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

Tuganishuri Jérémie[1], Chan-Young Yune[2], Gihong Kim[3], Seung Woo Lee[4], Manik Das Adhikari[5], and, Sang-Guk Yum[6*]

[1]Department of Civil and Environmental Engineering, Gangneung-Wonju National University, Gangneung, Gangwon 25457, South Korea

[2]Department of Civil and Environmental Engineering, Gangneung-Wonju National University, Gangneung, Gangwon 25457, South Korea

[3]Department of Civil and Environmental Engineering, Gangneung-Wonju National University, Gangneung, Gangwon 25457, South Korea

[4]Department of Civil and Environmental Engineering, Gangneung-Wonju National University, Gangneung, Gangwon 25457, South Korea

[5]Department of Civil and Environmental Engineering, Gangneung-Wonju National University, Gangneung, Gangwon 25457, South Korea

[6]Department of Civil and Environmental Engineering, Gangneung-Wonju National University, Gangneung, Gangwon 25457, South Korea

Correspondence to: Sang-Guk Yum (skyeom0401@gwnu.ac.kr)

Abstract. Landslides due to rainfall are among the 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. The objectives of this research are to construct a model using advanced data-driven algorithms (i.e., ordinary least squares or Linear regression (OLS), random forest (RF), support vector machine (SVM), extreme gradient boosting (EGB), generalized linear model (GLM), decision tree (DT), 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 trained and tested on South Korean landslide dataset, with the EGB predictions yielding the highest coefficient of determination ($R^2 = 0.8841$) and the lowest mean absolute error (MAE = 146.6120 m^3), followed by RF predictions ($R^2 = 0.8435$, MAE = 330.4876 m^3) on the holdout set. The results indicated that the DNN, EGB, and RF models exhibited R^2>0.8 on both the training and test sets. The difference in coefficient of determination R^2 on the training and holdout set were 1.75, 7.72, and 12.17% for RF, EGB and DNN, respectively, signifying that these models could yield reliable volume estimates in adjacent areas with similar geomorphological and environmental settings. The volume of landslides was strongly influenced by slope length, maximum hourly rainfall, slope angle, aspect, and altitude. The anticipated volume of landslides can be important for land use allocation and efficient landslide risk

management.

Keywords: Data-driven models, shallow landslides, predictive machine learning, Rainfall-Induced Slope Failure, Landslides Volume Prediction, South Korea

1. Introduction

Landslides due to rainfall are phenomena that dislocate a mass of soil from its natural position and slide downward along a slope due to gravity forces. Intense or long-duration rainfall infiltrates the soil and increases the pore pressure, resulting in soil saturation that leads to slope failure. The saturated soil becomes weak and loses cohesion, and the slope fails when rainfall crosses a certain threshold (Bernardie et al., 2014; Martinović et al., 2018; Lee et al., 2021). The heavy rainfall saturates a slope and triggers a landslide due to the reduction of the soil's shear strength and the increase of pore water pressure (Tsai and Chen, 2010; Lacerda et al., 2014; Chatra et al., 2019; Chen et al., 2021; Luino et al., 2022). For example, steep slopes with loose soils and even moderate rainfall can lead to the displacement of an enormous quantity of soil mass. On the contrary, in slopes with more stable, cohesive soils, the surface failure might be smaller (Tsai and Chen, 2010). 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 surface failure (Chang and Chiang, 2009; Bernardie et al., 2014; Chen et al., 2017). The landslide occurrences can also be influenced by human activities that weaken the slope, such as excavation at the slope toe and loading caused by construction and land use such as agriculture, mining etc. (Rosi et al., 2016). The rapid urbanization activities in mountainous regions affect the topography through hill cutting, deforestation and water drainage (Rahman et al., 2017); these activities disturb the slope structure and change the water flow, which exacerbates the effect of landslides in regions where human engineering activities are mostly located (Holcombe et al., 2016; Chen et al., 2019). Therefore, to mitigate landslide-induced risks in the runout regions, estimation of the volume of landslides due to rainfall (VLDR) plays a crucial role.

The quantification of the VLDR is essential for effective risk management (Tacconi Stefanelli et al., 2020), emergency response, engineering design (Cheung, 2021), economic assessment and environmental protection (Alcántara-Ayala and Sassa, 2023). With the estimates of VLDR, the morphologist can update hazard maps (Van Westen, 2000) to reflect the scale of potential mass movement in various regions to obtain regions with similar likelihood of landslides of similar soil mass to highlight risk zone levels, i.e., low, moderate and high. These classifications help engineers to apply appropriate slope stabilization techniques depending on the level of risk (Dahal and Dahal, 2017). Additionally, enhancing the precision of VLDR estimations and improving the predictive capabilities is essential for understanding and monitoring landscape evolution. Montgomery (2009) emphasized that the volume of landslides is a key factor in determining the extent of downstream damage, particularly for large debris flows or rock avalanches, which can drastically alter the landscape and affect surrounding ecosystems and infrastructure. Similarly, Korup (2004) further explored the long-term geomorphological effects of large-volume landslides, highlighting their importance in reshaping mountainous terrains and influencing sediment transport, which is critical for understanding both immediate and future landscape changes. However, the existing landslide susceptibility models mostly used for the identification of regions susceptible to landslides (i.e., landslide zonation) (Kim et al., 2014; Gutierrez-Martin, 2020; Chen et al., 2021; Li et al., 2022), which are essential in emergency management because they provide a general overview of zones with a higher probability of landslide occurrence without emphasizing on the determination of the approximate value of the volume of failing mass in relation to excessive rainfall events.

