n |Z(s_i) - \hat{Z}(s_i)| \quad (10)$$

where  $Z(s_i)$  and  $\hat{Z}(s_i)$  denote actual and calculated safety factors, and  $n$  is the amount of data.

**Geostatistical method.** The kriging method is used to obtain the value at an unknown point considering spatial variability through an interpolation technique and has been widely adopted for its advantage of reducing error variance<sup>42</sup>. The error variance ( $\sigma(P_0)$ ) at point ( $P_0$ ) is defined by Eq. (11) with semi-variance ( $\gamma(P_n, P_0)$ ) between the measured value at point ( $P_0$ ) and the predicted value at point ( $P_n$ ), and Lagrange multiplier ( $\psi(P_0)$ ).

$$\sigma^2(P_0) = \sum_{n=1}^n \lambda_n \gamma(P_n, P_0) + \psi(P_0) \quad (11)$$

where  $\lambda_n$  denotes the weighting factor.

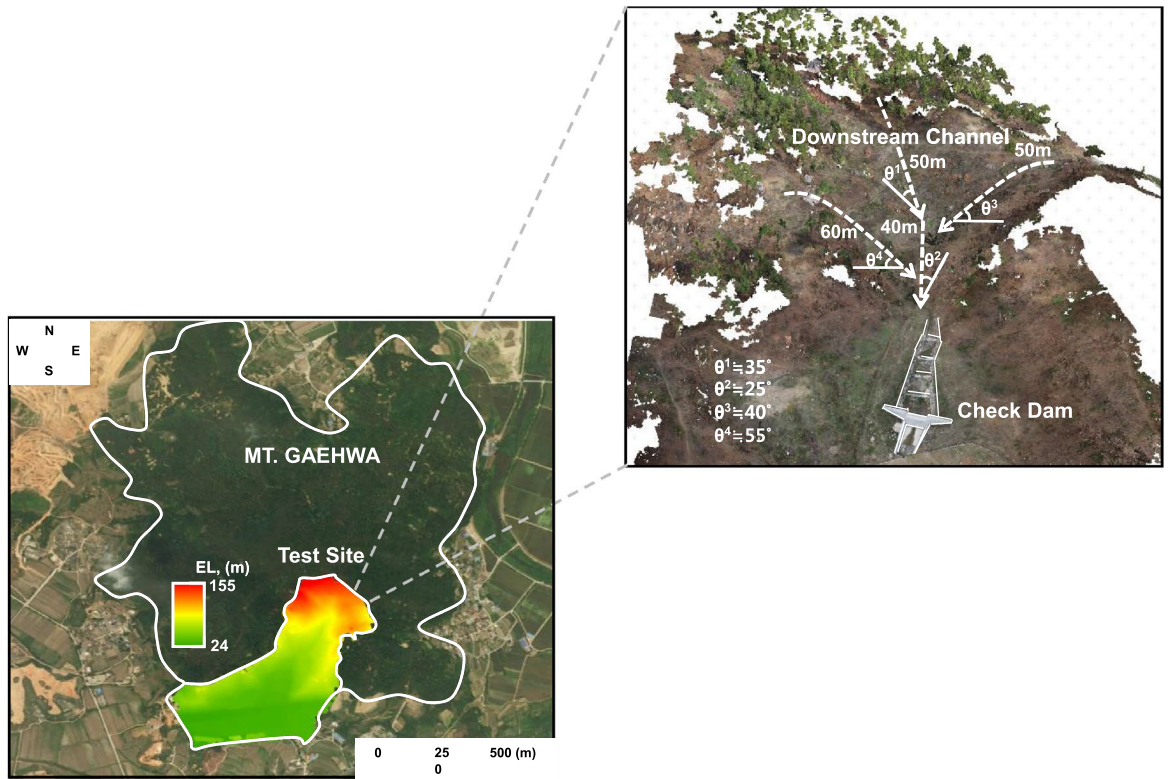
The variogram, which can reflect the variance of  $\lambda_n$  between  $P_0$  and  $P_n$ , was defined with a nugget, sill, and range to minimize the error. The nugget and sill are the non-zero variogram and maximum amplitude, respectively. The range is the distance between the variogram and the sill.

## Methodology

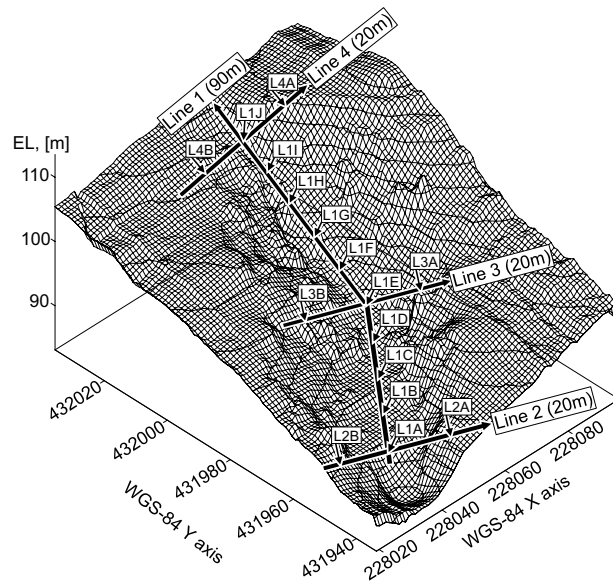
**Site description.** Gaehwa Mountain, located in the city of Sejong, South Korea, at 36° 29' 10.77" N–36°29'07.65" " N latitude and 127°18'46.03" E–127°18'47.73" E longitude, was selected as the experimental site because debris flow occurred in this area about five years ago. The whole view of this slope is shown in Fig. 2, and the range of the digital elevation model, derived by aerial surveying, is approximately 24–155 m. The study area is composed of a main stream 90 m in length and two branch streams 60 and 50 m in length, which are connected downstream of the main stream on the left and right. The slope of the main stream is about 25–35°, and the slope of the branch streams is relatively high, about 40–55°. In addition, a large amount of soil was deposited downstream due to the debris flow, and a check dam had already been installed to prevent disasters.

**Experiment.** Field and laboratory experiments were performed to obtain the input parameters of the empirical equation. Field experiments were conducted with a seismic survey, an electrical resistivity survey, a dynamic cone penetration test (DCPT), and time domain reflectometry (TDR). The disturbed samples were also extracted to perform sieve analysis in the laboratory. The seismic and electrical resistivity surveys used four profiles, including main and branch streams, as shown in Fig. 3. The geophone spacing for the seismic survey was set to 2 m, and the commonly used drop hammer was used as the source. Electrodes for measuring electrical resistivity were installed at the same intervals as the seismic survey, and the Wenner array method was applied. The

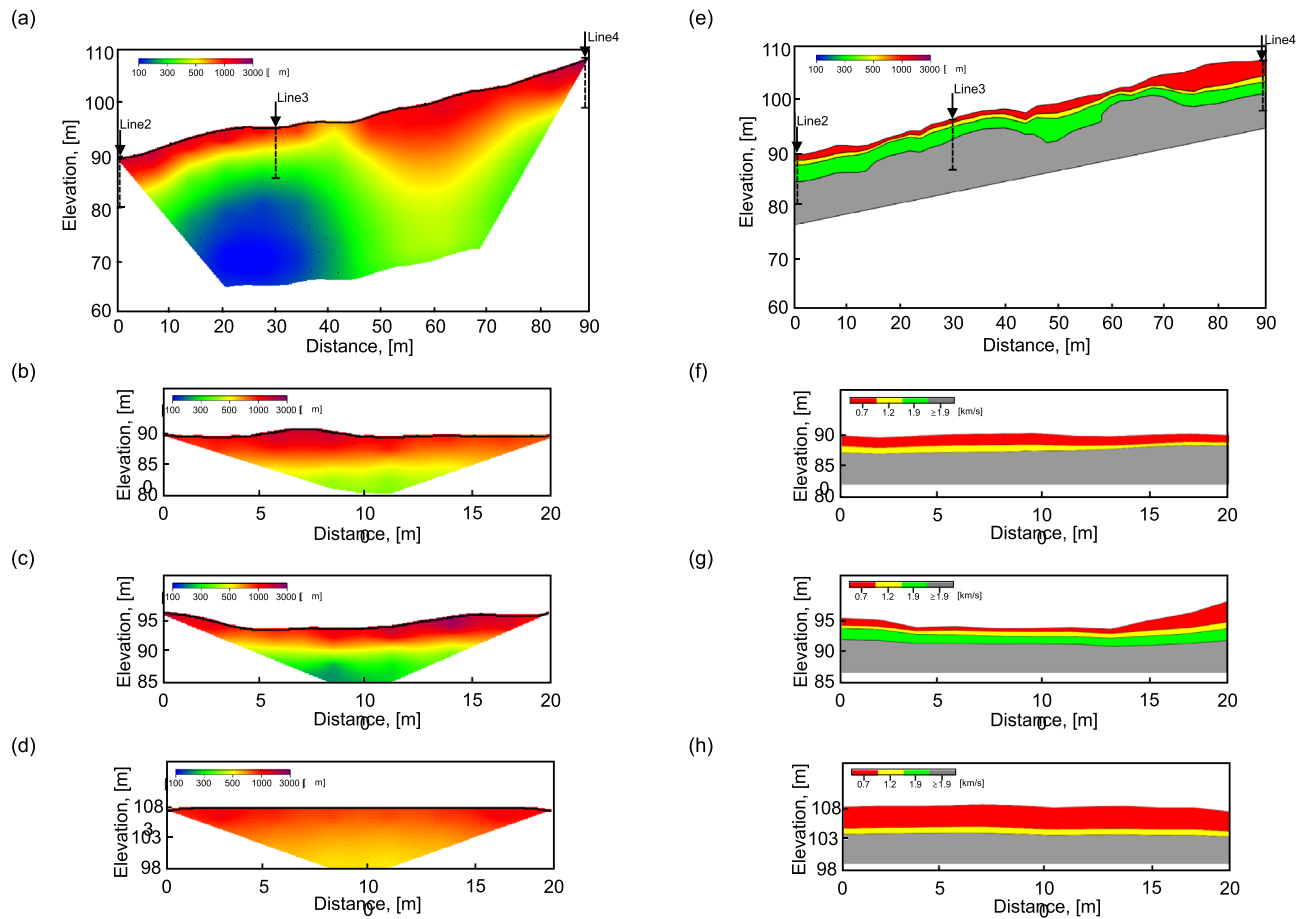




**Figure 2.** Site description. The picture of left side was captured through <https://map.kakao.com> and the detailed topography, placed on right side, was directly obtained by on-site geological survey with UAV survey. The agisoft viewer was applied to find the topography based on measured digital elevation model.



**Figure 3.** Experimental profiles. Lines 1, 2, 3, and 4 denote seismic and resistivity survey profiles. Arrows show points for performing penetration tests and extracting soil; arrow interval is 10 m. The only grid was built from software of Surfer. WGS: World Geodetic System.



