https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

Natural Hazards and Earth System Sciences
Discussions

Prediction of volume of shallow landslides due to rainfall using data-driven models

Jérémie Tuganishuri, Chan-Young Yune, Manik Das Adhikari, Seung Woo Lee, Gihong Kim, Sang-GukYum*

Department of Civil and Environmental Engineering, Gangneung-Wonju National University, Gangneung, Gangwon-do, Republic of Korea

*Corresponding author: Sang-Guk Yum; skyeom0401@gwnu.ac.kr

Abstract

Landslides due to rainfall are among most destructive natural disasters that cause property damages, huge financial losses and human deaths in different parts of the World. To plan for mitigation and resilience, the prediction of the volume of rainfall-induced landslides is essential to understand the relationship between the volume of soil materials debris and their associated predictors. Objectives of this research are to construct a model by utilizing advanced data-driven algorithms (i.e., ordinary least square or Linear regression (OLS), random forest (RF), support vector machine (SVM), extreme gradient boosting (EGB), generalized linear model (GLM), decision tree (DT), and deep neural network (DNN), K-nearest neighbor (KNN) and Ridge regression (RR)) for the prediction of the volume of landslides due to rainfall considering geological, geomorphological, and environmental conditions. Models were tested on the Korean landslide dataset to observe the best-performing model, and among tested algorithms, the extreme gradient boosting ranked high with the coefficient of determination (R^2=0.85) and mean absolute error (MAE=150.421m^3). The volume of landslides was strongly influenced by slope length, drainage status, slope angle, aspect, and age of trees. The anticipated volume of landslide can be important for land use allocation and efficient landslide risk management.

Keywords: Data-driven models, volume of landslide, prediction models, rainfall

1. Introduction

Landslides due to rainfall is a phenomenon in which a given volume of soil dislocates from its original high to lower point altitude due to gravity forces along a slope fragilized by rainfall that crosses a certain threshold (Bernardie et al., 2014; Martinović et al., 2018; Lee et al., 2021). This massive volume of soil causes enormous environmental degradation, infrastructure damage, and casualties, which is a hindrance to socio-economic aspect of the community (Van et al., 2021; Alcántara-Ayala, 2021). The rainfall quantity and duration influence the volume of the landslides; the higher the intensity and the longer the duration of rainfall, the larger the resulting volume of landslides (Chen et al., 2017; Bernardie et al., 2014; Chang and Chiang, 2009). The landslide occurrence can also be influenced by human activities that fragilize the slope, such as excavation at the slope toe and loading caused by construction (Rosi et al., 2016). Therefore, the accurate prediction of the volume of landslides due to rainfall is an important key for designing strategies for resilience and planning for the protection of the inhabitants of a particular region with certain landslide risks subjected to a predicted quantity of rainfall (Conte et al., 2022). Consequently, for the safety of communities, the efficient selection of infrastructure sites must be done in places where landslides cannot bury buildings (Fan et al., 2017). Further, for the protection of crops, the farmland location, and other land use activities, accurate landslide prediction taking into account real root causes through the analysis of triggering and influencing factors, is crucial to achieve a durable landslide safety management system (Paudel et al., 2003; Lee, 2009; Fan et al.,2017; Dai et al., 2019; Alcántara-Ayala, 2021).

The prediction of landslide volume due to rainfall is important for the analysis of infrastructure placement to protect against being buried in extreme landslide events. In South Korea, many infrastructures are placed at the foot of mountains, which makes them vulnerable to extreme landslides, which can bury villages, farm lands etc. The findings of Lee (2016) indicated that due to climate change, the average rainfall has increased by 271.23 mm for the period 1971-2100 based on future climate scenarios. Therefore, the efficient prediction of landslide volumes can be useful for land use management in such a way that locations with expected high volume of landslide may be used for other activities which do not get affected by landslide events, such as forest and gardens or activities that reduce water infiltration and non-continuous disturbance of subsoil to maintain groundwater stability and strengthen the topsoil.

Most researchers focused on the prediction of landslides runout and susceptibility (Giarola

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

Natural Hazards and Earth System Sciences
Open Access
EGU
Discussions

et al., 2024; Melo et al., 2019; Peruzzetto et al., 2020). Nevertheless, few researchers estimated the volume of landslides based on the statistical approach (Ju et al., 2023; Dai and Lee, 2001). Ju et al. (2023) constructed an area-volume power law model for the estimation of the volume of landslides using LiDAR data in Hong Kong. Razakova et al. (2020) calculated landslide volume using a digital elevation model and ground-based measurement. Dai and Lee (2001) found that the 12 hours of rainfall influenced the volume of landslides and frequency-volume followed the power law relation. It was observed that most of these studies did not consider detailed predisposing factors and their contribution to the prediction of the volume of landslides due to rainfall. Recently, Lee et al. (2021) applied an artificial neural network (ANN) model for the prediction of the volume of debris flow in the central region of South Korea based on the patterns from the already occurred landslide characteristics and the region morphometry. In the present study, the volume of landslides due to rainfall is predicted using OLS, RF, SVM, EGB, GLM, DT, DNN, KNN and RR algorithms, considering the details of triggering factors (i.e., rainfall) and predisposing factors (i.e., geological, geomorphological, and environmental).

In this study, we aim to construct a data-driven algorithm that combines input parameters for physical-based and empirical models and incorporate more complex non-linear features of input variables to predict the occurrence of associated events more accurately. The main assumption behind the data-driven algorithm is that the considered feature input of the model produces similar volume of landslides due to rainfall and follows the same pattern at a particular region with the same features under the same quantity of rainfall. Here, we examine different machine learning algorithms and compare their performance using the coefficient of determinations R^2 and mean square errors (MAE) resulting from the application of each algorithm. The model can be customized to be applied in other regions according to the regional settings.

2. Study area

The region for testing the model is South Korea, characterized by mountainous (63% of total land) relief, especially in the eastern part of the country (Lee et al., 2021). The Korean peninsula climate comprises cold and dry winters and humid summers. During the summer season, heavy rainfall from June to September causes 95% of all landslides due to rainfall each year (Lee et al., 2020). In addition, the landslides may be aggravated by typhoons, which mostly occur in August and September, and it is anticipated that frequency will increase due to climate change. The annual

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

Natural Hazards and Earth System Sciences
Open Access
Discussions

rainfall ranges between 1000 mm to 1400mm and 1800mm for the central region and southern region, respectively (Jung et al., 2017; Alcantara and Ahn, 2020). The geology of the Korean peninsula is composed of metamorphic (45%), igneous (30%) and 25% of sedimentary rocks (Lee and Winter, 2019). Subsequently, the influence of rainfall, environmental and geological factors frequently generated landslides across the country as depicted in Figure 1. The distribution of rainfall and volume is summarized in Fig 1.