Numerous researchers used landslide inventory, remote sensing data and numerical techniques to establish the relationship between landslide geometry and the influencing factors to determine the landslide volume quantitatively. For example, Saito et al. (2014) studied the relationship between rainfall-triggered landslides to test whether the volume of landslides across Japan that occurred between 2001 and 2011 can be directly predicted from rainfall metrics. The findings revealed that larger landslides occurred when rainfall exceeded certain thresholds, but there were significant discrepancies between peaks of rainfall metrics and maximum landslide volumes, and the total rainfall was the suitable predictor of landslides. Dai and Lee (2001) established the frequency-volume relation for landslides in Hong Kong and noticed that the relation for shallow landslides above $4m^3$ followed the power law. The 12-hour rolling rainfall contributed most to the prediction of the volume of landslides. Jaboyedoff et al. (2012) contributed by demonstrating the value of remote sensing technologies such as Light Detection and Ranging (LiDAR) in conjunction with field data to improve the accuracy of volume estimates and capture the geomorphological changes associated with landslides. Ju et al. (2023) constructed an area-volume power law model for the estimation of the volume of landslides using high-resolution LiDAR data collected between 2010 and 2020 in Hong Kong. The aim was to estimate accurately the volume of landslides on small-scale landslides. The reliance on localized datasets limits the model's applicability in regions with different geological settings, and the model does not consider all variabilities of landslide characteristics. Razakova et al. (2020) calculated landslide volume using remote sensing data to assess the efficiency of aerial photographs in environmental impact assessment and ground-based measurement. The study did not consider the effect of vegetation and topography and only focused on a single landslide case, which may be a source of bias due to differences in soil composition and environmental factors. Hovius et al. (1997) analyzed multiple sets of aerial photos and frequency-magnitude relations for landslides in New Zealand. The finding pinpointed that the landslides frequency-magnitude followed power law and infrequent large magnitude contributed to the landscape change. The study highlighted the importance of soil composition in landslide size, but the reliance on aerial photos, which may be inaccurate in dense forest areas, and the omission of climatic factors limit the generality of the findings. Guzzetti et al. (2008) applied statistical methods on regional landslide inventories and antecedent rainfall data ranging between 10 min to 35 days. The findings revealed that the slope angle and soil type significantly influence landslide volume estimates, and the rainfall intensity is more important than duration. Chatra et al. (2019) applied numerical methods to study the effect of rainfall duration and intensity on the generation of pore pressure in the soil; the finding revealed a higher instability in loose soil compared to medium soil slopes. Huang et al. (2020) introduced a hybrid machine-learning model combining support vector regression (SVR) with a genetic algorithm to estimate debris-flow volumes. The model was tested on real-world case studies, showing improved accuracy in volume predictions compared to traditional methods. However, it was noticed that the study relied on a limited dataset, which may reduce the model's generalizability to other regions of different geomorphology and environmental settings. Shirzadi et al.

(2017) compared the effectiveness of statistical and machine-learning models in simulating landslide volumes-areal relations, demonstrating that machine-learning techniques outperform traditional statistical methods in terms of accuracy. The study did not consider the climatic and geomorphic factors such as rainfall, vegetation, soil type, etc., triggering and influencing factors for the landslide occurrence. It was noted that existing models only treated the interaction of soil and rainfall without considering the environmental factors, human activity, and non-linear behavior of the triggering and influencing factors.

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., geomorphological, soil and environmental). Here, we aim to construct a data-driven algorithm that combines input parameters for physical-based and empirical models and incorporates 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 a 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 (ML) algorithms and compare their performance using the coefficient of determinations (R^2), mean square errors (MAE), Root mean square error (RMSE), Mean absolute percentage error (MAPE), and symmetric mean absolute percentage errors (SMAPE) of the predicted volume of landslides. The focus is to optimize the predictions of the volume of landslides due to rainfall, taking into account triggering and influencing factors with higher accuracy.