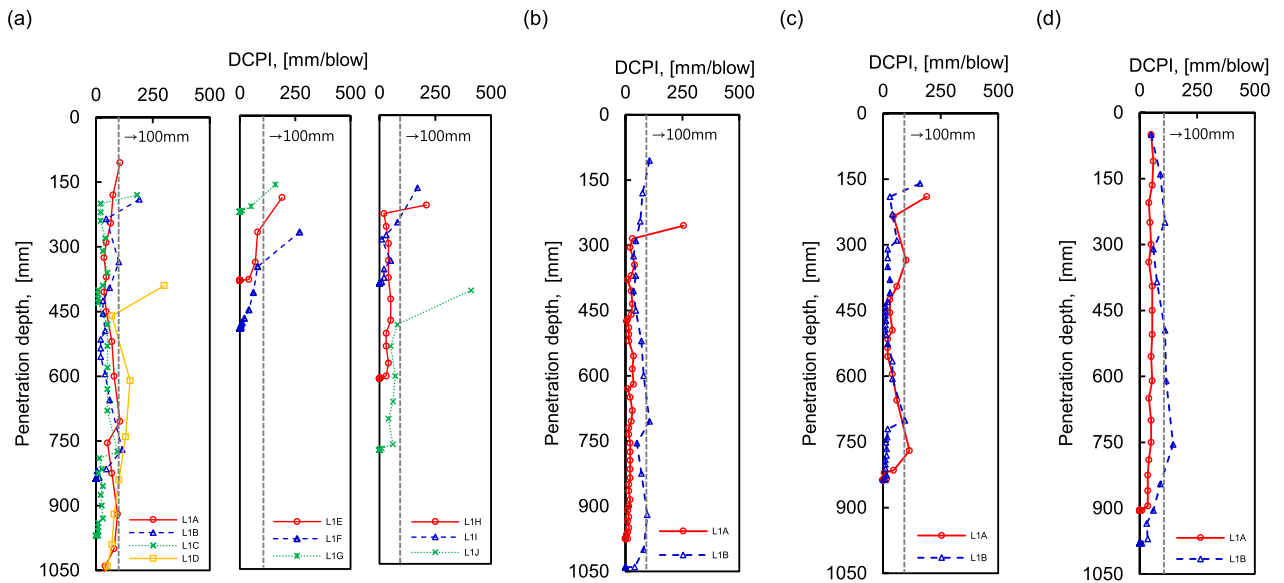
**Figure 4.** Geophysical survey results: (a) line 1; (b) line 2; (c) line 3; (d) line 4 for electrical resistivity (e) line 1; (f) line 2; (g) line 3; (h) line 4 for elastic wave.

detailed contents of seismic and electrical resistivity surveys were replaced with a reference<sup>43</sup>. DCPT and TDR were performed at about 10 m intervals at the same locations at which the geophysical survey was conducted, and 16 in situ tests were performed. The principle and experimental procedure of the DCPT and TDR are based on the following papers<sup>44–46</sup>. In addition, soil sampling was carried out in the same area as the in situ exploration. The sieve analysis was performed with a procedure suggested by ASTM<sup>47</sup>. The field test and sampling location based on world geodetic system (WGS) coordinates are shown in Fig. 3.

## Results

**Field and laboratory test results.** A cross-sectional diagram of resistivity is shown in Fig. 4a–d, and the relatively high range (over  $\approx 1000 \Omega\text{m}$ ) is distributed from the surface to a depth of about 10–15 m. High electrical resistivity was strongly demonstrated in the upper area of the slope (around Line 4) due to the high altitude. In the lower part of the slope, high amounts of groundwater can be collected due to the geological characteristics of the deep valley, as shown in Figs. 2 and 3. Therefore, a low electrical resistivity of about  $100 \Omega\text{m}$  appeared at a depth of 10 m at a slope distance of 20–40 m. The electrode spacing of lines 2, 3, and 4 was the same as that of line 1, but the surface depth was relatively less than that of line 1 due to the shorter distance used for performing the experiment. The results of lines 2, 3, and 4 are similar to those of line 1, with a high electrical resistivity of around  $1000 \Omega\text{m}$  on the surface. Low electrical resistivity, which was similar to line 1 at an 85 and 90 m depth, was deduced at a distance of 5–15 m of lines 2 and 3. The results of electrical resistivity were used to convert porosity and hydraulic conductivity into input parameters in the empirical equation. Archie's law was used to obtain porosity<sup>48,49</sup>, and the cementation factor ( $m$ ) and tortuosity factor ( $\alpha$ ) were set to 1.3 and 1 with reference to previous studies<sup>50,51</sup>. Hydraulic conductivity was deduced by the Kozeny-Carman equation through porosity, derived by Archie's law, and the effective diameter, derived by sieve analysis<sup>46</sup>.

The result of seismic wave velocity is shown in Fig. 4e–h. The strata were divided into a colluvium layer, 0.7 km/s, sedimentary rock layer, 1.2 km/s; sedimentary rock layer, 1.9 km/s; and a soft rock layer, 1.9 km/s or more, based on the literature<sup>52–54</sup>. The colluvium layer deposited on the surface was the most affected stratum by debris flow, and the thickness was estimated at 0.3–3 m in lines 1, 2, 3, and 4. Even though it is difficult to systematically compare electrical resistivity and seismic wave results, since the properties that affect each technique are different, the upper (0–10 m) and lower (70–90 m) distances of the slope were thick, and the middle distance (20–40 m) was thin at Line 1 with similar electrical resistivity. The thickness of the surface at lines 2, 3, and 4 also shows a



**Figure 5.** DCPT results: (a) line 1; (b) line 2; (c) line 3; (d) line 4.

similar trend to the that of areas with high electrical resistivity ranges. The distribution of elastic wave velocity at each location was used to calculate the elastic modulus as an input parameter in the empirical equation<sup>55–58</sup>.

The penetration depth of the DCPT at each line was plotted as shown in Fig. 5, and the results of line 1 were divided into lower (L1A to L1D), middle (L1E to L1G), and upper (L1H to L1J) parts of the slope. The initial penetration depth was more than 100 mm in the lower, middle, and upper parts in Line 1, and lines 2 and 3 also show the same results. Note that a large amount of soil was moved from the upper part to the ground due to debris flow, so a large penetration depth was recorded at initial impact. Therefore, it is assumed that the surface is weak. The initial penetration depth of line 4 was about 50 mm on average, which is smaller than the depth obtained by other lines, indicating relatively stiff ground. The reason for this is that the soil of the initial zone for the occurrence of debris flow, in the upper area, had already moved to the lower part, so the strength of the surface is high. The DCPT test results were used to estimate soil thickness as an input variable, and the experiment was performed until the penetration depth was  $\approx 0$  mm<sup>44,45,59</sup>. The final penetration depth was recorded as approximately 700, 1000, 800, and 950 mm on average for lines 1, 2, 3, and 4, respectively. The shear strength in the empirical equation was also estimated through the results of the DCPT, and the correlation equation, which was calculated in similar geological conditions of weathered soil on the slope, was applied<sup>60</sup>.

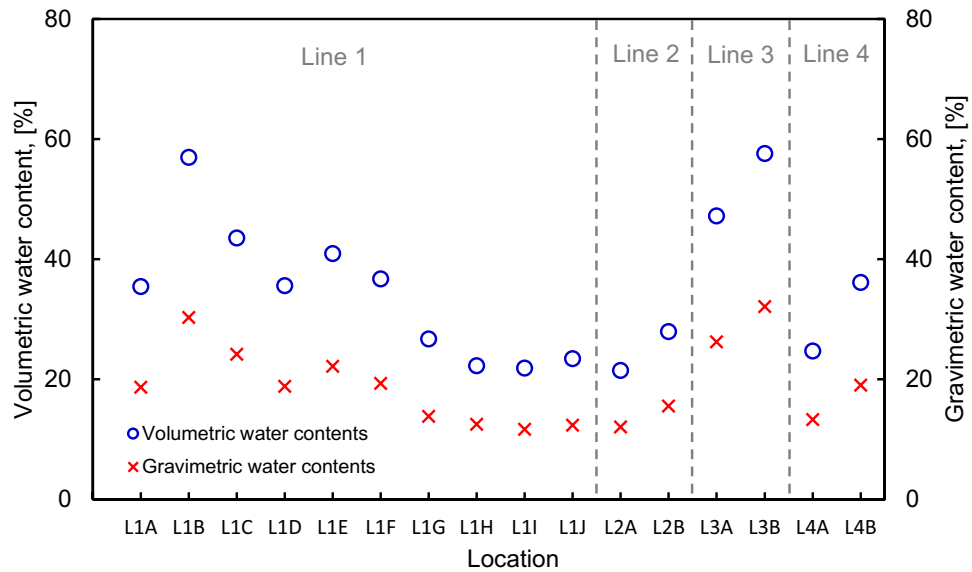
Figure 6 shows the volumetric water content measured by TDR and the gravimetric water content derived from laboratory experiments through extracted samples, which were applied to the input parameters in the empirical equation. L1A–L1F, L3A, and L3B, which are catchment basins formed by steep slopes, showed relatively high water content, the ranges of volumetric and gravimetric water content were 35–57% and 18–32%. For the opposite reason, the ranges of volumetric and gravimetric water content were found to be relatively low at 21–36% and 11–19% in L1G–L1J, L4A, and L4B, which are the upper areas of the slope. The volumetric water content was higher than the gravimetric water content under a three-phase soil system, and the difference was mainly due to the soil bulk density, assuming that the water density is uniform. The difference between volumetric and gravimetric water content was large in the catchment areas at L1A–L1F, L3A, and L3B (approximately  $1.53 \text{ g cm}^{-3}$  on average). The reason for this difference is that the weight per unit volume was increased due to the mixed fine particles from movement into the lower part of the slope.

Figure 7 shows the results of sieve analysis, which were deduced from well-graded soil (SW), based on the unified soil classification system (USCS). The fine content was estimated as the amount that passed through a #200 sieve, the catchment areas of L1A, L1C, and L2A and the initial area of L4B showed high values of 7.8%, 8.7%, 7.5%, and 7.7%, respectively. The fine content in this area was calculated as 6.3% on average, as shown in Table 1. The coefficients of uniformity ( $C_u$ ) and curvature ( $C_c$ ) are also shown in Table 1. All values show that  $C_u$  is greater than 4 and  $C_c$  is distributed in the range of 1–3 as a well-graded particle.

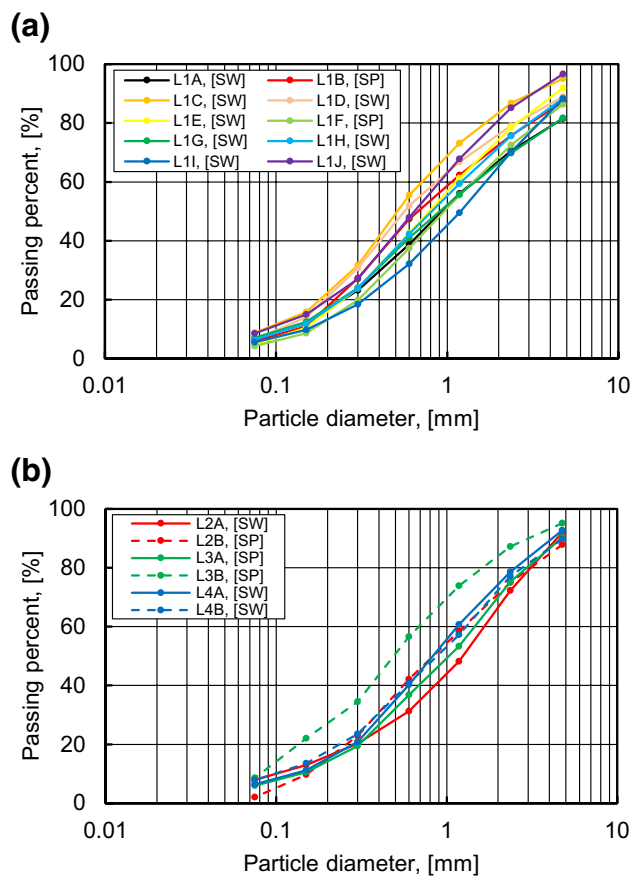
Finally, the eight input parameters in the empirical equation were derived through the previous experiments. Porosity and hydraulic conductivity were obtained by converting electrical resistivity, and elastic modulus was derived from elastic wave velocity. DCPI was used to estimate shear strength and soil thickness, and saturation was used as the volumetric water content converted to TDR. In addition, gravimetric water content and fine content were obtained through laboratory tests.