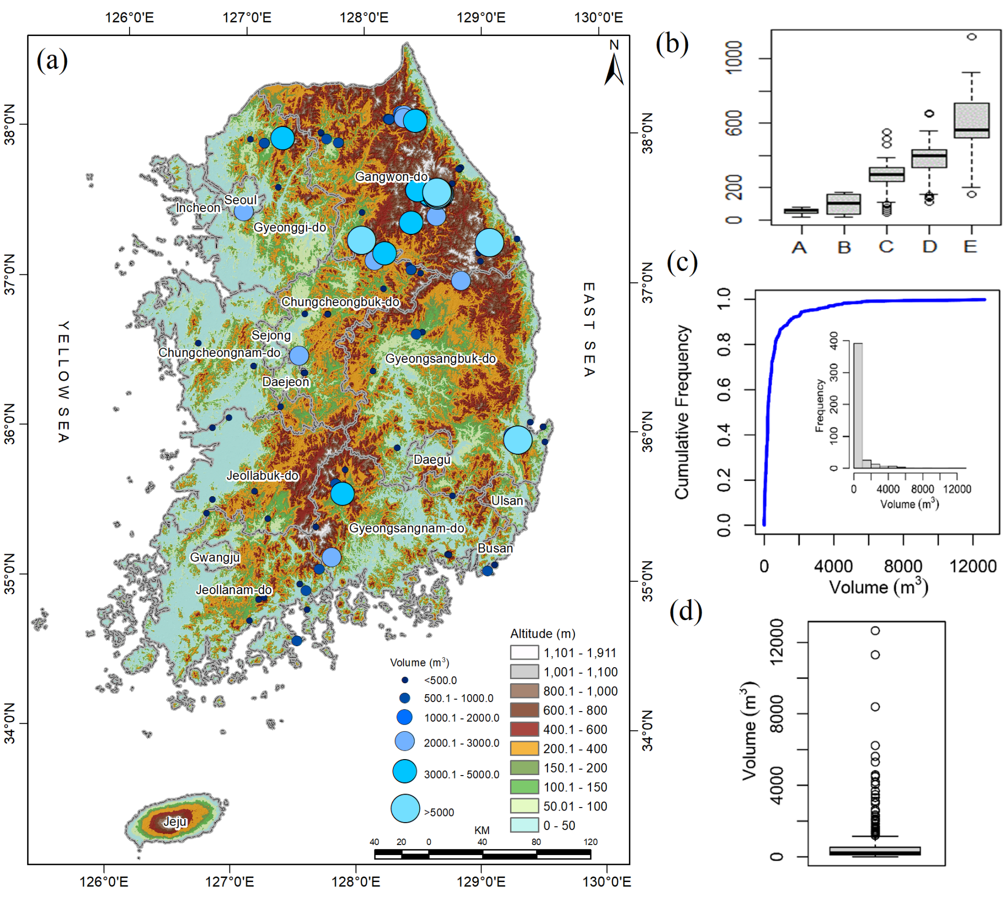

Figure 1. (a) Spatial distribution of landslides in South Korea, (b) temporal variation of rainfall, i.e., A: Maximum hourly rainfall, B: Four weeks rainfall, C: Three hours rainfall, D: Three days rainfall and E: Two weeks rainfall, (c) cumulative frequency distribution of volume of landslides and (d) box plot of volume of landslides.

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

Natural Hazards and Earth System Sciences
Discussions

3. Data and method

In this paper, we consider nine data-driven models, namely ordinary least square or Linear regression (OLS), random forest (RF), support vector machine (SVM), extreme gradient boosting (EGB), generalized linear model (GLM), decision tree (DT), and deep neural network (DNN), k-nearest neighbor (KNN) and Ridge regression (RR) to predict the volume of landslides due to rainfall. The model is tested on the South Korean landslides inventories and predisposing factors coupled with triggering factors, i.e., rainfall data. The detailed workflow is summarized in Figure 2. The steps for construction of these models can be briefly summarized as follows: a) the dataset for landslide inventories is cleaned and joined with rainfall dataset, b) the collinearity analysis is made using variance inflation factor, c) continuous feature are scaled (Z-score) (Bonamutial and Prasetyo, 2023) to facilitate algorithms to converge fast, d) the dataset is split into training and test set, e) all models are tested on the same training set, and the model evaluation on the test set using MAE and R^2 for the comparison of actual and predicted volume by each model, f) variable importance is calculated for most performing model, and g) the distance correlation is calculated for each continuous feature, and Kruskal-Wallis and Dunn test are conducted to examine the similarity of the effect of each category on the landslide volume.

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

Natural Hazards and Earth System Sciences
Discussions

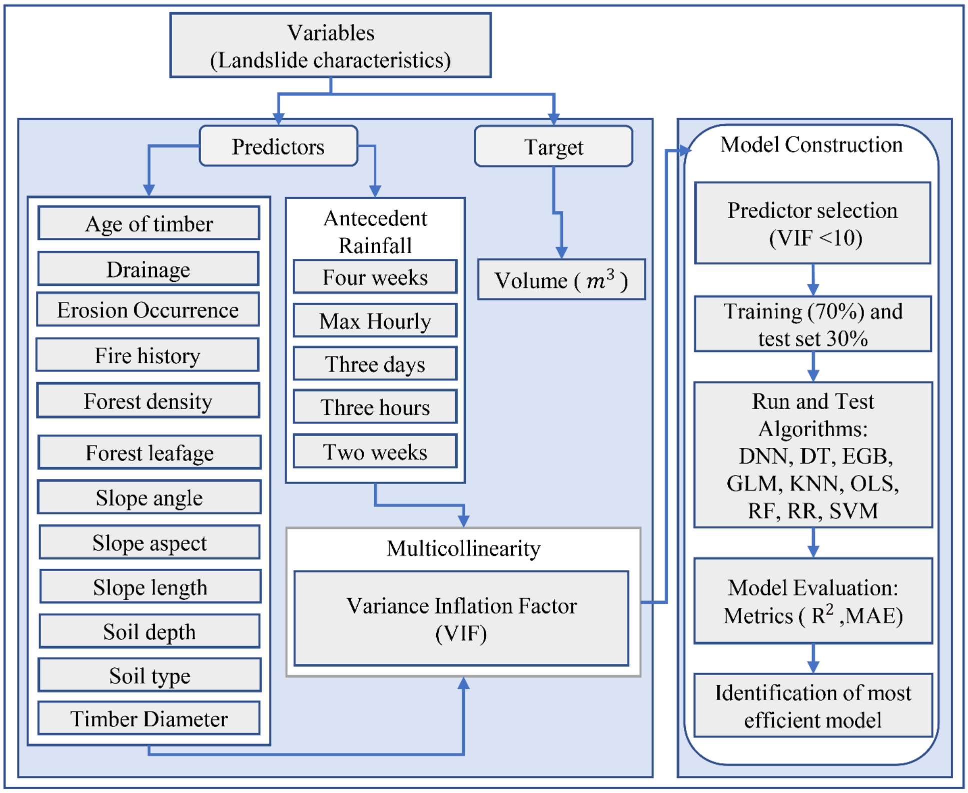

Figure 2.Workflow for the prediction of volume of landslide due to rainfall.