2. Data and Study Region

2.1. Study Region

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., 2022). South Korea is located on the southern part of the Korean Peninsula, bordered by the Yellow Sea to the west coast and the East Sea (Sea of Japan) to the East. According to the Korean Meteorological Administration (https://www.kma.go.kr/), the country has a temperate climate characterized by four distinct seasons: hot and humid summers, cold winters, and springs and falls with moderate temperatures. The annual rainfall varies between 1000 mm to 1400 mm and 1800 mm for the central region and southern region, respectively (Jung et al., 2017; Alcantara and Ahn, 2020). During the summer, heavy rainfall from June to September leads to significant surface runoff, increases landslide risk, and causes approximately 95% of all landslides each year (Lee et al., 2020; Park and Lee, 2021). 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 (Kim and Park, 2021). The rainfall trend analysis from 1971 to 2100 predicted an increase in rainfall of 271.23mm, which indicates the growing risk of landslides associated with climate change (Lee, 2016). Temperature variations are influenced by its geographical location; the average summer temperatures vary between 25 and 30°C, while winter temperatures can drop to -10°C in some parts of the country (https://web.kma.go.kr/). The South Korean geologically is mainly composed of granitic and metamorphic rocks, such as gneiss, schist, and granite, which influence the stability of the landscape (Jung et al., 2024). The geomorphology is characterized by rugged mountains, river valleys, and coastal plains, with the Taebaek Mountains running along the eastern edge (Kim et al., 2020). The influence of rainfall, environmental, geomorphology, and geological factors increase the vulnerability to landslides across the country, especially in the northeastern mountainous region, as depicted in Figure 1. The predominant soil types in South Korea include clay, sandy, and loamy soils, each with different characteristics affecting water infiltration, retention and erosion (Kang et al., 2022; Lee et al., 2023). Clay soils, being more stable, can become highly saturated, increasing landslide risk during heavy rains. On the other hand, sandy soils are loose and more prone to shallow landslides during light rainfall. Regions with steep topography and poorly consolidated soil (loose) are mostly at risk, especially after prolonged rainfalls (Kim et al., 2015).

The combination of heavy summer rainfall, geological composition, and geomorphological factors makes South Korea particularly vulnerable to shallow landslides. Thus, continuous monitoring and research are vital to understanding the complex interactions between climate, geology, soil types, and landslide occurrences in this region. Understanding the collective effects of meteorological, environmental, geological stability, and geomorphological features is crucial for developing effective disaster management strategies and enhancing public safety in landslide-prone areas. As climate change continues to impact rainfall patterns, South Korea faces ongoing challenges in mitigating landslide risks and protecting vulnerable communities.