**Dataset.** The objective area was divided into 1800 grid cells of 1 m each to examine the relationship between dependent and independent variables considering spatial variability. The variables were set to the eight factors of the empirical equation and safety factor to perform deep learning, including LR, an NN, and an RF. Datasets of dependent and independent variables were constructed at the experimental grids. The dependent variable





**Figure 6.** Distributions of volumetric and gravimetric water contents.



**Figure 7.** Grain-size distribution curves: (a) line 1; (b) line 2, 3 and 4.

was calculated with Eq. (1) and soil cohesion, slope angle, soil density, and friction angle were determined referring to previous results in the same area<sup>61</sup>. The vertical water table height and root cohesion were fixed at 0.5 m and 100 N m<sup>-2</sup>, respectively, the same as the values used in the previous study<sup>61</sup>. For soil depth, the same value as the empirical equation based on the field experiment was applied. However, the field penetration and laboratory tests were conducted on only 16 positions, thus, more data were required to perform LR, an NN, and

Location		Fine content (%)	Coefficient of uniformity ( $C_u$ )	Coefficient of curvature ( $C_c$ )
Line 1	L1A	7.8	15.0	1.1
	L1B	5.8	8.5	0.7
	L1C	8.7	8.2	1.3
	L1D	6.1	8.5	1.1
	L1E	4.4	7.3	0.9
	L1F	4.1	9.4	0.8
	L1G	6.9	14.3	1.0
	L1H	6.3	10.9	1.2
	L1I	5.3	12.0	1.1
	L1J	8.3	10.0	1.5
Line 2	L2A	7.5	17.8	2.2
	L2B	1.8	8.0	0.8
Line 3	L3A	5.8	10.0	0.9
	L3B	7.3	8.8	0.9
Line 4	L4A	6.7	8.0	1.1
	L4B	7.7	15.0	1.1

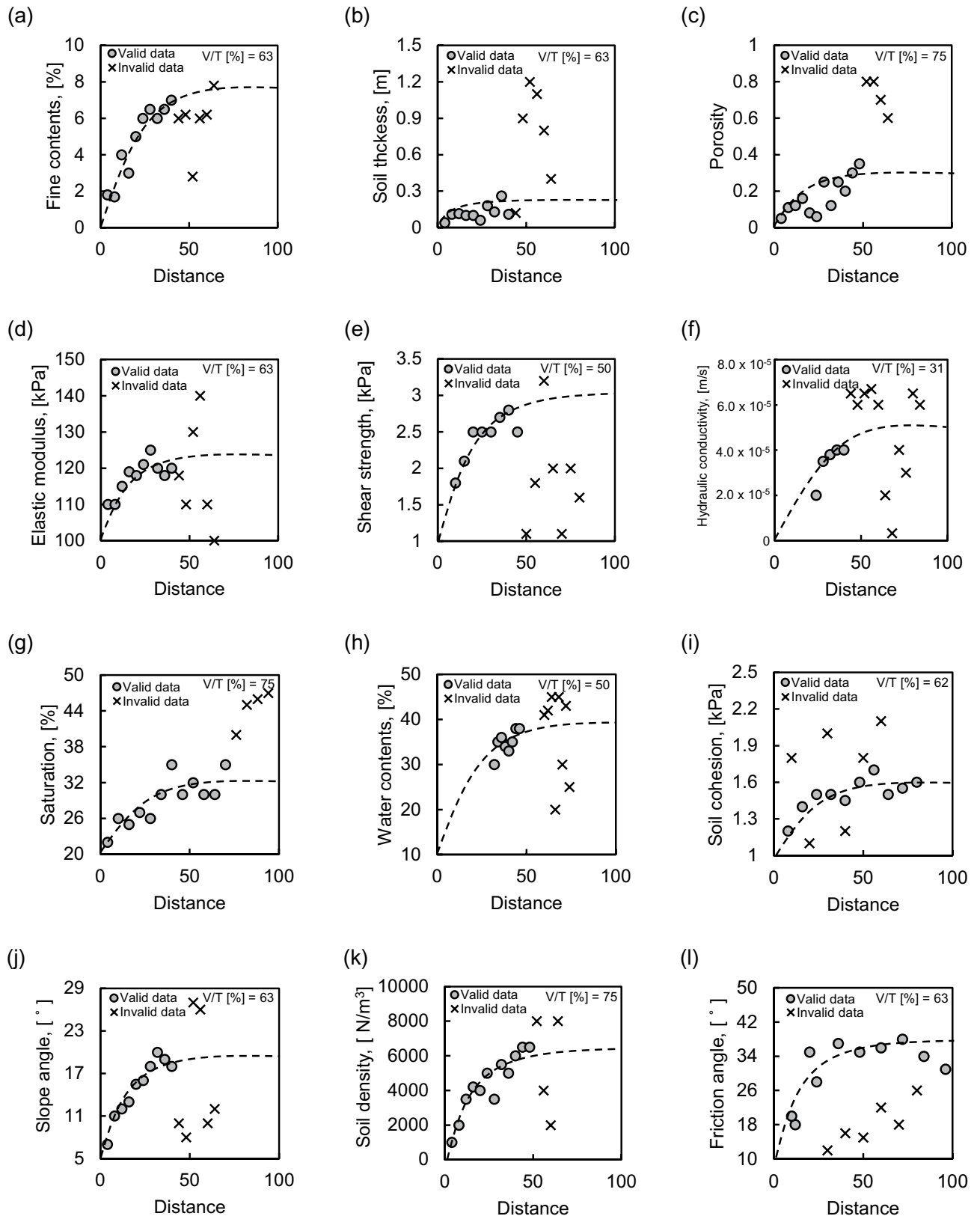
**Table 1.** Detailed sieve test results.

	Nugget effect	Sill	Type	Range
Fine content	0	2	Spherical	16
Soil thickness	0	0.1	Spherical	9.6
Porosity	0	0.01	Spherical	12.8
Elastic modulus	0.0001	0.0002	Spherical	80
Shear strength	30,000	120,000	Gaussian	92
Hydraulic conductivity	0	5.5E-09	Spherical	80
Saturation	0	300	Spherical	130
Water content	0	80	Spherical	132
Soil cohesion	0	22,000	Spherical	75
Slope angle	0	20	Spherical	38
Soli density	0	4000	Spherical	12
Friction angle	8	30	Spherical	50

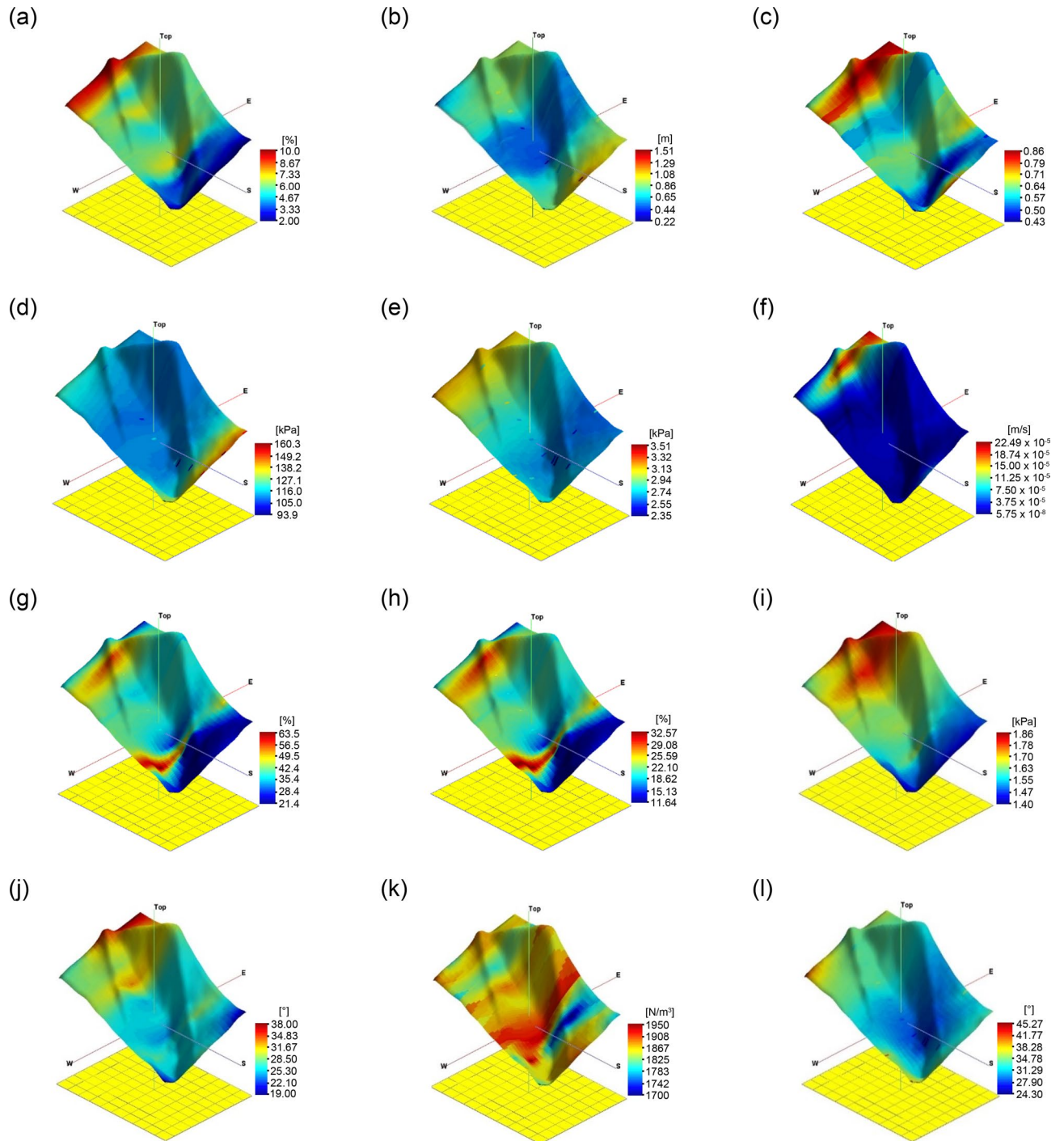
**Table 2.** Factors of variogram in each input variable.

an RF. Therefore, the geostatistical technique, a specialized interpolation method with kriging, was applied to determine the dependent and independent variables of the remaining grids<sup>62</sup>. The sill, nugget, and range values were derived from the distribution of each dataset based on geostatistics, and the constant values are shown in Table 2. The Stanford Geostatistical Modeling Software (SGeMS) beta (V2.5b), which is an open-source computer package for solving spatially related variables, was used to perform kriging analysis. The resolution of a variogram consisting of a sill, nugget, and range is known to be superior, as the nugget-to-sill ratio is smaller<sup>63</sup>. Among exponential, spherical, and Gaussian models, spherical and Gaussian models with the smallest nugget-to-sill ratio were selected. Note that the spatial distributions of the measurement result well correlated with the spherical and Gaussian models. Figure 8 shows 12 variograms, including 8 and 4 input parameters for the empirical equation and Eq. (1), respectively. Among the 16 tests, the data used for kriging in accord with the variogram were expressed as valid data, and the rest were indicated as invalid data. The valid data to total data ratio was calculated to be about 31–75%. Porosity, saturation, and soil density showed the highest ratio, while hydraulic conductivity showed the lowest ratio. The reason for this is that the spatial variability in this area is different for each parameter. The 3D kriging result based on the derived variogram is shown in Fig. 9, showing the distribution of each parameter. Finally, 1800 datasets including artificial data based on the geostatistical technique were constructed.

**Machine learning results.** *Linear regression.* The multi-LR model was applied to find the relationship between independent and dependent variables with gradient descent, Levenberg–Marquardt (LM), and Bayesian regularization (BR). The flow chart of performing linear regression was expressed in Fig. 10. The ratio of training and test data was determined to be 7:3 to perform cross-validation based on the hold-out validation technique, thus, 1260 and 540 pieces of data were randomly used for training and test procedures, respectively. The MSE value according to the validation ratio is shown in Fig. 11, and the 7:3 ratio showed the smallest value.

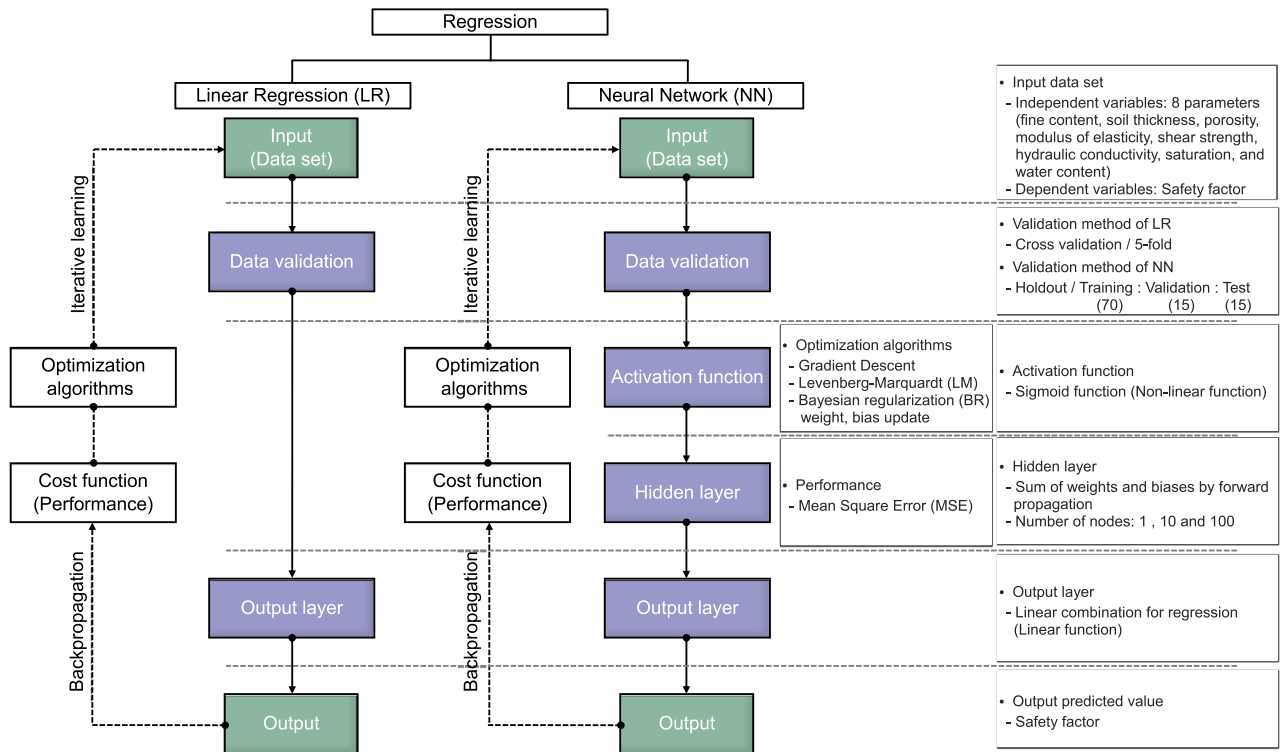


**Figure 8.** Variograms of input variables: (a) fine content; (b) soil thickness; (c) porosity; (d) elastic modulus; (e) shear strength; (f) hydraulic conductivity; (g) saturation; (h) water content; (i) soil cohesion; (j) slope angle; (k) soil density; (l) friction angle.