3.1 Data

The landslide inventory dataset contains 450 landslide record information from 2011 to 2012, which was collected from different locations in South Korea by Korean Forest Services. This dataset tabulates landslide location, volume, slope length, soil type, drainage situation, fire history, and vegetation features such as age, diameter of timber, leafage, and forest density. The outcome variable (volume) to be predicted was estimated as a product of the area affected by landslides and its depth. The estimation of volume of flown away material by landslides is important as it help to assess risks the estimated damage can cause at the valley at the bottom of the failed slope, such as blocking transportation network, burying crops or farmland, damage-built environment near landslide risks area (Evans et al.,2007; Rotaru et al.,2007; Intrieri et al., 2019).

Landslides due to rainfall occur as a result of slope failure over-saturation from groundwater and rainfall infiltration that destabilize the slope (Kafle, 2022). Therefore, slope

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

length, slope angle and slope aspect play an important role in the determination of the volume of geological material uprooted by landslides (Zaruba and Mencl, 2014; Khan et al., 2021). The slope stability depends on the properties of composing material which have different soil permeability index which indicates water infiltration capability (Chen et al., 2015). From surveyed regions three main soil types, namely, sandy loam, loam, and silt loam, were observed, and their coefficient of permeability is 1.7, 1.65 and 1.5, respectively (Lee et al., 2013), were used as numerical predictor variables. In addition, the drainage network that channeling rainwater in hilly terrain drains soil and reduces the saturation which minimizes the likelihood of landslide occurrence as a result of groundwater discharge and rainfall water flow (Hovius et al.,1998; Wei et al., 2019). Furthermore, the occurrence of forest fires can contribute to the occurrence of landslides due to the burning of vegetation covering the area and can also change soil property and increase soil pH (Lee et al., 2013). Moreover, the vegetation type, leafage, roots, age and density can be predictors of the occurrence and the volume of landslides. The vegetation covers the topsoil, prevents drying and the direct hit of rain drops which automatically dig holes on the ground due to the force of gravity acting on the raindrop combined with the soil permeability (Omwega,1989; Keefer, 2000). The absence of vegetation allows rainwater to seep away fine topsoil, causing shallow landslides (Gonzalez-Ollauri and Mickovski, 2017). Thus, planting vegetation is recommended as a better practice to improve soil cohesion and prevent potential landslides due to soil root interaction (Gong et al., 2017; Phillips et al., 2021). The density of forest and leafage type (broad, pines or mixture) determine the quantity of raindrop intercepted and prevented from hitting directly the soil which emphasizes the vegetation's landslides mitigation role. The rainfall, a triggering factor of landslides which consists of rainfall at the time of landslide event and antecedent rainfall are critical factors that influence the occurrence of landslides (Yune et al., 2010; Khan et al.,2012; Kim et al., 2021). In this study, we consider time-based aggregated rainfall. The considered variables are illustrated in Table 1.

Table 1. Considered variables for data-driven model construction.

Group	Features	Description	Reference
Vegetation	Fire history	The burning of the vegetation intensifies the mass movement of soil near uncovered burned	(Highland and Bobrowsky, 2008; Culler et al.,2021)

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

Natural Hazards and Earth System Sciences
Open Access
EGU
Discussions

Group	Features	Description	Reference
		stem of trees and free movement on uncovered soil due to post-fire rainfall and storm.	
	Age of tree Forest leafage Forest density Timber diameter (m)	The age of tree combined with the quantity of rainfall may generate higher landslide intensity especially in in trees of age below 10 years. The disturbance of vegetation significantly impacts the susceptibility of landslides in forested regions.	Turner et al., 2010 ; Scheidl et al.,2020
Geomorphology	Drainage	The drainage has a significant effect on the slope stability and promote the efficient control of the influence of rainfall on the ground water fluctuation. The presence of drainage increases the threshold of landslides due to rainfall.	Yan et al., 2019 ; Sun et al.,2010 ; wei et al.,2019
	Erosion	The presence of erosion increases and contributes to the destructive capability of landslides by increasing the volume of transported materials.	Korup et al., 2007
	Slope angle (degree) Slope aspect Slope length (m)	There exists an established relationship between the slope morphology and volume of landslide due to rainfall. The volume increases as the slope length increases. the steeper slopes have lower presence of landslide due to low transportable materials	Qiu et al.,2016 ; Donnarumma et al., 2013
	Soil depth (m) Soil type	Soil properties, depth and texture have a significant difference in infiltration rates which generate different influence on the occurrence of landslides.	Kitutu et al., 2009 ; McKenna et al., 2012
Rainfall	Maximum rain Four weeks rain	Rainfall intensity has an effect on the volume and frequency of landslides being the major	Wieczorek, 1987;

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

Group	Features	Description	Reference
	Three hours rain Three days rain Two weeks rain	triggering factor. The antecedent rainfall and duration of rainfall influence the volume, and deep landslides happen due to rainfall of long duration.	Dai and Lee, 2001; Bernardie et al., 2014; Gariano et al.2017

Variable selection procedure was carried out based on previous literature and applied in the model using variance inflation factor (VIF) (O'Brien, 2007) to eliminate collinear variables. The variable with VIF<10 was considered as non-colinear and hence used in the model. The summary statistics of variables with VIF <10 were summarized in Table 2. The training and test set was scaled (Z-score or variance stability scaling) to solve convergence issues that are associated with running the model without feature scaling (Singh and Singh, 2022). To run the model on the data-data driven methods that accept numerical features, the test and training set was one-hot-encoded to create a feature matrix (Seger, 2018).

Table 2: Summary statistics continuous variables.

Variables	units	N	Min	Mean	Median	Max	Sd
Maximum hourly rain	mm	450	18.5	52.596	61.72	78.5	17.221
Three hours rainfall	mm	450	15	95.044	105.5	171	55.596
Three days rainfall	mm	450	44.5	282.31	283.5	549.5	79.295
two weeks rainfall	mm	450	111.5	388.342	399.5	663	82.854
Four weeks rainfall	mm	450	157.9	586.075	561	1135	158.6
Slope angle	Degree (°)	450	10	34	34.004	65	7.982
Slope length	m	450	1.8	21.246	13	180	22.57
Soil depth	m	450	10	60.311	75	75	20.219
Soil type	constant	450	1.5	1.675	1.7	1.7	0.051
Timber diameter	m	450	0.11	0.227	0.23	0.35	0.146
Age of timber	years	450	15	35.2	35	60	13.392