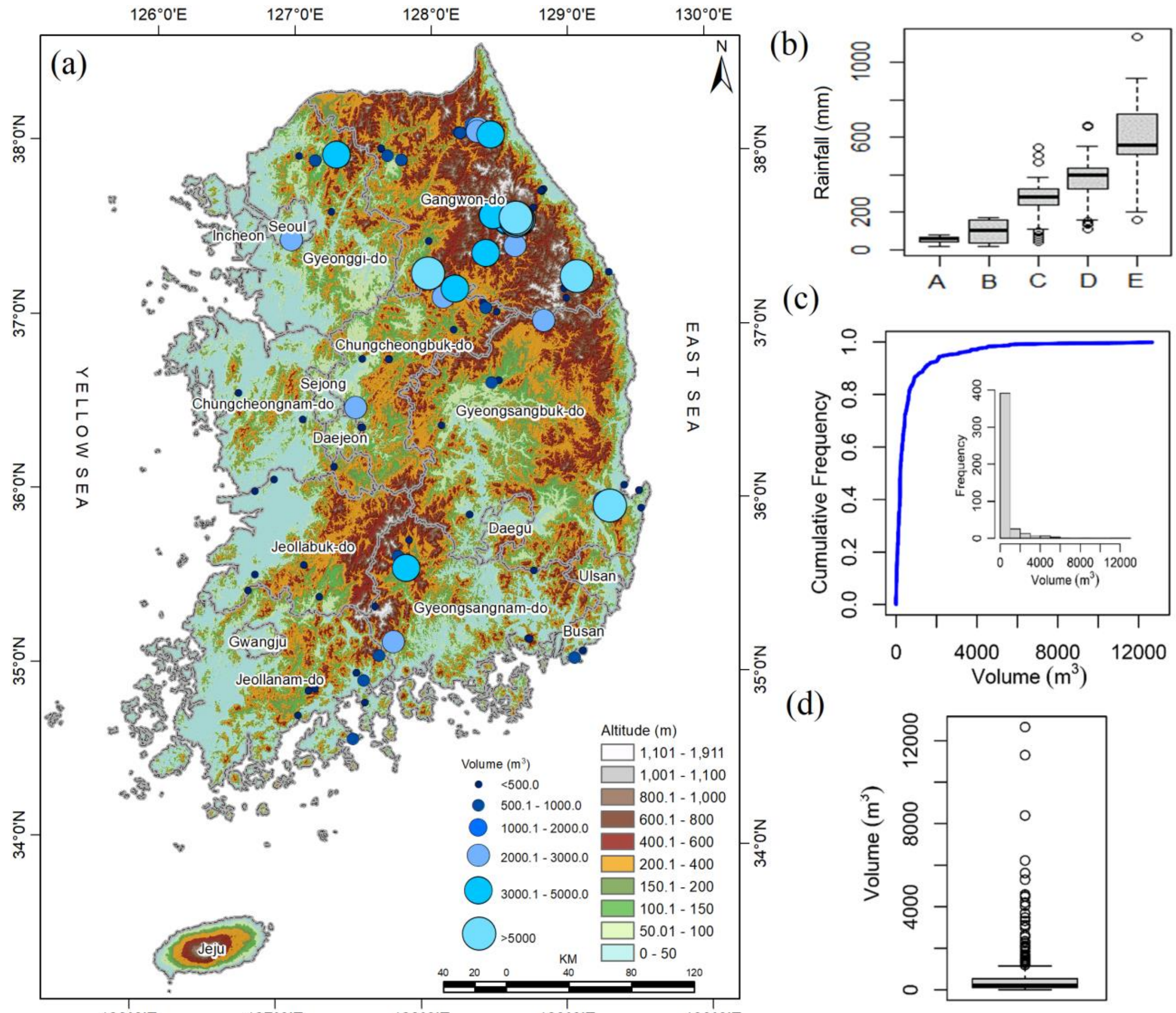

Figure 1: (a) Spatial distribution of landslides in South Korea, (b) Temporal variation of rainfall, i.e., A: Maximum hourly rainfall, B: Four weeks rainfall, C: Three hours rainfall, D: Three days rainfall and E: Two weeks rainfall, (c) Cumulative frequency distribution of the volume of landslides, and (d) Box plot of the volume of landslides. (The elevation data presented in Fig. 1a is sourced SRTM DEM, downloaded from https://earthexplorer.usgs.gov/).

2.2 Data

The landslide inventory dataset contains 455 landslide record information from 2011 to 2012, collected from different locations in South Korea through field survey, and the vegetation and forest fire features were obtained from Korean Forest Services database. The combined dataset tabulates information on landslide geometry, such as runout length, width, depth, and volume of the affected area, along with geomorphological composition, vegetation, and antecedent rainfall prior to landslide events. The details regarding landslide predisposing and triggering factors are summarized in Table 1.

The majority of landslides in this region were shallow, translational slope failures (Kim et al., 2001). The occurred landslides had a volume varying between 1.5m^3 to 12,663m^3 and predominantly occurred in the northeastern and southeastern region (Figs.1a,c-d). The landslides that occurred exhibited a hollowed morphology and a rightward skew in the distribution of their volumes with 2570.7m^3 as 95th quantile, with the largest volume 12,663m^3, and the aggregate mass of landslide due to rainfall was 276,986.62m^3. The estimation of the volume of removed material by landslides is important as it helps to assess risks the estimated damage can cause down at the toe of the failed slope, such as blocking transportation network, burying crops or farmland, the damage-built environment near landslide risks area, and post-disaster recovery planning (Evans et al., 2006; Rotaru et al., 2007; Intrieri et al., 2019).

Table 1: Landslide influencing and triggering factors.

Group	Features	Feature Relevance	References
Vegetation	Fire history	The burning of the vegetation intensifies the mass movement of soil near the uncovered burned stem of trees and free movement on uncovered soil due to post-fire rainfall and storms. The sliding may also be due to loss of vegetation , altered soil property and structure. These lead to soil degradation and higher infiltration, which increase the pore pressure, and change in hydrology by concentrating water flow in places that exacerbate landslides.	Highland and Bobrowsky, 2008; Stoof et al., 2012; Hyde et al., 2016; Culler et al., 2021
	Age of tree	Mature forests have more resistance to shallow landslides due to highly developed roots, which improve soil cohesion and leaves that prevent direct contact of raindrops with the soil surface.	Sato et al., 2023; Lann et al., 2024
	Forest density	The presence of forest reduces the likelihood of landslides about three times compared to grassland. Grassland has been revealed to be three times more vulnerable to shallow landslides than broadleaf, coniferous, and secondary forests.	Greenwood et al., 2004; Turner et al., 2010; Scheidl et al., 2020; Asada and Minagawa, 2023; Lann et al., 2024
	Timber diameter (m)	Tree spacing and size were used to investigate the effect of root and tree in shallow landslide control. High root density generally enhances slope stability, and specific tree placement and root sizes between 5 to 20 mm effectively prevent landslides.	Wang et al., 2016; Cohen and Schwarz, 2017
Geomorphology	Drainage	The drainage significantly affects slope stability and promotes efficient control of rainfall's influence on groundwater fluctuation. The presence of drainage increases the threshold of landslides due to rainfall.	Korup et al., 2007; Sun et al., 2010; Yan et al., 2019; Wei et al., 2019
	Slope angle (°)	The steeper slopes have a lower presence of landslides due to the low transportable materials. Slopes between 20-40 degrees are most vulnerable to greater landslides as rainfall intensity and duration increase. Generally, the average angle of the terrain at the landslide location provides valuable insight into the region's overall	Donnarumma et al., 2013; Duc, 2013; Qiu et al., 2016