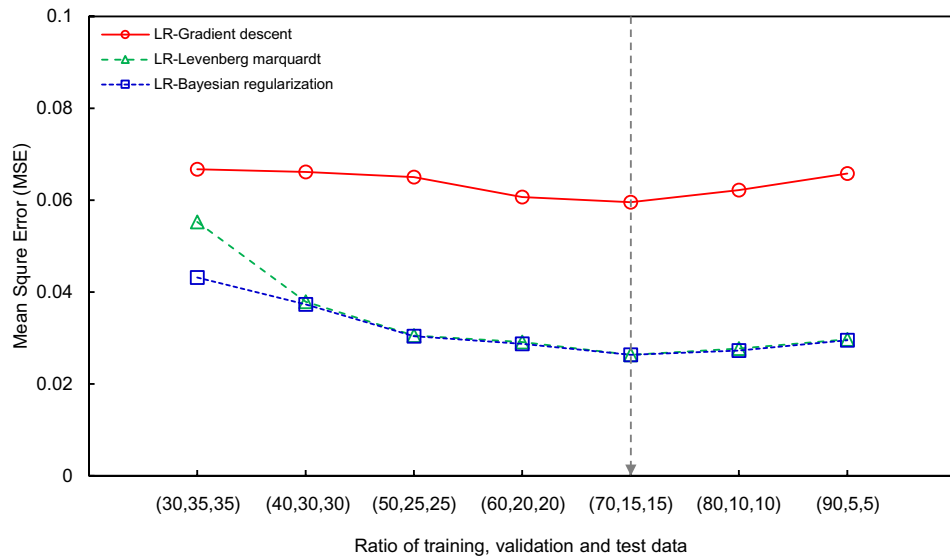


**Figure 9.** Kriging results of input variables: (a) fine content; (b) soil thickness; (c) porosity; (d) elastic modulus; (e) shear strength; (f) hydraulic conductivity; (g) saturation; (h) water content; (i) soil cohesion; (j) slope angle; (k) soli density; (l) friction angle. The software of SGeMS beta (v.2.5b) was applied to draw this figure.

The learning rate and epoch were set to 0.001 and 1000, respectively. The comparison between the actual and predicted safety factor based on LR is addressed in Fig. 12. In the gradient descent algorithm, the relationship showed almost linear behavior when it was lower than  $\approx 1.8$ , but the safety factor was underestimated at over  $\approx 1.8$ . The actual safety factor showed a linear relationship with the predicted safety factor when using LM and BR algorithms. Therefore, the coefficient of determination ( $R^2$ ) was 0.96 for training and test data in LM and BR. Although BR has a slightly higher  $R^2$  than LM, they both provide more reliable results than the gradient descent algorithm. In addition, root mean square error (RMSE) and mean square error (MSE) were calculated through Eqs. (9) and (10) to verify the reliability of the artificial dataset dividing experimental and interpolation data, and the results are shown in Fig. 13. The average RMSE and MSE values based on interpolation were 13–22% and 4–69% smaller, respectively, than those based on the experimental data in all algorithms. This result shows that



**Figure 10.** Flow chart of linear regression and the neural network.



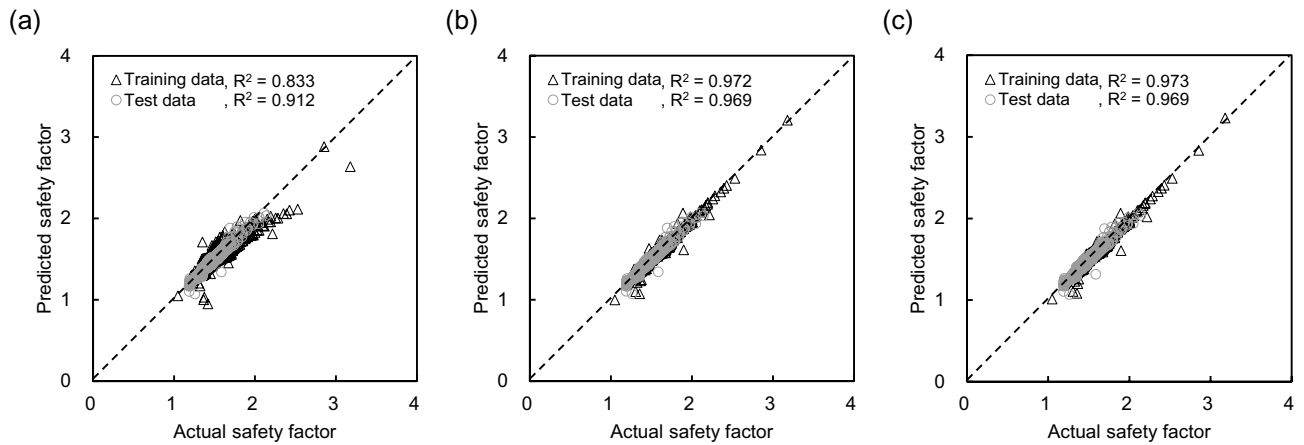
**Figure 11.** MSE according to selected ratio in linear regression.

the reliability of 1800 pieces of data, constructed by the interpolation method, is high and there are no outlier data. In addition, BR shows lower RMSE and MSE than any other algorithm with the same trend of  $R^2$ . Note that BR is the most reliable algorithm, the relationship between dependent and independent variables with BR based on the multi-LR model is shown in Eq. (12):

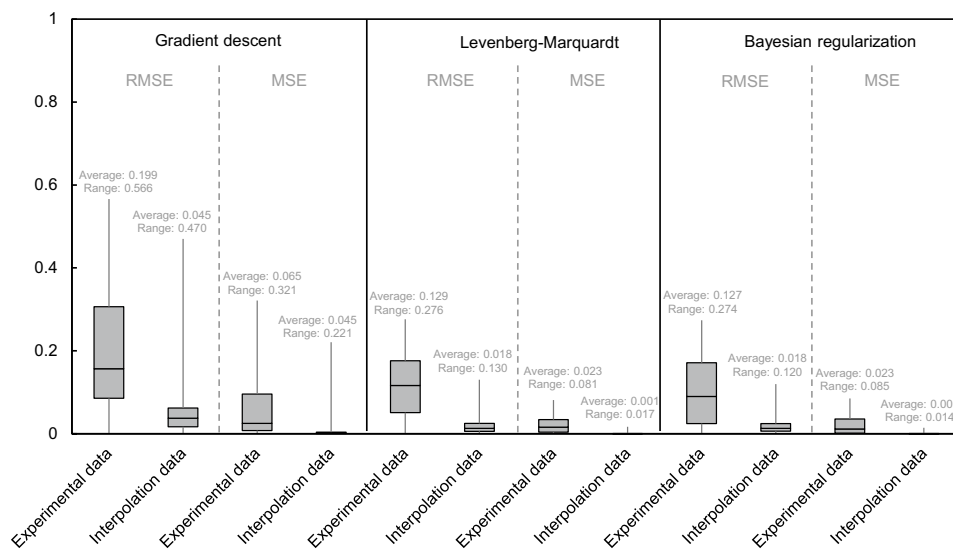
$$Y = 13.9362 + 0.3744 \cdot X_1 - 0.9758 \cdot X_2 + 6.3219 \cdot X_3 - 0.1350 \cdot X_4 - 5.0673 \cdot X_5 - 0.003072 \cdot X_6 + 1.5053 \cdot X_7 - 3.0019 \cdot X_8 \quad (12)$$

where  $X_1, X_2, X_3, X_4, X_5, X_6, X_7,$  and  $X_8$  denote the fine content, soil thickness, porosity, the modulus of elasticity, shear strength, hydraulic conductivity, saturation, and moisture content as independent variables.  $Y$  represents the safety factor as the dependent variable.





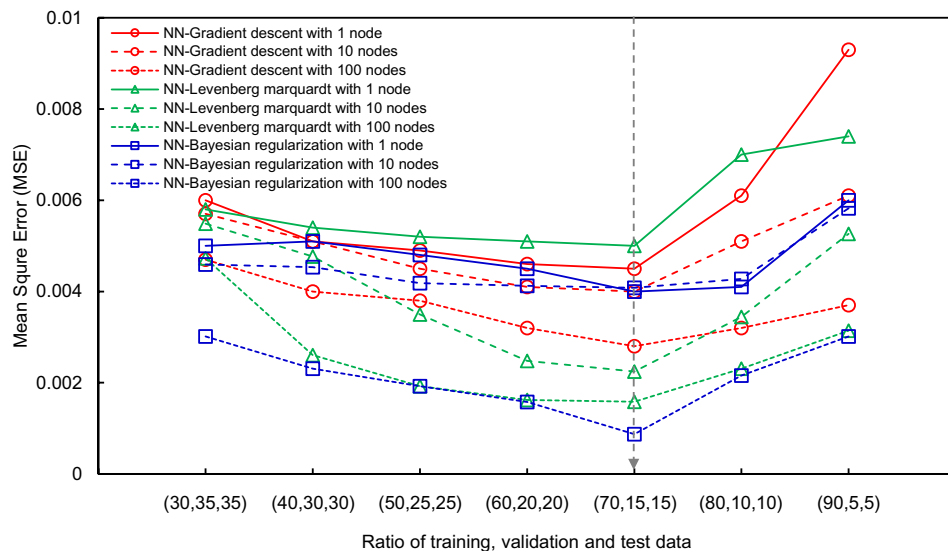
**Figure 12.** Comparison between the predicted and actual safety factors based on linear regression: (a) gradient descent algorithm; (b) Levenberg–Marquardt algorithm; (c) Bayesian regularization algorithm.



**Figure 13.** Distribution of root mean square error (RMSE) and mean square error (MSE) based on linear regression.

**Neural network.** The eight input factors and safety factors were designated as independent and dependent variables, respectively, as multi-LR, and the relationships between two variables were then investigated by applying a neural network. The procedures of regression based on an NN was plotted in Fig. 10. The NN focused on the nonlinear relationships between variables, and a sigmoid function was applied. The ratio between training and test data was 7:3 as multi-LR, and the reliability of prediction was improved with cross-validation. The MSE value according to the validation ratio in the NN is shown in Fig. 14. The learning rate and the number of learning iterations were set to 0.001 and 1000, respectively. The number of nodes and the optimization method were changed as hyperparameters, and the number of nodes was divided into 1, 10, and 100. The gradient descent, LM, and BR of the optimization method were selected in the same way as LR.

The relationship between actual and predicted safety factors is compared in Fig. 15 and was classified with the optimization method and the number of nodes. As the number of nodes increases,  $R^2$  increases in all optimization methods and the linearity is excellent. When the number of nodes is 100,  $R^2$  is 0.945, 0.995, and 0.992, on average, in gradient descent, LM, and BR, respectively. In addition, the  $R^2$  of LM and BR was larger than the  $R^2$  of gradient descent, similar to the result of multi-LR. With the NN, LM showed excellent performance with a slight difference. RMSE and MSE values of each optimization method according to the number of nodes are shown in Fig. 16. Similar to the  $R^2$  result, the calculated value tends to decrease as the number of nodes increases, and LM and BR show relatively smaller values than gradient descent. In addition, LM has a smaller range than BR when there are 100 nodes. Therefore, LM was selected as the best optimization method, and the number of nodes was fixed to 100 to operate the NN.