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

Volume	m^3	450	1.5	599.59	211.68	12663	1237.128

3.2 Method

In this study, nine data-driven methods were selected and tested on a Korean dataset. This section contains a brief introduction to each tested method. The first considered method is OLS, which is applied to estimate parameters of multilinear regression that yield the minimum residual sum of squares errors from the data (Dismuke and Lindrooth 2006) under assumptions of no correlation in independent variables and in error term, constant variance in error terms, non-linear collinearity of predictors, and normal distribution of error terms. The RF-regression is a supervised data-driven technique based on the ensemble learning which construct many decision trees during training time of a model by combining multiple decision trees to produce an improved overall result of the model outcome. The RF-regression is more efficient in the analysis of multidimensional dataset (Borup et al., 2023). RF is an effective predictive model due to non-overfitting characteristics based on the law of large numbers (Breiman, 2001). The decision tree regression is a predictive modeling technique in a form flowchart-like tree structure of all possible results, output, predictor costs, and utility. The DT simplifies the decision-making due to its algorithm that mimic human brain decision making patterns (Rathore and Kumar, 2016). The KNN technique draws an imaginary boundary in which prediction outcomes are allocated as the average of k nearest point predictors and averaging their output variable (response). The KNN calculates Euclidian distances to identify likeness between datapoints and then it groups points that have smaller distances between them (Kramer and Kramer, 2013). The RR is an improved form of ordinary least square which serves to respond to the case where the collinearity is found in predictor variables. The estimated coefficients of ridge are biased estimators of true coefficients and are generated after adding a penalty on the OLS model. The RR has always lower variances compared to OLS (Saleh et al., 2019). The advantage of the GLM over OLS is that the dependent variable need not follow the normal distribution. The GLM is composed by random and systematic components, and the link function that links the two. In this study, the GLM with Gaussian link function was applied. GLM are fitted using maximum likelihood estimation (Dobson and Barnett, 2018). The DNN are among data-driven models that revolutionized different fields; the DNN learns via multi-processing layers and identifies intricate patterns in the data to predict the outcome (LeCun et al.,

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

2015). Here, the backpropagation algorithm was used to predict the estimated outcome. The advantage of DNN is to discover the complex structures in the data using a back propagation algorithm with the capability to change the internal parameter (weight update). The SVM is popular for balanced predictive performance which makes it capable to train model on small sample size. (Pisner and Schnyer, 2020). SVM has been applied in many different landslide studies (Pham et al., 2018; Miao et al., 2018). SVM methods identify the optimal hyperplane in multi-dimensional space that separates different groups in the output values. The EGB is the most powerful and leading supervised machine learning method in solving regression problems. It can perform parallel processing on windows and Linux (Chen et al., 2015). The gradient boosting trains of differentiable loss function, and the model fits when the gradient is minimized. In this paper, both traditional statistical predictive models and machine learning models were used. The firsts are known for high clarity and explain ability, and the second is famous for handling non-linearity in features. In some cases, the performance of advanced data-driven algorithms is almost similar (Chowdhury, 2023).

4. Results

Prior to the construction of the model, the collinearity analysis was performed and variable with less variance inflation factor were retained for training and testing models. Figure 3 depicts retained features and corresponding VIF values. The retained features have VIF less than 10 (O'brien, 2007). All predictors except three days rainfall exhibited VIF less than 5 and still less than 10. Accordingly, all depicted variables were considered for predictive model construction.

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

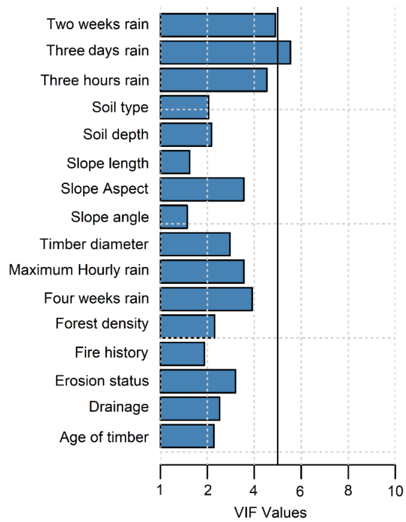

Figure 3. Variance inflation factor bar plot for explanatory variables.

The model was developed in R with different libraries, as discussed below. The DNN regression model was constructed using dnn() function from cito library (Amesoeder et al.,2023), with three hidden layers of (50,50,50) nodes. Model was trained on 208L epochs, learning rate (lr = 0.1), and loss = "mae". The decision tree regression model was constructed with tree() function from tree library, with recursive-partition method.The ridge regression model was constructed using glmnet() function from glmnet library(Jerome et al., 2010). the optimal lambda was obtained by performing 10-fold cross-validation. The EGB model was built using xgboost() function in xgboost packagages (Chen et al., 2022). The optimal model was obtained at 357^{th} boosting iteration with all parameters set to default. The GLM regression model was constructed using glm() function(Team, 2022) with family gaussian and identity link. The KNN regression was constructed using knnreg() function from caret package (Kuhn, 2022, with number of neighbors (k=7). The OLS model was constructed lm() from stats package (Team, 2022). The RF model was run using randomForest() from randomforest package (Liaw and Wiener, 2002), with default parameters and the optimal model was reached at 63^{rd} iteration. The ridge regression model was constructed using glmnet() from glmnet package (Jerome et al., 2012), with ridge penalty(alpha=0). The SVM regression model with linear kernel was built using e1071 package (Meyer et al.,2021) and other parameters set to default.

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

The predictive performance of all tested models was summarized in Fig. 4. The red line represents the perfect prediction. The scatter plot of actual and predicted values of tested models shows that OLS performed least compared to other models with R^2=0.27, that is, 27% of variance in the model could be explained by predictor variables. The second least performing was GLM with R^2=0.29 that is 2% improvement compared to OLS. Among all models five out of nine, namely, OLS, KNN, GLM, SVM, and RR, performed below 50%; however, these models predicted well small values of volume (below 2000m^3). The MAE of these five models was higher than the remaining four models, namely DT, RF, DNN and EGB. Among these lasts, the most performing was EGB with R^2=0.85 of variance explained by predictors and MAE=245.6 m^3. The summary of coefficients of determination and mean absolute errors for tested models are summarized in Table 3.