Group	Features	Feature Relevance	References
		steepness and geomorphic characteristics, which are crucial factors for landslide susceptibility and risk modeling.	
	Slope aspect	The effect of rainfall on slope differs by slope angle and slope aspect, which leads to unevenly distributed landslides.	Panday and Dong, 2021; Cellek, 2021
	Slope length (m)	The volume increases as the slope length increases. A complex interplay exists between rainfall, length of slope and slope angle in the occurrence of landslides.	Turner et al., 2010
	Soil depth (m)	Soil properties, depth, and texture have significant differences in infiltration rates, which have different influences on the occurrence of landslides.	Kitutu et al., 2009; McKenna et al., 2012
	Soil type	Soil types, namely, Sandy loam, silt loam and loam, with their coefficient of permeability 1.7, 1.65 and 1.5, respectively, retain water differently, leading to different saturation times. The soil with higher permeability tends to drain water more efficiently, making it less prone to saturation. In contrast, the soil with lower permeability, the pore pressure may rapidly increase leading to shallow landslide initiation during intense rainfall events.	Chen et al., 2015; Liu et al., 2021a
Location	Altitude	Regional variability of elevation and mountain steepness affect the quantity of rainfall and associated landslides.	Um et al., 2010; Hyun et al, 2010; Yoon and Bae, 2013; Park, 2015
	Maximum hourly rainfall	The rainfall infiltrates the slope and increases pore water pressure, which reduces soil shear strength and leads to soil saturation, that causes surface failure.	Wieczorek, 1987; Dai and Lee, 2001; Smith et al., 2023
Rainfall	Continuous rainfall	Sudden intense rainfall concentrated in short periods is responsible for shallow landslides and debris flow.	Zhang et al., 2019
	Three hours rainfall		
	Three days rainfall	The antecedent rainfalls increase moisture in the soil and weaken soil cohesion.	Bernardie et al., 2014; Chen et al., 2015; Gariano et al., 2017; Zhang et al., 2019; Ran et al., 2022
	Two weeks rainfall		
	Four weeks rainfall		

Location parameters such as altitude, latitude and longitude are essential elements that determine the microclimate of a given region, influencing rainfall patterns (Hyun et al., 2010; Yoon and Bae, 2013; Park, 2015). The northeastern region is characterized by high-elevation terrain, such as the Taebaek and Sobaek ranges, which dry air and lead to orographic precipitation (Yun et al., 2009). The windward mountain versants receive a substantial amount of rainfall, which can increase the likelihood of landslides (Jin et al., 2022). This variation of rainfall with respect to the direction highlights the importance of including slope aspect variables in landslide studies (Kunz and Kottmeier, 2006). Figure 2(a) depicts the relationship between the slope aspect and the volume of landslides and slope aspect, altitude and fire history and shows that larger volumes were localized in regions that faced forest fire and altitudes between 500 and 1000m. Additionally, the topographical features such as slope length and slope angle affect the size of the landslide (Panday and Dong, 2021), slope failure due to over-saturation from groundwater and rainfall infiltration that destabilize the slope (Kafle et al., 2022). Furthermore, slope 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 soil composition properties, including soil permeability indices that affect water infiltration and saturation level (Chen et al., 2015). In the study 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). Moreover, to reduce infiltration, the drainage network channels rainwater, drains the soil, and reduces saturation, which minimizes the likelihood of landslide occurrence due to groundwater discharge and surface runoff (Hovius et al., 1997; Wei et al., 2019). Furthermore, the vegetation protects the topsoil from the direct impact of raindrops hitting the ground, which causes erosion due to the force of gravity and reduces infiltration (Omwega, 1989; Keefer, 2000). The absence of vegetation allows rainwater to seep away fine topsoil, causing shallow landslides (Gonzalez-Ollauri and Mickovski, 2017). On the contrary, vegetation improves soil cohesion and prevents potential shallow landslides due to soil-root interaction (Gong et al., 2021; Phillips et al., 2021). The density of vegetation (forest) and leafage type (broad, pines or mixture) directly affects the quantity of raindrops intercepted and prevented from directly hitting the soil, which emphasizes the contributions of vegetation in the landslides mitigation. Further, the occurrence of forest fires can contribute to the occurrence of landslides due to the burning of vegetation covering the area, changing soil properties and increasing soil pH (Lee et al., 2013).

The rainfall, a triggering factor of landslides, is the immediate cause of slope instability and failure due to infiltration that leads to saturation resulting from increased pore water pressure that reduces soil shear strength (Yune et al., 2010; Khan et al., 2012; Kim et al., 2021; Lee et al., 2021). The antecedent rainfall increases the moisture in the soil, which accelerates the soil saturation; the cumulative effect is essential to understand the saturation levels (Ran et al., 2022). In this study, rainfall variables are grouped based on time, namely, continuous rainfall, which is the accumulative value of rainfall on the day of a landslide from rainfall start hour to the landslide event, maximum hourly rainfall, rainfall during the fixed period such as three hours, one day, three days, two weeks etc. (Fig. 1b). The histograms for rainfall considered in this study are depicted in Figure 2(b-g). The descriptive statistics for all continuous variables are illustrated in Table 2.