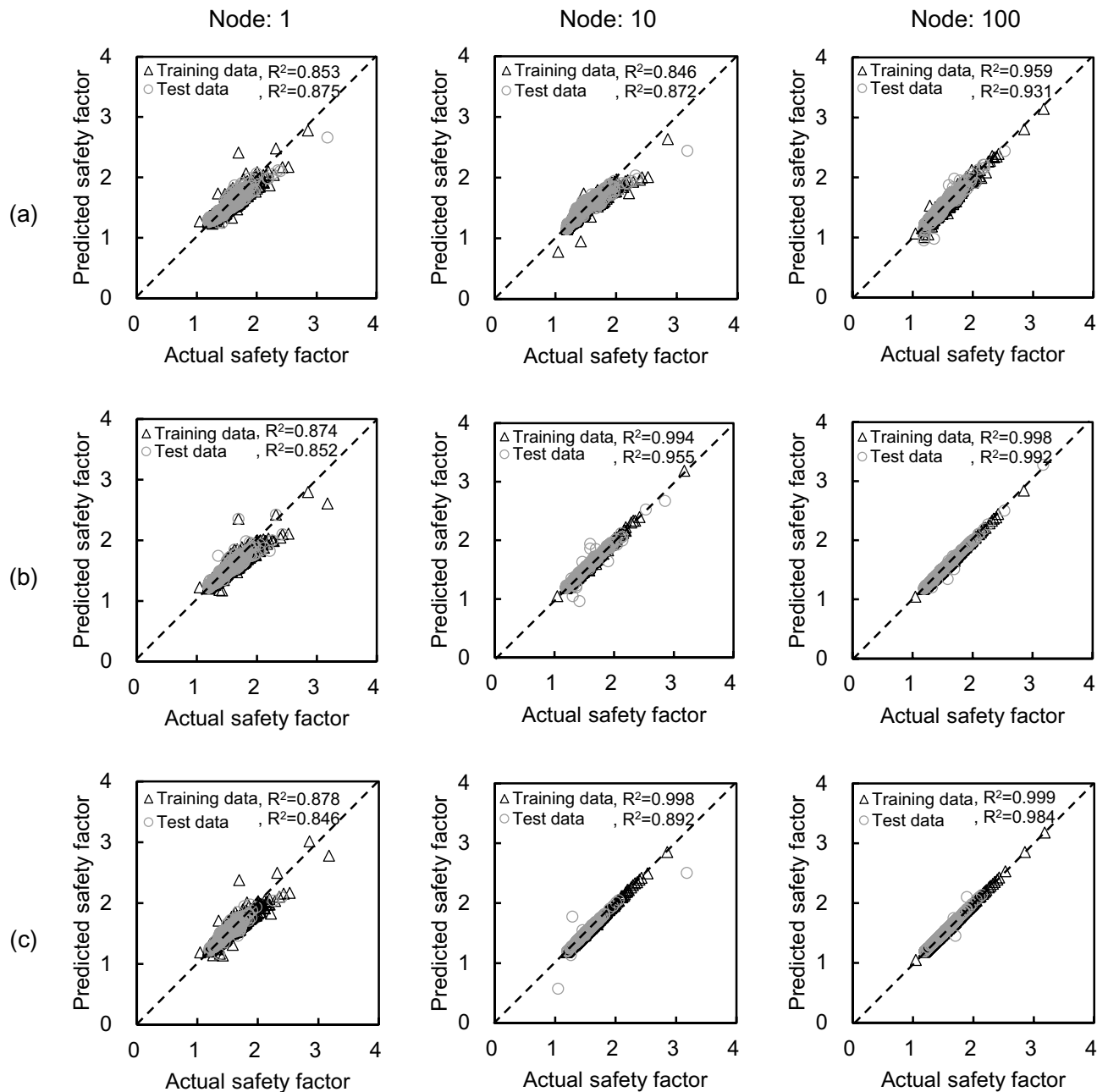
**Figure 14.** MSE according to selected ratio in neural network.

**Random forest.** The random forest (RF) algorithm was applied to estimate the importance of the eight independent variables. The random variable ( $m$ ) was determined as 3, referring to a previous study selecting at least one-third of the independent variables<sup>64</sup>. The flowchart of RF is summarized in Fig. 17 and the number of decision trees was considered by calculating the OOB error according to the number of trees, as shown in Fig. 18. The error (shown in Fig. 18) was found to have converged from about 20 decision trees or more, so the decision tree was set to 20. The importance of each factor was calculated as a score for predicting the safety factor through Eq. (3), and the result is shown in Fig. 19. Based on the quantitative score, the parameters with a strong effect on the safety factor were shear strength (0.380 point), soil thickness (0.225 point), elastic modulus (0.221 point), fine content (0.056 point), hydraulic conductivity (0.053 point), porosity (0.028 point), water content (0.020 point), and saturation (0.017 point). Additionally, the correlations between factors were calculated by Eq. (4) and plotted as GRG in Fig. 20. GRG shows the relationships between variables, and results close to +1 (proportion) or -1 (inverse proportion) mean a high correlation. Note that the values in the diagonal direction, which are the results of comparing the same variable, are all +1, and values based on +1 to the left and right are the same. The variables with the highest positive correlation are saturation and water content, because they represent the characterization of fluid. Among the positive correlations, the smallest value is 0.27 for shear strength and soil thickness. The closest value to -1 (inverse proportion) is -0.54, which is the relationship between saturation and elastic modulus, and the GRG of hydraulic conductivity and elastic modulus is -0.073, indicating a relatively small relationship. Saturation and water content are highly correlated and thus important, as shown in Fig. 20, with a low score because they can be linked.

## Discussion

Contour maps of the safety factor are plotted in Fig. 21 through LR and NN models with eight variables. Note that the contour map can accurately check the distribution of the safety factor while connecting the same safety factors. In addition, the safety factor calculated by Eq. (1) is shown to be a true value in Fig. 21a. Note that the overall distribution of the safety factor based on LR and NN models is mostly similar to the contour map of the safety factor deduced by Eq. (1). In the linear regression, the distribution of the safety factor at the top and bottom of the slope is more widely spread at a constant value ( $FS = 1.0$ ), and the color is darker blue. It is also distributed at a constant value ( $FS = 2.0$ ) in the middle of the slope, thus, the LR method has a limitation in reflecting variations of the safety factor. However, the contour map analyzed by the NN shows similar trends of color and distribution at the upper, middle, and lower slopes with the results based on Eq. (1). In particular, the NN has a characteristic of reflecting various changes in the safety factor in the left and middle areas of the slope. A similar behavior was demonstrated by RMSE and MSE of LR and the NN, which were quantitatively compared with Eq. (1), as shown in Figs. 13 and 16. Therefore, an additional analysis of important input variables based on the NN was attempted due to the relatively high resolution.

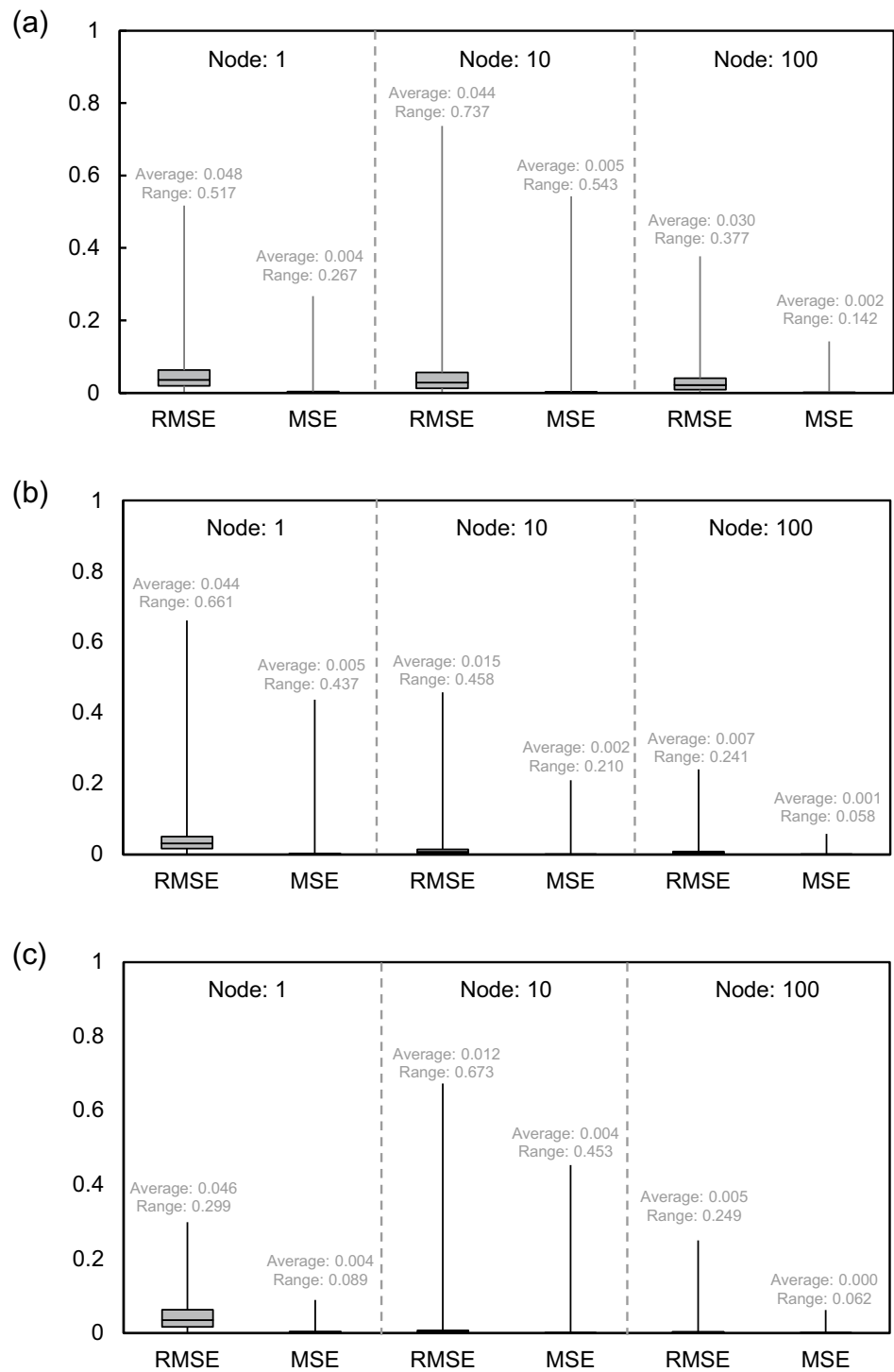
Although the deterministic method proposed in this study is also composed of eight variables, it also has a limitation in obtaining all input variables. By adding variables one by one based on the importance results of the RF, the contour maps in Fig. 21 were constructed, in order to find the smallest input variable compared with the resolution through Eq. (1) and to increase the utility of the deterministic method. Soil thickness, elastic modulus, fine content, hydraulic conductivity, porosity, water content, and saturation were added step by step to shear strength based on Fig. 19. The same hyperparameters were applied, referring to the results of Fig. 16. Figure 21k, which is the result of adding all eight input variables, shows the same contour map as Fig. 21c because it is based on the NN with all input variables. Although shear strength and soil thickness are the most important factors in the NN from RF, the contour maps of shear strength and shear strength + soil thickness roughly reflect the



**Figure 15.** Comparison between predicted and actual safety factors based on neural network: (a) gradient descent algorithm; (b) Levenberg–Marquardt algorithm; (c) Bayesian regularization algorithm.

accrued value of Fig. 21a. In addition, the constant safety factor at the top and bottom is widely distributed with slight differences from the actual values. However, Fig. 21f, which is the result of applying three important factors (shear strength + soil thickness + elastic modulus) as input variables, shows a similar distribution of the safety factor with the contour map of Fig. 21a. It can be seen that the areas in which the safety factor of the upper left and lower right parts is 2.5 are also reflected. When four important factors (shear strength + soil thickness + elastic modulus + fine content) are selected as input values, the resolution is more similar to Fig. 21a, and the contour map from five input variables shows even more similarity, as shown in Fig. 21h–k.