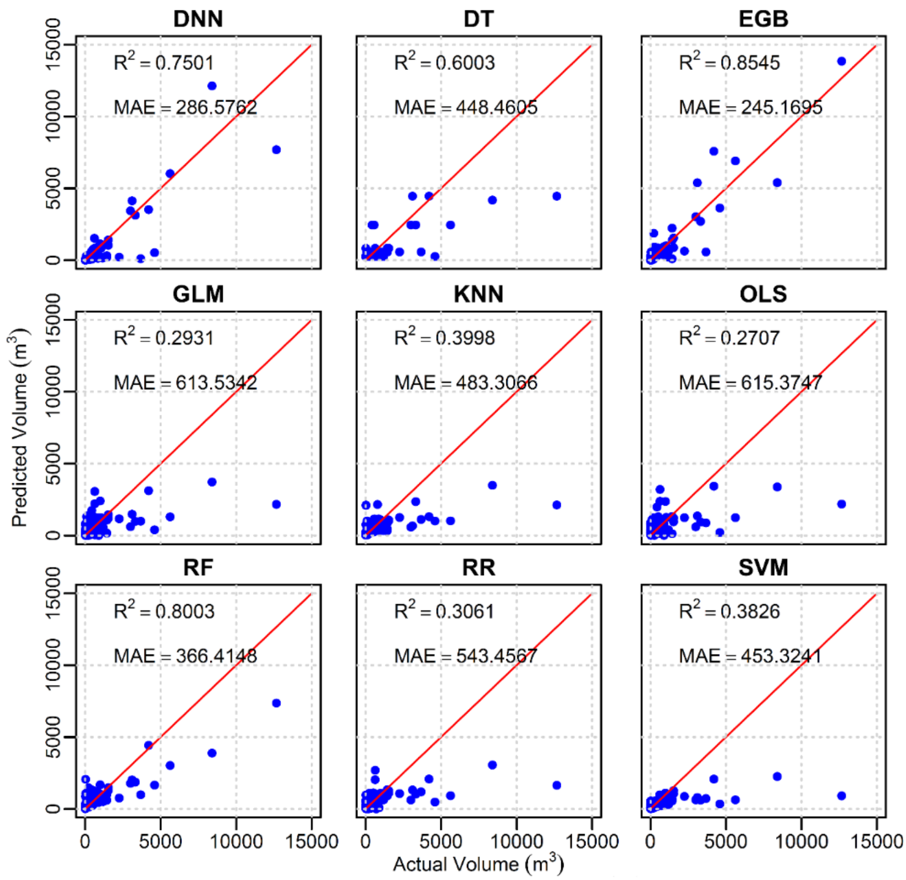

Figure 4. Scatterplot of actual and predicted values for nine tested models.

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

Table 3. Summary of R2 and MAE for tested models.

Models	DNN	DT	EGB	GLM	KNN	OLS	RF	RR	SVM
R2	0.7501	0.6003	0.8545	0.2931	0.3998	0.2707	0.8003	0.3061	0.3826
MAE	286.5762	448.4605	245.1695	613.5342	483.3066	615.3747	366.4148	543.4567	453.3241

To dive deep into the prediction performance of the EGB model, we analyzed variables importance in the prediction of the volume. It was observed that the slope length was the most contributing predictor in the performance of the EGB model, followed by the slope aspect. The presence and quality of drainage ranked the third most contributor in the prediction of the volume of rainfall due to landslides. In addition, age of timber (age of trees that were planted on the area that faced landslides) and maximum hourly rainfall have also shown a significant contribution in the prediction of volume of landslide due to rainfall. Figure 5 illustrates a list of independent variables that had a significant impact in the prediction of the volume.

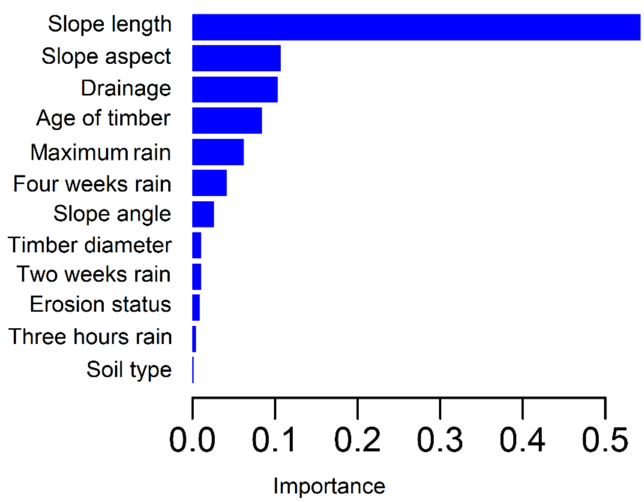

Figure 5. Variable importance for the EGB model.

The variable importance plot depicts the overall contribution of a given variable however, it does not provide detailed information. To get more insight into the relationship between the volume of landslides and predictors, statistical tests for normality, namely, Shapiro-Wilk's test,

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

Natural Hazards and Earth System Sciences
Open Access
Discussions

Kruskal-Wallis test, and Dunn's test were conducted. The Shapiro-Wilk's test (Dudley,2023) results revealed that the distribution of volume was non-normal ($W = 0.40642$, p-value < 0.001). Noting that the volume distribution was non-normal, we opted for the non-parametric tests, which do not rely on normality to conduct the distance correlation (Székely et al.,2007) test (dcor) for continuous independent features. Figure 6 illustrates that the slope length exhibited higher value (dcor=0.51) followed by rainfall features. This highlights the role of current and antecedent rainfall as triggering factor in the prediction of volume of landslides.

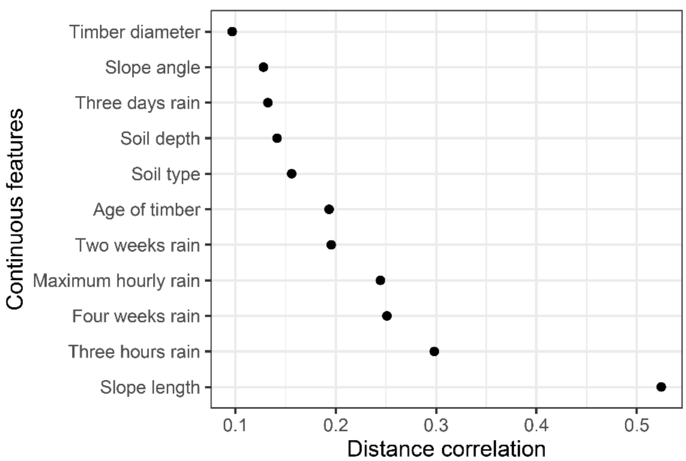

Figure 6. Distance correlation plot for the volume and continuous features.

Furthermore, to test for categorical features, Kruskal-Wallis test (McKight and Najab, 2010) was used to check whether the volume of landslide was different in each category and Dunn's tests (Dinno, 2015) were applied to examine which categories had similar means of the volume of landslides due to rainfall in different categories. The H_0 (null hypothesis) was that the mean volume of landslides in different categories is the same, and the H_1 (alternative hypothesis) was that the means of landsides are different in some categories. For the slope aspect, the second most significant predictor for the EGB model, the results of Kruskal-Wallis test (chi-squared = 20.889, df = 7, p-value = 0.003938) showed that there is a significant difference in median of volume in some categories of slope aspects. To know which classes of slope aspects had significantly different mean volumes, the Dunn's test results at 95% confidence interval, pairs (East-South west, East-South East, East-South, East-North West and North West-South East) had significantly

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

Natural Hazards and Earth System Sciences
Discussions

different means of landslides' volume (with p-value <0.05). Figure 7 depicts that the southwest and southeast aspects had a higher frequency of landslides.