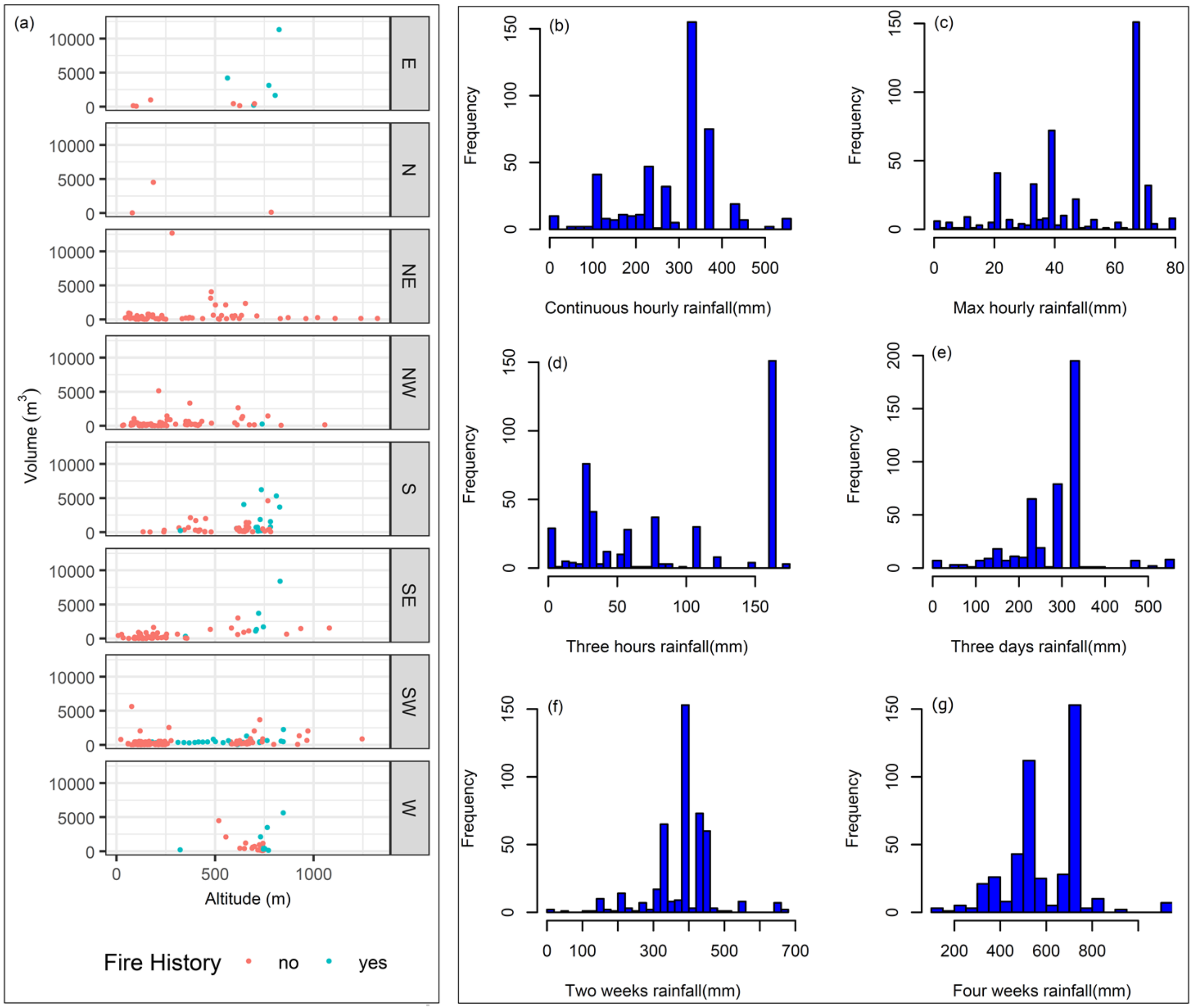

Figure 2: (a) The scatter plot showing the variation of landslide volumes with respect to slope aspect, fire history and altitude, and (b-g) Histograms of rainfall distribution.

Table 2: Summary statistics for continuous variables.

Variables	Units	N	Min	Mean	Median	Max	Std dev
Max Hourly rain	mm	455	0	48	48	78	20
Continuous rainfall	mm	455	0	285	327	550	106
Three hours rainfall	mm	455	0	88	80	171	60
Twelve Hours rainfall	mm	455	0	150	99	447	95
One day rainfall	mm	455	0	202	162	538	112
Three days rain	mm	455	0	280	284	550	86
Seven days rain	mm	455	0.5	323	330	634	88
Two weeks rain	mm	455	0.5	385	400	663	90
Three weeks rain	mm	455	86	504	533	914	115