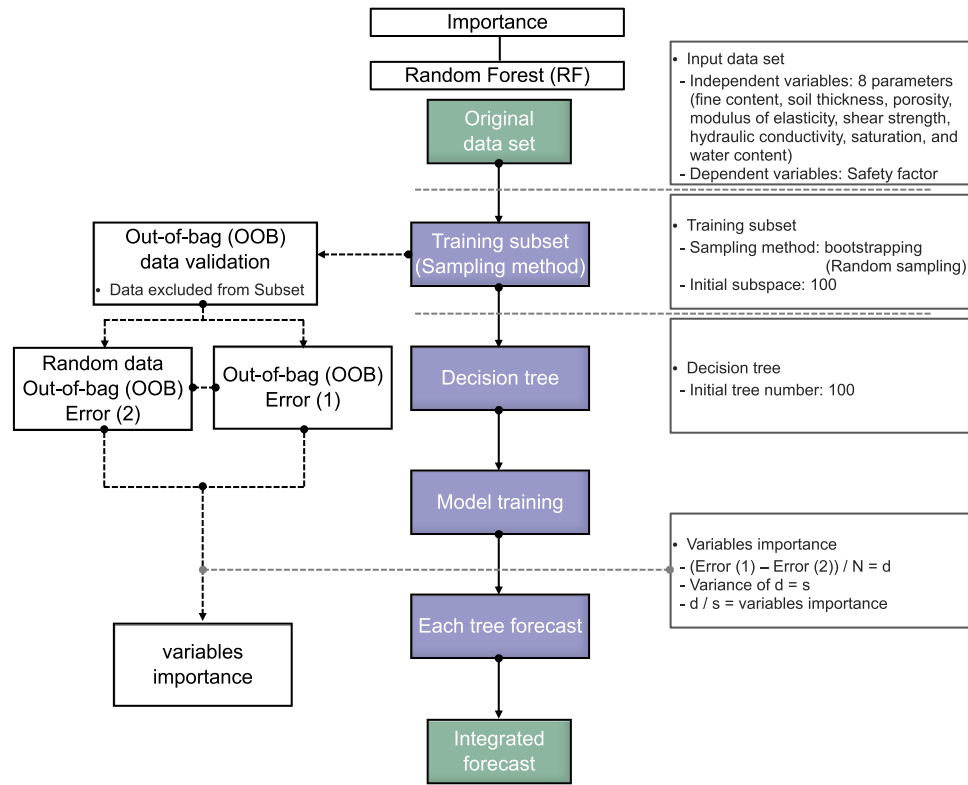
RMSE using one to eight important factors as input variables is shown in Fig. 22 to quantitatively compare the difference in the safety factor according to the input factors. The average RMSE based on one important factor of shear strength was calculated as 0.091, and when all eight variables were entered, the average RMSE was calculated as 0.001, which is 98% less. The more important factors added, the lower the RMSE. A similar range of average RMSE was estimated to be about 0.003–0.005 when four important factors (shear strength + soil thickness + elastic modulus + fine content) and seven important factors (shear strength + soil thickness + elastic modulus + fine content + hydraulic conductivity + porosity + water content) were applied. These results suggest that, if it is difficult to obtain all eight input variables when using an NN, a reliable contour map can be derived



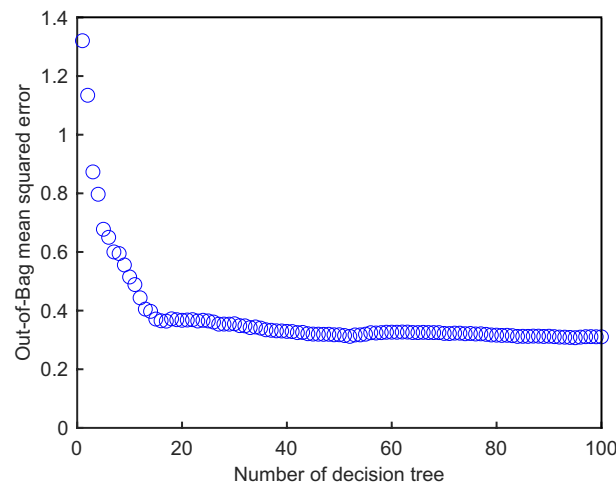
**Figure 16.** Distribution of RMSE and MSE based on neural network: (a) gradient descent; (b) Levenberg–Marquardt; (c) Bayesian regularization.

by capturing the order of important factors. In addition, it is judged that the assessment of quantitative susceptibility is possible by using at least four important factors: shear strength, soil thickness, elastic modulus, and fine content.

The RMSE was also used to verify results of this study and the comparison of RMSE was demonstrated in Fig. 23 with previously studied values. To reliably compare RMSE, previous studies that recently performed to estimate landslide susceptibility through machine learning algorithm were selected. And thus, studies conducted by Pham et al.<sup>65</sup>, Tien Bui et al.<sup>17</sup>, Van Dao et al.<sup>14</sup>, Ghasemian et al.<sup>18</sup> and Nhu et al.<sup>66</sup> were selected because they



**Figure 17.** Flow chart of the random forest.



**Figure 18.** Out-of-bag refers to the squared error according to the number of trees.

used to ensemble technique, novel hybrid function, spatially explicit deep learning neural network, reduced error pruning tree, and combined ensemble model, respectively for mapping risk area of landslide. The average value of RMSE shown in each paper is 0.336, 0.452, 0.387, 0.210 and 0.321 for Pham et al.<sup>65</sup>, Tien Bui et al.<sup>17</sup>, Van Dao et al.<sup>14</sup>, Ghasemian et al.<sup>18</sup> and Nhu et al.<sup>66</sup> respectively. However, the RMSE of this study was found to be 0.005 and 0.001 on average, which are relatively low values when four main parameters and eight parameters were used. These results show different ratios of 41–89.4% and 209–451% based on the four main parameters and eight parameters, respectively. Even though the input variable and selected algorithms were different between this study and the previous studies, the comparison of RMSE show that the methodology used in the study is reliable when evaluating landslide susceptibility through machine learning. The relatively low RMSE is dominated by two factors: 1) input data and 2) machine learning algorithm with hyperparameter. It shows that the eight main variables in empirical equation, constructed by the opinions of expert group, excellently reflect landslide susceptibility considering characterizations of solid, structural failure, and water flow, and high quality data was



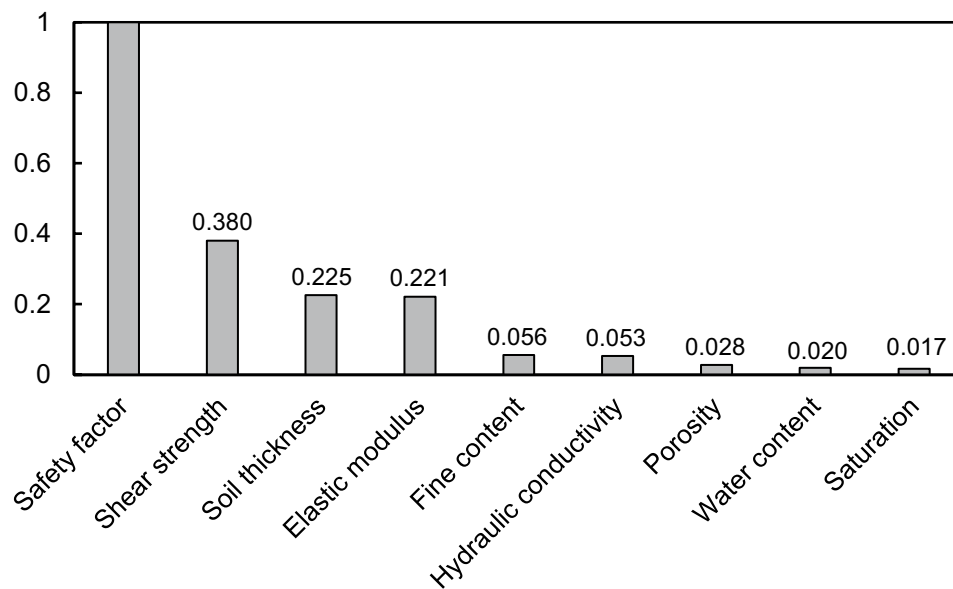


Figure 19. Importance of input variables based on random forest.