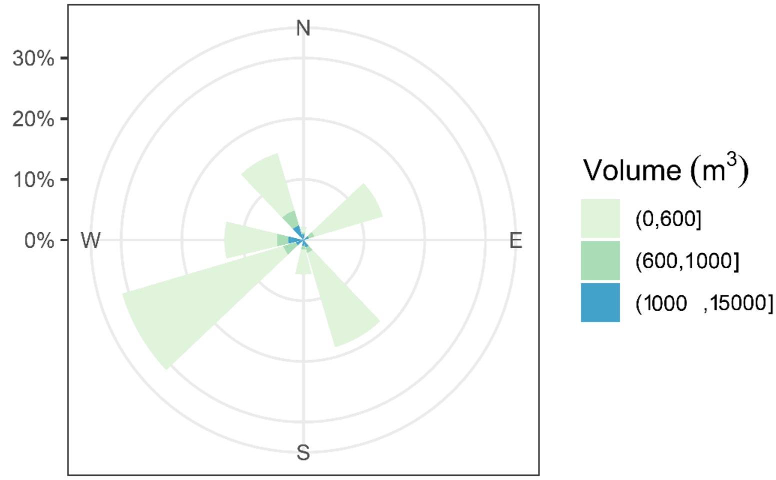

Figure 7. The distribution of volume of landslides due to rainfall with respect to the slope aspect.

The Kruskal-Wallis test for the difference in mean of drainage classes the result was: chi-squared = 15.792, df = 2, p-value = 0.000372 which shows that the means of volume per classes were different. This was clarified by Dunn's test results were p-values were less than 0.05 in all pairwise mean difference comparisons. The results of these tests highlighted that the drainage has a remarkable influence on the occurrence of rainfall-induced landslides in the Korean Peninsula.

5. Discussion

This study aim was to construct data data-driven algorithm that predict the volume of landslide due to rainfall. The result of nine different tested algorithms revealed a tremendous difference between classical regression models (OLS, RR, and GLM) and other data driven machine learning models. In this study, apart from SVM regression and KNN, other machine learning models (DNN, DT, RF, and EGB) exhibited high prediction capability with R^2 above 50% (Fig.3). The random forest model performed well in predicting smaller volume however as the volume increased the model underpredicted volume values. The DNN model performed quite well with low MAE compare to random forest however the model did not perform on well moderate volume values which resulted in reduction of R^2. The EGB model tested on South Korean landslide inventory

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

coupled with rainfall data at the time of landslide events and antecedent rainfall within one month of the event exhibited the highest performance compared to other constructed algorithms.

The slope aspect played an important role in prediction of the volume and the landslide mostly occurred on location oriented toward south west and south east. That may be due to the direction taken by typhoon which hit the south west versants of mountains upon landfall on the Korean peninsula toward North East Pacific (Ha, 2022, Lee et al., 2013). The findings of this research are congruent with Lee et al. (2013) who also highlighted that the mountain versant oriented to strong wind direction may face more landslides. The study also highlighted that the efficacy of drainage plays an important role in the prevention of landslides which due to the stabilizing effect.

The occurrence of landslides triggered by rainfall is a complex phenomenon which involve many interrelated environmental setting human activity, geological conditions and climatic conditions. Moreover, the occurrence of typhoons is known to aggravate the landslides impacts on communities (Chang et al., 2008), incorporating typhoon variables in future studies to customize for regional setting may improve the accuracy of the model. The advantage of his research is that the constructed model has high predictive accuracy and can handle the non-linearity of predisposing factors. The model came to fill the gap of few literatures related to the prediction of volume of landslides using data-driven techniques. This model can be a better tool to help policy makers to integrate the landslides volume risks in in policy to protect infrastructure and inhabitants dwelling near foot of mountains with high risks of being buried by geological materials resulting from landslides.

6. Conclusions

In this paper, the aim was to construct the data driven model that predict the volume of landslides due to rainfall. To this, nine different classical regression models and machine learning algorithms were tested on South Korean landslide data set containing features of landslides that occurred between 2011 and 2012. Among tested models, Extreme gradient boosting (EGB) produced most accurate prediction. This is proven by the evaluation of the difference between actual and predicted values were R^2 was 0.8545, and MAE was 245.1695m^3. The analysis of feature variables in the contribution to the prediction of the model, revealed that the slope length was the most influencing predictor. The EGB model can be a promising tool for the prediction of the volume of landslide due to its high predictive performance. The model can be customized on different environmental settings. The model can be applied to estimate the expected volume of landslides based on forecasted rainfall once the model is well-adjusted to fit the geomorphological and environmental settings of the region of interest. Therefore, this model can be a better tool for planning for resilience and infrastructure pre-construction risk assessment to ensure the new infrastructure is placed in stable regions free from severe landslides.

Competing Interests

The contact author has declared that none of the authors has any competing interests.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1A6A1A03044326), and the National Research Foundation of Korea (NRF) grant (2021R1C1C2003316) funded by the Korea government (Ministry of Science and ICT).

Reference

Alcantara, A. L., and Ahn, K. H. (2020). Probability distribution and characterization of daily precipitation related to tropical cyclones over the Korean Peninsula. Water, 12(4), 1214.

Alcántara-Ayala, I. (2021). Integrated landslide disaster risk management (ILDRiM): the challenge to avoid the construction of new disaster risk. Environmental Hazards, 20(3), 323-344.

Amesoeder, C., Hartig, F., and Pichler, M. (2023). cito: An R package for training neural networks using torch. arXiv e-prints, arXiv-2303.

Bernardie, S., Desramaut, N., Malet, J.-P., Gourlay, M., and Grandjean, G. (2014). Prediction of changes in landslide rates induced by rainfall. Landslides, 12(3), 481–494. doi:10.1007/s10346-014-0495-8

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

Bonamutial, M., and Prasetyo, S. Y. (2023, August). Exploring the Impact of Feature Data Normalization and Standardization on Regression Models for Smartphone Price Prediction. In 2023 International Conference on Information Management and Technology (ICIMTech) (pp. 294-298). IEEE.

Borup, D., Christensen, B. J., Mühlbach, N. S., and Nielsen, M. S. (2023). Targeting predictors in random forest regression. International Journal of Forecasting, 39(2), 841-868.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Chang, K. T., and Chiang, S. H. (2009). An integrated model for predicting rainfall-induced landslides. Geomorphology, 105(3-4), 366-373.

Chang, K. T., Chiang, S. H., and Lei, F. (2008). Analysing the relationship between typhoon-triggered landslides and critical rainfall conditions. Earth Surface Processes and Landforms: The Journal of the British Geomorphological Research Group, 33(8), 1261-1271

Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, Li M, Xie J, Lin M, Geng Y, Li Y, Yuan J (2022). _xgboost: Extreme Gradient Boosting_. R package version 1.6.0.1, <https://CRAN.R-project.org/package=xgboost>.

Chen, C. W., Oguchi, T., Hayakawa, Y. S., Saito, H., and Chen, H. (2017). Relationship between landslide size and rainfall conditions in Taiwan. Landslides, 14, 1235-1240.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... and Zhou, T. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1-4.