Variables	Units	N	Min	Mean	Median	Max	Std dev
Four weeks rain	mm	455	108	587	561	1135	160
Soil depth	m	455	0.2	0.6	0.75	0.75	0.19
Soil type	-	455	1.5	1.6	1.5	1.7	0.087
Timber diameter	m	455	0.15	0.27	0.23	0.35	0.086
Age of tree	Years	455	10	34	35	60	14
Slope length	m	455	1.8	21	13	180	23
Slope angle	Degree (°)	455	10	34	34	65	7.9
Altitude	m	455	9	391	272	1324	273

3. Methods

In this paper, we consider nine data-driven models, namely OLS, RF, SVM, EGB, GLM, DT, DNN, KNN and 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 3. The steps for construction of these models can be briefly summarized as follows: a) the dataset for landslide inventories is cleaned and combined with rainfall dataset, b) the collinearity analysis is performed 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 mean absolute error (MAE), coefficient of determination (R^2), root mean square error (RMSE), symmetric mean absolute percentage error (SMAPE) and mean absolute percentage error (MAPE) for the comparison of actual and predicted volume by each model, f) variable importance is calculated for the optimal 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.

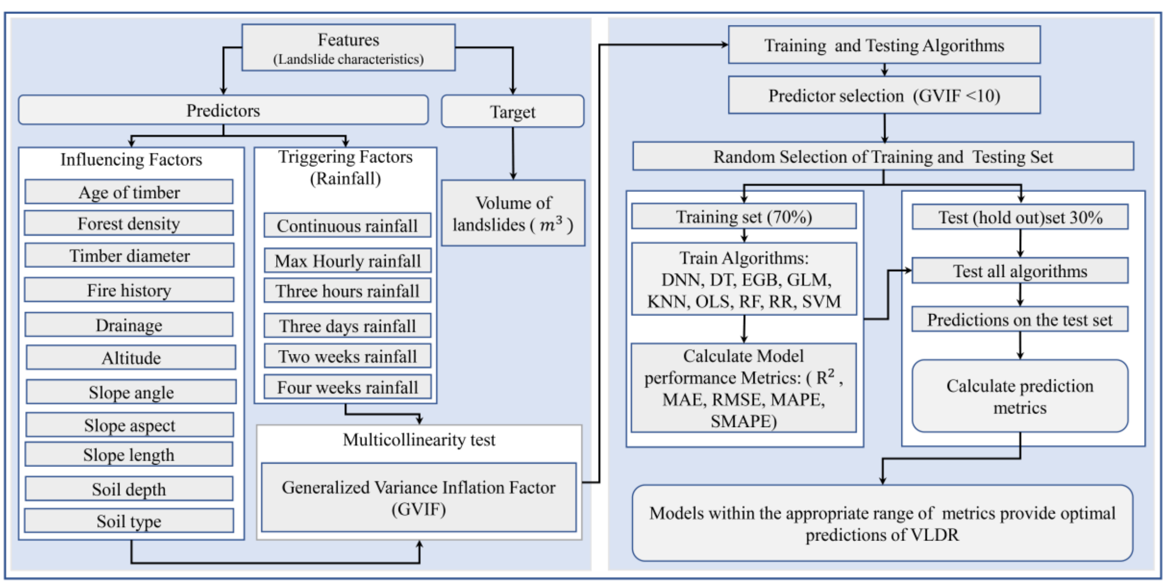

Figure 3: Workflow for the prediction of the volume of landslides due to rainfall.

3.1 Model Construction

In the present investigation, we aimed to predict landslide volume using models that minimize error with interpretability and scalability. Since one model can not have all properties simultaneously, we selected some widely used models due to their inherent interpretability and scalability properties. The OLS, GLM, and DT were widely used for their high interpretability, which helps to understand the influence of individual features on predictions (Gelman and Hill, 2007; Breiman, 2017). On the other hand, the EGB, RF, SVM, RR, and KNN were used due to their robust performance in capturing complex patterns in data, which is essential for accurate predictions of landslide volumes (Liaw and Wiener, 2002; Hastie, 2009; Chen et al., 2022). Additionally, considering that the model will be used on a regional scale, which will require big data, the EGB, RF, and DNN are designed to efficiently handle large datasets, making them suitable for the regional scale analysis. These last models can be scaled to incorporate more data from different geographical areas without significant adjustments, enhancing their applicability in future research (Krizhevsky et al., 2012). Accordingly, nine data-driven methods were selected and tested on a Korean dataset to predict VLDR.

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 (Kotsakis, 2023) under assumptions of no correlation in independent variables and 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 ensemble learning, which constructs many decision trees during the 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 datasets (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 DT regression is a predictive modeling technique in the form of a flowchart-like tree structure that includes all possible results, output, predictor costs, and utility. The DT simplifies the decision-making due to its algorithm that mimics 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 the 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 squares, which serves to respond to cases where 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 is fitted using maximum likelihood estimation (Dobson and Barnett, 2018). The DNN is 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., 2015). Here, the backpropagation algorithm was used to predict the estimated outcome. The advantage of DNN is that it can discover the complex structures in the data using a back propagation algorithm capable of changing 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). Subsequently, SVM has been applied in many different landslide studies (Pham et al., 2018; Miao et al., 2018). SVM methods identify the optimal hyperplane in multidimensional 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., 2022). 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 ML models were used. The firsts are known for high clarity and explainability, 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 et al., 2023).