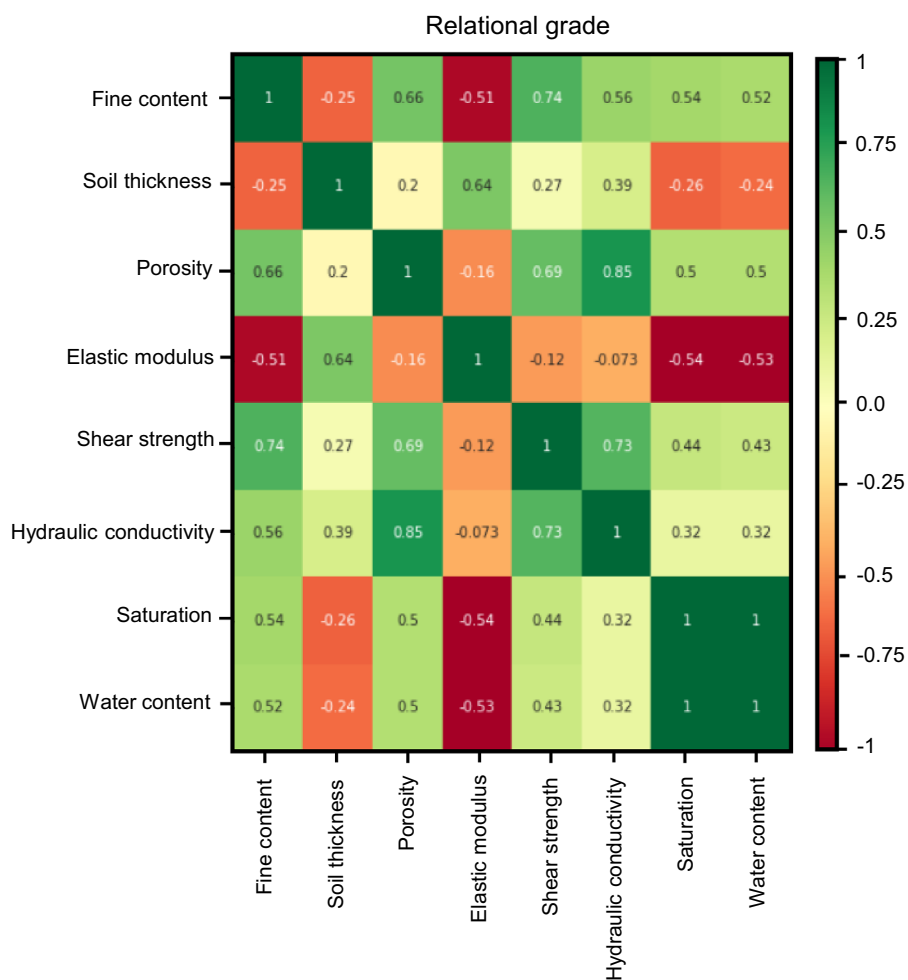
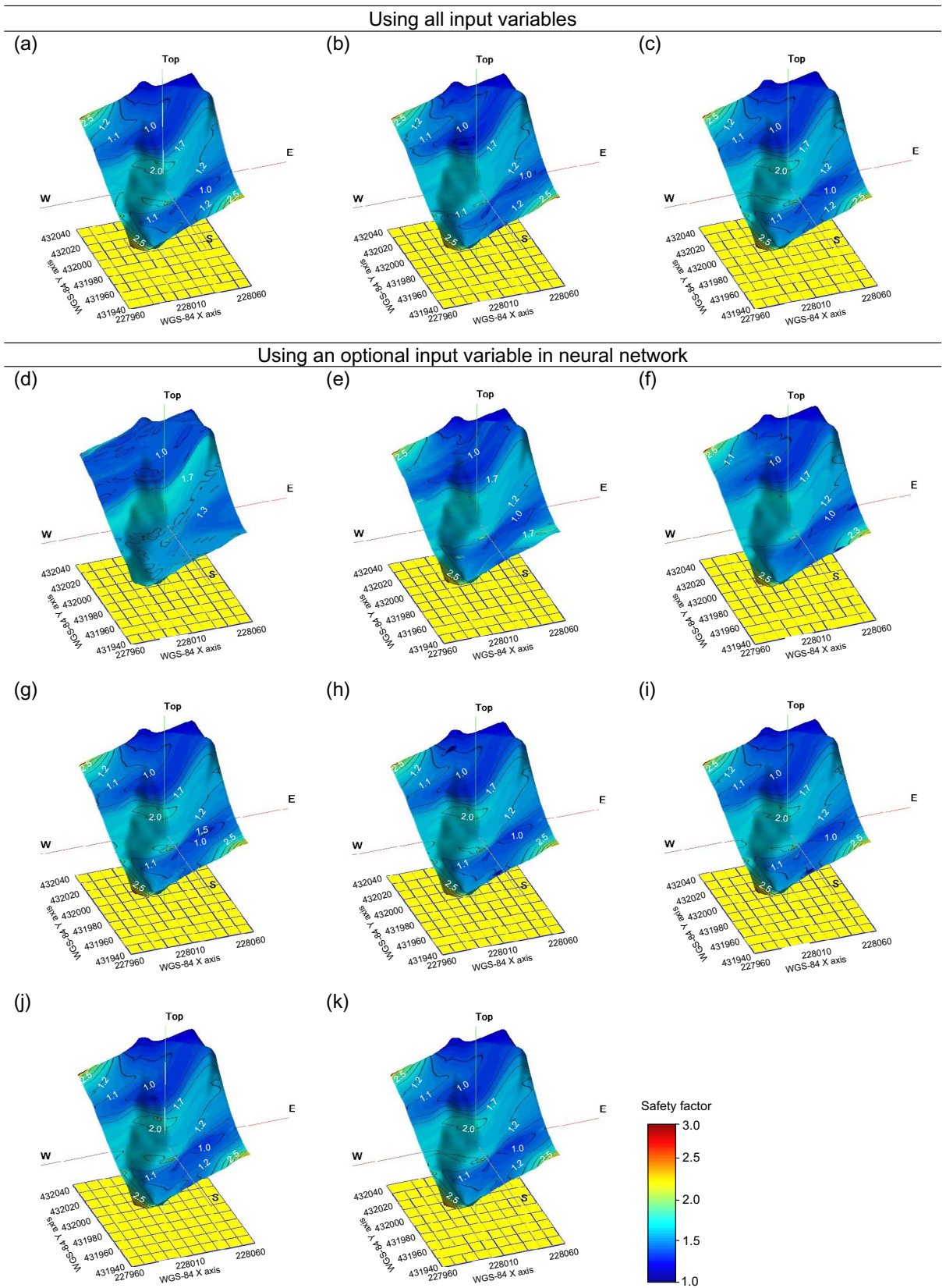
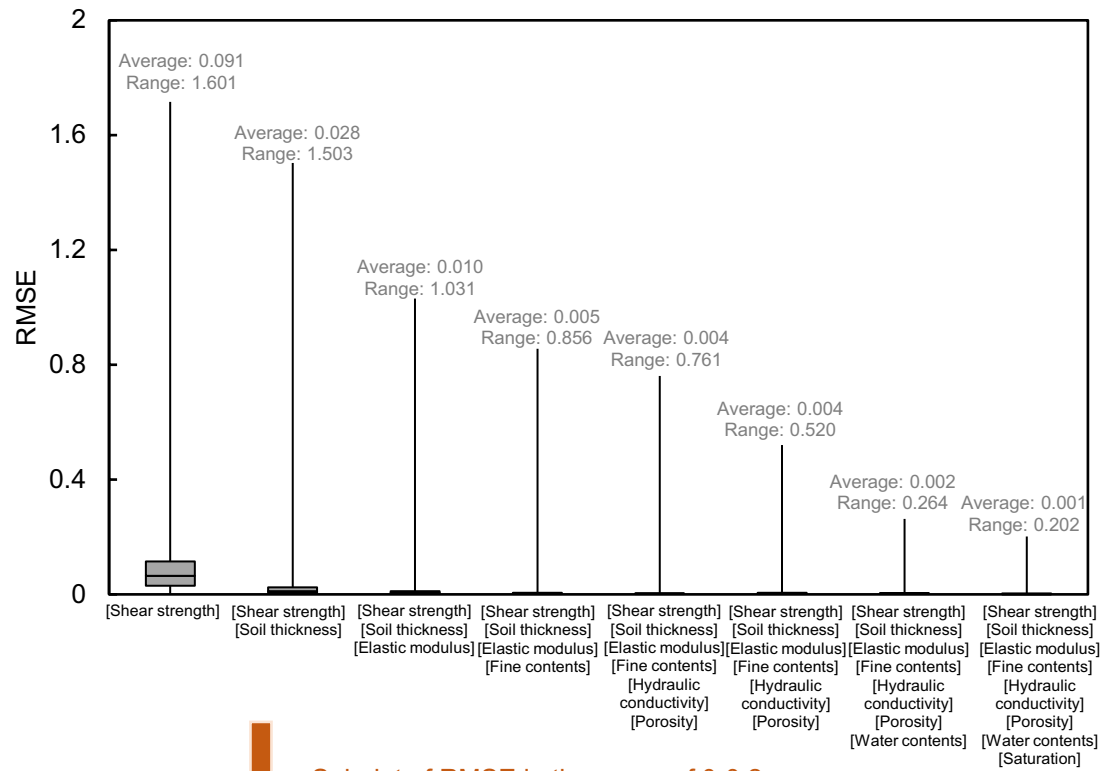


Figure 20. Gray relational grade (GRG) results based on random forest.



**Figure 21.** Safety factor modeling based on: (a) Eq. (1), (b) linear regression; (c) neural network; (d) 1 parameter; (e) 2 parameters; (f) 3 parameters; (g) 4 parameters; (h) 5 parameters; (i) 6 parameters; (j) 7 parameters; (k) 8 parameters. The software of SGeMS beta (v.2.5b) was applied to draw this figure.



Subplot of RMSE in the range of 0-0.2

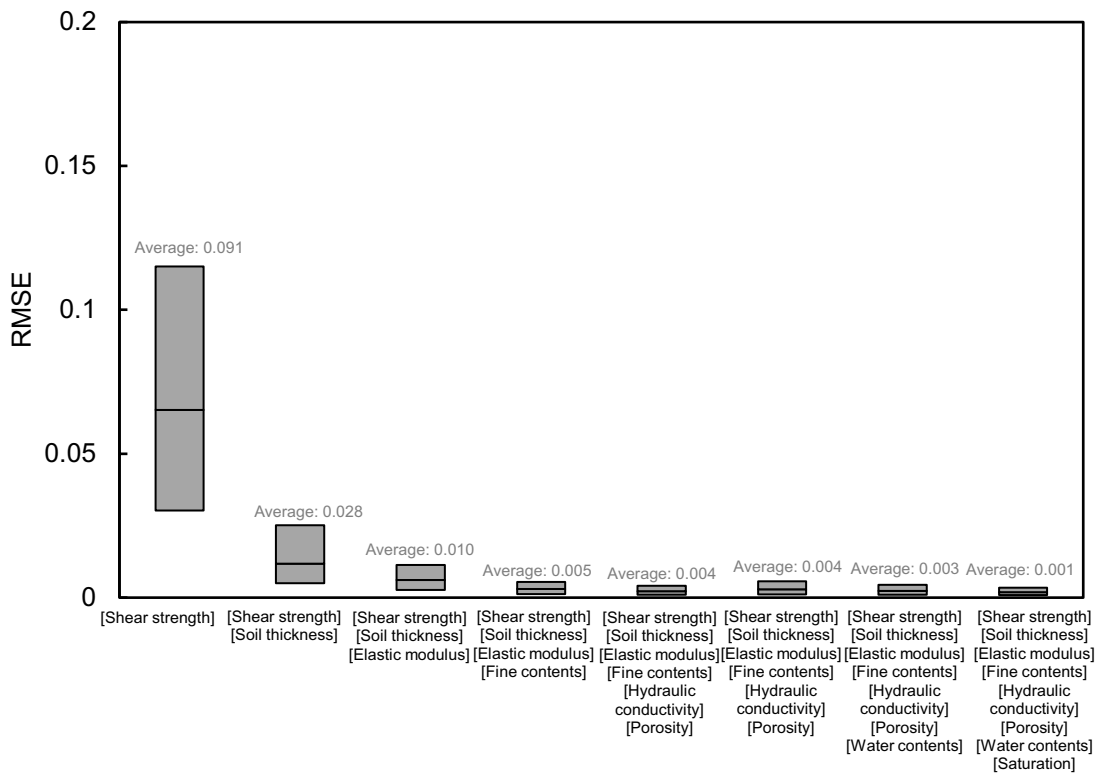
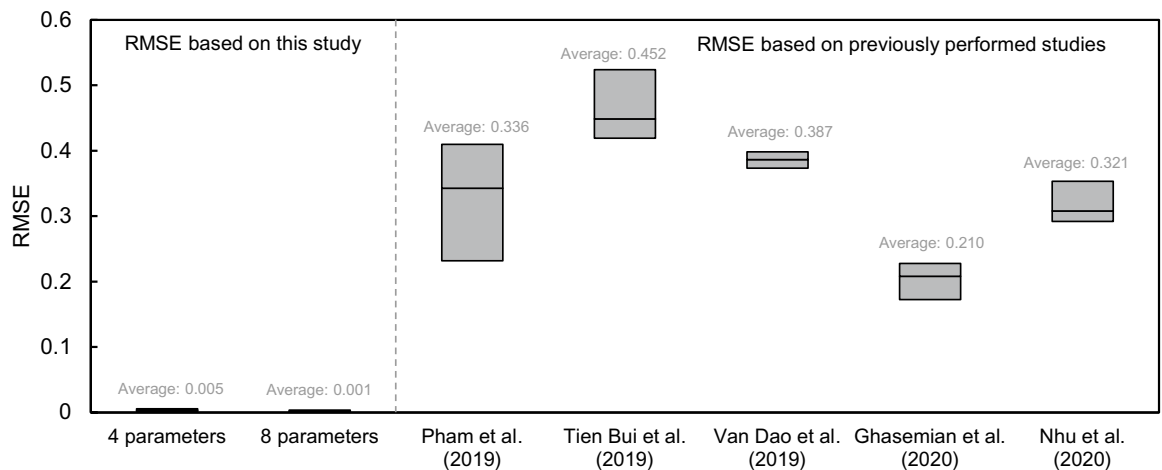


Figure 22. Box plot of RMSE according to the optional input variable in the neural network.



**Figure 23.** Comparison of RMSE based on this study and previously performed studies.

used as input variables through various experiments. In addition, optimally selected machine learning algorithms and hyperparameter produced reliability result as shown in Figs. 12 and 15. Even though the accuracy depends on geological condition, the number of data and characterization of soil, the applied method in this study is a wise option to estimate landslide susceptibility.

It is difficult to find 8 variables through laboratory and field tests. However, the elastic wave survey is a prior investigation method to understand the entire range of the mountain, and the elastic modulus, which is one of the four main factors, can be obtained through an elastic wave survey. It can also provide the location of the DCPT to obtain soil thickness and shear strength. Finally, fine content is also captured after sampling at the same location with the penetration test. In this way, all four main factors can be obtained. Of course, obtaining hydraulic conductivity, porosity, saturation, and water content, if possible, is also necessary to improve reliability.

## Conclusion

In this study, the weight of each variable in the empirical equation was estimated using LR and NN, and the distribution of the safety factor was addressed. The detailed conclusions are as follows:

- Field and laboratory experiments were conducted to obtain eight variables (fine content, soil thickness, porosity, elastic modulus, shear strength, hydraulic conductivity, saturation, and water content), and the dataset was built through the interpolation method.
- The gradient descent, Levenberg–Marquardt (LM), and Bayesian regularization (BR) methods were used to more precisely perform back propagation, and LM and BR showed excellent performance.
- The safety factor estimated by the NN was most similar to the true value, and the safety factor calculated by the four main factors of shear strength, soil thickness, elastic modulus, and fine content also showed high reliability. Although it is difficult to obtain all eight variables used in the study, the quantitative susceptibility of slope can be evaluated by using only these four main factors.