Chen, Z., Luo, R., Huang, Z., Tu, W., Chen, J., Li, W., ... and Ai, Y. (2015). Effects of different backfill soils on artificial soil quality for cut slope revegetation: Soil structure, soil erosion, moisture retention and soil C stock. Ecological engineering, 83, 5-12.

Chowdhury, M. Z. I., Leung, A. A., Walker, R. L., Sikdar, K. C., O'Beirne, M., Quan, H., and Turin, T. C. (2023). A comparison of machine learning algorithms and traditional regression-based statistical modeling for predicting hypertension incidence in a Canadian population. Scientific Reports, 13(1), 13.

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

Natural Hazards and Earth System Sciences
Open Access
Discussions

Conte, E., Pugliese, L., and Troncone, A. (2022). A simple method for predicting rainfall-induced shallow landslides. Journal of Geotechnical and Geoenvironmental Engineering, 148(10), 04022079

Culler, E. S., Livneh, B., Rajagopalan, B., and Tiampo, K. F. (2021). A data-driven evaluation of post-fire landslide susceptibility. Natural Hazards and Earth System Sciences Discussions, 1-24.

Dai, F. C., and Lee, C. F. (2001). Frequency–volume relation and prediction of rainfall-induced landslides. Engineering geology, 59(3-4), 253-266.

Dai, K., Xu, Q., Li, Z., Tomás, R., Fan, X., Dong, X., ... and Ran, P. (2019). Post-disaster assessment of 2017 catastrophic Xinmo landslide (China) by spaceborne SAR interferometry. Landslides, 16, 1189-1199.

Dinno, A. (2015). Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. The Stata Journal, 15(1), 292-300.

Dismuke, C., and Lindrooth, R. (2006). Ordinary least squares. Methods and designs for outcomes research, 93(1), 93-104.

Dobson, A. J., and Barnett, A. G. (2018). An introduction to generalized linear models. CRC press

Donnarumma, A., Revellino, P., Grelle, G., and Guadagno, F. M. (2013). Slope angle as indicator parameter of landslide susceptibility in a geologically complex area. Landslide Science and Practice: Volume 1: Landslide Inventory and Susceptibility and Hazard Zoning, 425-433.

Dudley, R. (2023). The Shapiro–Wilk test for normality

Evans, S. G., Mugnozza, G. S., Strom, A., and Hermanns, R. L. (Eds.). (2007). Landslides from massive rock slope failure (Vol. 49). Springer Science and Business Media.

Fan, J. R., Zhang, X. Y., Su, F. H., Ge, Y. G., Tarolli, P., Yang, Z. Y., ... and Zeng, Z. (2017). Geometrical feature analysis and disaster assessment of the Xinmo landslide based on remote sensing data. Journal of Mountain Science, 14(9), 1677-1688.

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

Gariano, S. L., Rianna, G., Petrucci, O., and Guzzetti, F. (2017). Assessing future changes in the occurrence of rainfall-induced landslides at a regional scale. Science of the total environment, 596, 417-426.

Giarola, A., Meisina, C., Tarolli, P., Zucca, F., Galve, J. P., and Bordoni, M. (2024). A data-driven method for the estimation of shallow landslide runout. Catena, 234, 107573.

Gong, Q., Wang, J., Zhou, P., and Guo, M. (2021). A regional landslide stability analysis method under the combined impact of rainfall and vegetation roots in south China. Advances in Civil Engineering, 2021, 1-12.

Gonzalez-Ollauri, A., and Mickovski, S. B. (2017). Hydrological effect of vegetation against rainfall-induced landslides. Journal of Hydrology, 549, 374-387.

Ha, K. M. (2022). predicting typhoon tracks around Korea. Natural Hazards, 113(2), 1385-1390.

Highland, L. and Bobrowsky, P.: The Landslide Handbook: A Guide to Understanding Landslides, United States Geological Survey, Reston, VA, Circular 1325, https://pubs.usgs.gov/circ/1325/ (last access: 6 March 2023), 2008. a, b

Hovius, N., Stark, C. P., Tutton, M. A., and Abbott, L. D. (1998). Landslide-driven drainage network evolution in a pre-steady-state mountain belt: Finisterre Mountains, Papua New Guinea. Geology, 26(12), 1071-1074.

Intrieri, E., Carlà, T., and Gigli, G. (2019). Forecasting the time of failure of landslides at slope-scale: A literature review. Earth-science reviews, 193, 333-349.

Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL:<https://www.jstatsoft.org/v33/i01/>.

Ju, L. Y., Zhang, L. M., and Xiao, T. (2023). Power laws for accurate determination of landslide volume based on high-resolution LiDAR data. Engineering Geology, 312, 106935.

Jung, Y., Shin, J. Y., Ahn, H., and Heo, J. H. (2017). The spatial and temporal structure of extreme rainfall trends in South Korea. Water, 9(10), 809.

Kafle, L., Xu, W. J., Zeng, S. Y., & Nagel, T. (2022). A numerical investigation of slope stability influenced by the combined effects of reservoir water level fluctuations and precipitation: A case study of the Bianjiazhai landslide in China. Engineering Geology, 297, 106508.

Keefer, R. F. (2000). Handbook of soils for landscape architects. Oxford University Press.

Khan, M. A., Basharat, M., Riaz, M. T., Sarfraz, Y., Farooq, M., Khan, A. Y., ... and Shahzad, A. (2021). An integrated geotechnical and geophysical investigation of a catastrophic landslide in the Northeast Himalayas of Pakistan. Geological Journal, 56(9), 4760-4778.

Khan, Y. A., Lateh, H., Baten, M. A., and Kamil, A. A. (2012). Critical antecedent rainfall conditions for shallow landslides in Chittagong City of Bangladesh. Environmental Earth Sciences, 67, 97-106.

Kim, S. W., Chun, K. W., Kim, M., Catani, F., Choi, B., and Seo, J. I. (2021). Effect of antecedent rainfall conditions and their variations on shallow landslide-triggering rainfall thresholds in South Korea. Landslides, 18, 569-582.

Kitutu, M. G., Muwanga, A., Poesen, J., and Deckers, J. A. (2009). Influence of soil properties on landslide occurrences in Bududa district, Eastern Uganda. African journal of agricultural research, 4(7), 611-620.

Korup, O., Clague, J. J., Hermanns, R. L., Hewitt, K., Strom, A. L., and Weidinger, J. T. (2007). Giant landslides, topography, and erosion. Earth and Planetary Science Letters, 261(3-4), 578-589.

Kramer, O., and Kramer, O. (2013). K-nearest neighbors. Dimensionality reduction with unsupervised nearest neighbors, 13-23.

Kuhn M (2022). _caret: Classification and Regression Training_. R package version 6.0-92, <https://CRAN.R-project.org/package=caret>

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

Lee, D. H., Cheon, E., Lim, H. H., Choi, S. K., Kim, Y. T., and Lee, S. R. (2021). An artificial neural network model to predict debris-flow volumes caused by extreme rainfall in the central region of South Korea. Engineering Geology, 281, 105979.