Received: 2 November 2020; Accepted: 10 March 2021

Published online: 23 March 2021

## References

1. Dikshit, A., Sarkar, R. & Satyam, N. Probabilistic approach toward Darjeeling Himalayas landslides: A case study. *Cogent Eng.* 5(1), 1537539–1537550 (2018).
2. Lee, S., Jang, J., Kim, Y., Cho, N. & Lee, M. J. Susceptibility analysis of the Mt. Umyeon landslide area using a physical slope model and probabilistic method. *Remot. Sens.* 12(16), 2663–2680 (2020).
3. Lee, S., Lee, M. J., Jung, H. S. & Lee, S. Landslide susceptibility mapping using naïve bayes and bayesian network models in Umyeonsan, Korea. *Geocarto Int.* 35(15), 1665–1679 (2020).
4. Wang, X. *et al.* Uncertainty assessment in quantitative rockfall risk assessment. *Landslides* 11(4), 711–722 (2014).
5. Fu, S. *et al.* Landslide hazard probability and risk assessment at the community level: A case of western Hubei, China. *Nat. Hazard. Earth Syst. Sci.* 20(2), 581–601 (2020).
6. Salciarini, D., Volpe, E. & Cattoni, E. Probabilistic vs. deterministic approach in landslide triggering prediction at large-scale. In national conference of the researchers of geotechnical engineering. *Springer* 1, 62–70 (2019).
7. Chen, L. K., Chang, C. H., Liu, C. H. & Ho, J. Y. Application of a three-dimensional deterministic model to assess potential landslides, a case study: Antong Hot Spring Area in Hualien. *Taiwan. Water* 12(2), 480–495 (2020).
8. Ba, Q., Chen, Y., Deng, S., Yang, J. & Li, H. A comparison of slope units and grid cells as mapping units for landslide susceptibility assessment. *Earth Sci. Inf.* 11(3), 373–388 (2018).
9. Jun, H., Min, D. H. & Yoon, H. K. Determination of monitoring systems and installation location to prevent debris flow through web-based database and AHP. *Mar. Georesour. Geotechnol.* 35(8), 1049–1057 (2017).
10. Yousefi, S. *et al.* Assessing the susceptibility of schools to flood events in Iran. *Sci. Rep.* 10(1), 1–15 (2020).

11. Bui, D. T., Nhu, V. H. & Hoang, N. D. Prediction of soil compression coefficient for urban housing project using novel integration machine learning approach of swarm intelligence and multi-layer perceptron neural network. *Adv. Eng. Inform.* **38**(1), 593–604 (2018).
12. Chang, K. T., Merghadi, A., Yunus, A. P., Pham, B. T. & Dou, J. Evaluating scale effects of topographic variables in landslide susceptibility models using GIS-based machine learning techniques. *Sci. Rep.* **9**(1), 1–21 (2019).
13. Di, B. *et al.* Assessing susceptibility of debris flow in southwest China using gradient boosting machine. *Sci. Rep.* **9**(1), 1–12 (2019).
14. Van Dao, D. *et al.* A spatially explicit deep learning neural network model for the prediction of landslide susceptibility. *CATENA* **188**, 104451 (2020).
15. Kuradusenge, M., Kumaran, S. & Zennaro, M. Rainfall-induced landslide prediction using machine learning models: The case of Ngororero District, Rwanda. *Int. J. Environ. Res. Public Health* **17**(11), 4147 (2020).
16. Wang, Z., Liu, Q. & Liu, Y. Mapping landslide susceptibility using machine learning algorithms and GIS: A case study in Shexian County, Anhui Province, China. *Symmetry* **12**(12), 1954 (2020).
17. Tien Bui, D. *et al.* Shallow landslide prediction using a novel hybrid functional machine learning algorithm. *Remot. Sens.* **11**(8), 931 (2019).
18. Ghasemian, B. *et al.* Shallow landslide susceptibility mapping: A comparison between classification and regression tree and reduced error pruning tree algorithms. *Vietnam J. Earth Sci.* **42**(3), 208–227 (2020).
19. Ilbeigi, M., Ghomeshi, M. & Dehghanbanadaki, A. Prediction and optimization of energy consumption in an office building using artificial neural network and a genetic algorithm. *Sustain. Cities Soc.* **61**, 102325 (2020).
20. Khan, I. *et al.* Design of neural network with Levenberg–Marquardt and Bayesian regularization backpropagation for solving pantograph delay differential equations. *IEEE Access* **8**, 137918–137933 (2020).
21. Hammond, C. J., Prellwitz, R. W., and Miller, S. M. Landslide hazard assessment using Monte Carlo simulation. In *Proceedings of 6th international symposium on landslides*, 251–294 (Christchurch, 1992).
22. Ietto, F., Perri, F. & Cella, F. Geotechnical and landslide aspects in weathered granitoid rock masses (Serre Massif, southern Calabria, Italy). *CATENA* **145**, 301–315 (2016).
23. Meisina, C. & Scarabelli, S. A comparative analysis of terrain stability models for predicting shallow landslides in colluvial soils. *Geomorphology* **87**(3), 207–223 (2007).
24. Hobbs, P. R. N., & Boon, D. P. (2013). Great Fryup landslide: Slope stability analyses. British Geological Survey, Internal Report IR/13/002. British Geological Survey, Nottingham.
25. Uhlemann, S. *et al.* Assessment of ground-based monitoring techniques applied to landslide investigations. *Geomorphology* **253**, 438–451 (2016).
26. Calcaterra, D. & Parise, M. Landslide types and their relationships with weathering in a Calabrian basin, Southern Italy. *Bull. Eng. Geol. Environ.* **64**(2), 193–207 (2005).
27. Schilirò, L., Montrasio, L. & Mugnozza, G. Prediction of shallow landslide occurrence: Validation of a physically-based approach through a real case study. *Sci. Total Environ.* **569**, 134–144 (2016).
28. Paulin, G. L., Bursik, J., Lugo-Hubp, M. & Orozco, J. Z. Effect of pixel size on cartographic representation of shallow and deep-seated landslide, and its collateral effects on the forecasting of landslides by SINMAP and Multiple Logistic Regression landslide models. *Phys. Chem. Earth A/B/C* **35**(3), 137–148 (2010).
29. Dweiri, F., Kumar, S., Khan, S. A. & Jain, V. Designing an integrated AHP based decision support system for supplier selection in automotive industry. *Expert Syst. Appl.* **62**, 273–283 (2016).
30. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65**(6), 386–408 (1958).
31. Chauvin, Y. & Rumelhart, D. E. *Backpropagation: Theory, Architectures, and Applications* 1–475 (Psychology press, 1995).
32. Breiman, L. Random forests. *Mach. Learn.* **45**(1), 5–32 (2001).
33. Zhang, P., Yin, Z. Y., Jin, Y. F. & Chan, T. H. A novel hybrid surrogate intelligent model for creep index prediction based on particle swarm optimization and random forest. *Eng. Geol.* **265**, 105328–105350 (2020).
34. Zeiler, M. D. *Adadelta: An Adaptive Learning Rate Method* 1–6 (Cornell University, 2012).
35. Roweis, S. *Levenberg–Marquardt Optimization* (University of Toronto, 1996).
36. MacKay, D. J. The evidence framework applied to classification networks. *Neural Comput.* **4**(5), 720–736 (1992).
37. Sariev, E. & Germano, G. Bayesian regularized artificial neural networks for the estimation of the probability of default. *Quant. Financ.* **20**(2), 311–328 (2020).
38. Bouktif, S., Fiaz, A., Ouni, A. & Serhani, M. A. Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. *Energies* **11**(7), 1636–1656 (2018).
39. Rymarczyk, T., Kłosowski, G., Kozłowski, E. & Tchórzewski, P. Comparison of selected machine learning algorithms for industrial electrical tomography. *Sensors* **19**(7), 1521–1542 (2019).
40. Yaseen, Z. M. *et al.* Non-tuned machine learning approach for hydrological time series forecasting. *Neural Comput. Appl.* **30**(5), 1479–1491 (2018).
41. Yoon, H. K. Relationship between aspect ratio and crack density in porous-cracked rocks using experimental and optimization methods. *Appl. Sci.* **10**(20), 7147–7168 (2020).
42. Oliver, M. A. & Webster, R. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *CATENA* **113**, 56–69 (2014).
43. Park, C. H., Byun, J. H., Won, K. S., Cho, H. T. & Yoon, H. K. Characterization of alluvium soil using geophysical and sounding methods. *Mar. Georesour. Geotechnol.* **35**(1), 127–135 (2017).
44. Choo, H., Jun, H. & Yoon, H. K. Application of elastic wave velocity for estimation of soil depth. *Appl. Sci.* **8**(4), 600–614 (2018).
45. Choo, H., Jun, H. & Yoon, H. K. Porosity estimation of unsaturated soil using Brutsaert equation. *Soil Dyn. Earthquake Eng.* **104**, 33–39 (2018).
46. Lee, S. J. & Yoon, H. K. Hydraulic conductivity of saturated soil medium through time-domain reflectometry. *Sensors* **20**(23), 7001–70018 (2020).
47. ASTM Committee D-18 on Soil and Rock. *Standard Test Methods for Particle-Size Distribution (Gradation) of Soils Using Sieve Analysis*. ASTM International (2009).
48. Archie, G. E. The electrical resistivity log as an aid in determining some reservoir characteristics. *Trans. AIME* **146**(01), 54–62 (1942).
49. Lee, J. S. & Yoon, H. K. Theoretical relationship between elastic wave velocity and electrical resistivity. *J. Appl. Geophys.* **116**, 51–61 (2015).
50. Gorman, T. & Kelly, W. E. Electrical-hydraulic properties of unsaturated Ottawa sands. *J. Hydrol.* **118**(1–4), 1–18 (1990).
51. Keller, G. V. & Frischknecht, F. C. *Electrical Methods in Geophysical Prospecting* (Springer, 1966).
52. Dunkelmann, T. J., Rosendahl, B. R. & Karson, J. A. Structure and stratigraphy of the Turkana rift from seismic reflection data. *J. Afr. Earth Sci.* **8**(2–4), 489–510 (1989).
53. Le Douaran, S., Burrus, J. & Avedik, F. Deep structure of the north-western Mediterranean Basin: Results of a two-ship seismic survey. *Mar. Geol.* **55**(3–4), 325–345 (1984).
54. Whiteley, R. J. & Greenhalgh, S. A. Velocity inversion and the shallow seismic refraction method. *Geoexploration* **17**(2), 125–141 (1979).



55. Azimian, A. & Ajalloeian, R. Empirical correlation of physical and mechanical properties of marly rocks with P wave velocity. *Arab. J. Geosci.* **8**(4), 2069–2079 (2015).
56. Santamarina, J. C., Klein, K. A. & Fam, M. A. *Soils and Waves* (Wiley, 2001).
57. Schultheiss, P. J. Simultaneous measurement of P & S wave velocities during conventional laboratory soil testing procedures. *Mar. Georesour. Geotechnol.* **4**(4), 343–367 (1981).
58. Yoon, H. K., Kim, D. H., Lee, W. & Lee, J. S. Field elastic wave and electrical resistivity penetrometer for evaluation of elastic moduli and void ratio. *J. Korean Soc. Civ. Eng.* **30**(2C), 85–93 (2010).
59. ASTM D 6951/D 6951M. Standard test method for use of the dynamic cone penetrometer in shallow pavement applications. In *American Society for Testing and Materials* (2018).
60. Tsuchida, T., Athapaththu, A. M. R. G., Kano, S. & Suga, K. Estimation of in-situ shear strength parameters of weathered granitic (Masado) slopes using lightweight dynamic cone penetrometer. *Soils Found.* **51**(3), 497–512 (2011).
61. Choo, H., Min, D. H., Sung, J. H. & Yoon, H. K. Sensitivities of input parameters for predicting stability of soil slope. *Bull. Eng. Geol. Env.* **78**(8), 5671–5685 (2019).
62. Ejigu, B. A., Wencheke, E., Moraga, P. & Giorgi, E. Geostatistical methods for modelling non-stationary patterns in disease risk. *Spat. Stat.* **35**, 100397–100415 (2020).
63. Taiwo, S. M. & Yoon, H. K. Estimation of elastic wave velocity and DCPI distributions using outlier analysis. *Eng. Geol.* **247**, 129–144 (2018).
64. Ließ, M., Glaser, B. & Huwe, B. Uncertainty in the spatial prediction of soil texture: Comparison of regression tree and Random Forest models. *Geoderma* **170**, 70–79 (2012).
65. Pham, B. T. *et al.* Landslide susceptibility modeling using Reduced Error Pruning Trees and different ensemble techniques: Hybrid machine learning approaches. *CATENA* **175**, 203–218 (2019).
66. Nhu, V. H. *et al.* Landslide susceptibility mapping using machine learning algorithms and remote sensing data in a tropical environment. *Int. J. Environ. Res. Public Health* **17**(14), 4933 (2020).

## Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R1A2C2012113).

## Author contributions

D.-H.M. managed data and H.-K.Y. analysis data. All authors reviewed final version manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to H.-K.Y.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021