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

Natural Hazards and Earth System Sciences
Open Access
Discussions

Lee, D. H., Kim, Y. T., and Lee, S. R. (2020). Shallow landslide susceptibility models based on artificial neural networks considering the factor selection method and various non-linear activation functions. Remote Sensing, 12(7), 1194.

Lee, M. J. (2016). Rainfall and landslide correlation analysis and prediction of future rainfall base on climate change. In Geohazards Caused by Human Activity. IntechOpen.

Lee, S. G. (2009). The Effects of Landslide in South Korea and Some Issues for Successful Management and Mitigation. 한국토양비료학회 학술발표회 초록집, 181-191.

Lee, S. G., and Winter, M. G. (2019). The effects of debris flow in the Republic of Korea and some issues for successful risk reduction. Engineering geology, 251, 172-189.

Lee, S. W., Kim, G., Yune, C. Y., and Ryu, H. J. (2013). Development of landslide-risk assessment model for mountainous regions in eastern Korea. Disaster advances, 6(6), 70-79.

Liaw, A., and Wiener, M., (2002). Classification and regression by randomForest. R News 2(3), 18--22.

Martinović, K., Gavin, K., Reale, C., and Mangan, C. (2018). Rainfall thresholds as a landslide indicator for engineered slopes on the Irish Rail network. Geomorphology, 306, 40-50.

McKenna, J. P., Santi, P. M., Amblard, X., and Negri, J. (2012). Effects of soil-engineering properties on the failure mode of shallow landslides. Landslides, 9, 215-228.

McKight, P. E., and Najab, J. (2010). Kruskal-wallis test. The corsini encyclopedia of psychology, 1-1.

Melo, R., Zêzere, J. L., Rocha, J., and Oliveira, S. C. (2019). Combining data-driven models to assess susceptibility of shallow slides failure and runout. Landslides, 16, 2259-2276.

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2021). _e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien_. R package version 1.7-9, <https://CRAN.R-project.org/package=e1071>.

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

Miao, F., Wu, Y., Xie, Y., and Li, Y. (2018). Prediction of landslide displacement with step-like behavior based on multialgorithm optimization and a support vector regression model. Landslides, 15, 475-488.

O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. Quality and quantity, 41, 673-690.

Omwega, A. K. (1989). Crop cover, rainfall energy and soil erosion in Githunguri (Kiambu District), Kenya. The University of Manchester (United Kingdom).

Paudel, P. P., Omura, H., Kubota, T., and Morita, K. (2003). Landslide damage and disaster management system in Nepal. Disaster Prevention and Management: An International Journal, 12(5), 413-419.

Peruzzetto, M., Mangeney, A., Grandjean, G., Levy, C., Thiery, Y., Rohmer, J., and Lucas, A. (2020). Operational estimation of landslide runout: comparison of empirical and numerical methods. Geosciences, 10(11), 424.

Pham, B. T., Tien Bui, D., and Prakash, I. (2018). Bagging based support vector machines for spatial prediction of landslides. Environmental Earth Sciences, 77, 1-17.

Phillips, C., Hales, T., Smith, H., and Basher, L. (2021). Shallow landslides and vegetation at the catchment scale: A perspective. Ecological Engineering, 173, 106436.

Pisner, D. A., and Schnyer, D. M. (2020). Support vector machine. In Machine learning (pp. 101-121). Academic Press.

Qiu, H., Regmi, A. D., Cui, P., Cao, M., Lee, J., and Zhu, X. (2016). Size distribution of loess slides in relation to local slope height within different slope morphologies. Catena, 145, 155-163.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.

Rathore, S. S., and Kumar, S. (2016). A decision tree regression-based approach for the number of software faults prediction. ACM SIGSOFT Software Engineering Notes, 41(1), 1-6.

Razakova, M., Kuzmin, A., Fedorov, I., Yergaliev, R., and Ainakulov, Z. (2020). Methods of calculating landslide volume using remote sensing data. In E3S Web of Conferences (Vol. 149, p. 02009). EDP Sciences.

Rosi, A., Peternel, T., Jemec-Auflič, M., Komac, M., Segoni, S., and Casagli, N. (2016). Rainfall thresholds for rainfall-induced landslides in Slovenia. Landslides, 13, 1571-1577.

Rotaru, A., Oajdea, D., and Răileanu, P. (2007). Analysis of the landslide movements. International journal of geology, 1(3), 70-79.

Saleh, A. M. E., Arashi, M., and Kibria, B. G. (2019). Theory of ridge regression estimation with applications. John Wiley and Sons.

Scheidl, C., Heiser, M., Kamper, S., Thaler, T., Klebinder, K., Nagl, F., ... and Seidl, R. (2020). The influence of climate change and canopy disturbances on landslide susceptibility in headwater catchments. Science of the total environment, 742, 140588.

Seger, C. (2018). An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing.

Singh, D., and Singh, B. (2022). Feature wise normalization: An effective way of normalizing data. Pattern Recognition, 122, 108307.

Sun, H. Y., Wong, L. N. Y., Shang, Y. Q., Shen, Y. J., and Lü, Q. (2010). Evaluation of drainage tunnel effectiveness in landslide control. Landslides, 7, 445-454.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances.

Turner, T. R., Duke, S. D., Fransen, B. R., Reiter, M. L., Kroll, A. J., Ward, J. W., ... and Bilby, R. E. (2010). Landslide densities associated with rainfall, stand age, and topography on forested landscapes, southwestern Washington, USA. Forest Ecology and Management, 259(12), 2233-2247.

Van Tien, P., Luong, L. H., Duc, D. M., Trinh, P. T., Quynh, D. T., Lan, N. C., ... & Loi, D. H. (2021). Rainfall-induced catastrophic landslide in Quang Tri Province: the deadliest single landslide event in Vietnam in 2020.

https://doi.org/10.5194/nhess-2024-90
Preprint. Discussion started: 11 July 2024

Natural Hazards and Earth System Sciences
Discussions

Wei, Z. L., Shang, Y. Q., Sun, H. Y., Xu, H. D., and Wang, D. F. (2019). The effectiveness of a drainage tunnel in increasing the rainfall threshold of a deep-seated landslide. Landslides, 16, 1731-1744.

Wieczorek, G. (1987). In central Santa Cruz Mountains, California. Debris flows/avalanches: process, recognition, and mitigation, 7, 93.

Yan, L., Xu, W., Wang, H., Wang, R., Meng, Q., Yu, J., and Xie, W. C. (2019). Drainage controls on the Donglingxing landslide (China) induced by rainfall and fluctuation in reservoir water levels. Landslides, 16, 1583-1593.

Yune, C. Y., Jun, K. J., Kim, K. S., Kim, G. H., and Lee, S. W. (2010). Analysis of slope hazard-triggering rainfall characteristics in Gangwon Province by database construction. Journal of the Korean Geotechnical Society, 26(10), 27-38.

Zaruba, Q., and Mencl, V. (2014). Landslides and their control. Elsevier.