3.2 Feature Selection and Data Splitting

The variable selection procedure was based on previous literature and applied in the model using generalized variance inflation factor (GVIF) (O'Brien, 2007) to eliminate collinear variables. The variable with GVIF<10 was considered non-colinear and used in the model. Figure 4 depicts retained features and corresponding GVIF values. The retained features have GVIF less than 10 (O'brien, 2007). Accordingly, all depicted variables were considered for the model training. Further, to train the model, the datasets were split randomly, with 70% of the data for the training set and 30% for testing (Nguyen et al., 2021). The 10-fold cross-validation was performed to obtain an optimal model. 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 using driven methods that accept numerical features only, the test and training set was one-hot-encoded to create a feature matrix (Seger, 2018).

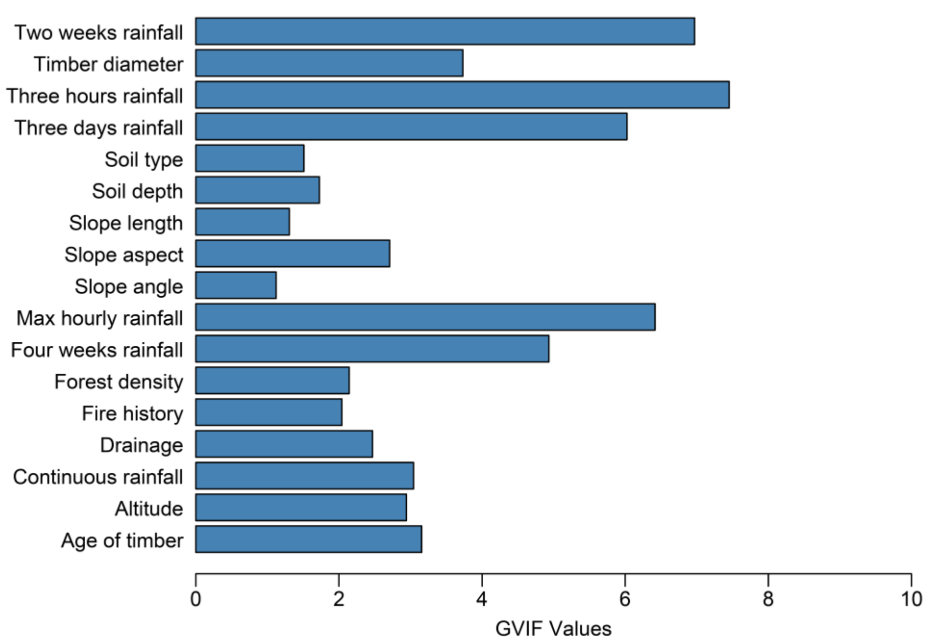

Figure 4: Generalized Variance Inflation Factor (GVIF) bar plot for features.

3.3 Model Evaluation Metrics

The model performance evaluation is a process of quantifying the difference between the observed value not used in the modeling process and the predicted value by the model. Different metrics are applied depending on the type of task, whether it is a classification or a regression problem. Subsequently, the widely used evaluation metrics for regression models, namely, R^2, MAE, RMSE, MAPE and SMAPE, were utilized to evaluate the model performances. The metric formulae and evaluation criteria are summarized in Table 3.

Table 3: Model evaluation metrics.

Metrics	Evaluation	References
$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	• Measures the square root of the average squared differences between predicted and actual values. • Lower values indicate better model performance.	Hyndman and Koehler, 2006
$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	• The average of the absolute differences between predicted and actual values. • Lower values indicate better model performance.	Willmott and Matsuura, 2005
$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert$	• Measures the accuracy of a model as a percentage, which can be more interpretable. • Lower values indicate better model performance.	Armstrong, 2001
$SMAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{\lvert y_i - \hat{y}_i\rvert}{\lvert y_i\rvert - \lvert\hat{y}_i\rvert}$	• Unlike MAPE, which can be skewed by very small actual values, SMAPE accounts for both the actual and predicted values, making it symmetric. • SMAPE is expressed as a percentage • Mitigates the impact of small actual values on the error metric, providing a more balanced assessment. • Lower values indicate better model performance.	Hyndman and Koehler, 2006
$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	• Represents the proportion of variance in the dependent variable that can be explained by the independent variables. • Values closer to 1 indicate a better fit	Darlington, 1990; Chicco et al., 2021

***y_i and $\hat{y}_i$ representing the actual and predicted value and, $\bar{y}$ and n standing for the mean of actual value and number of observations in the dataset, respectively.**

4. Results

This section details how all analyses and model development were performed in R using various libraries. The DNN regression model was constructed using dnn() function from the cito library (Amesoeder et al., 2023), with two hidden layers of (50, 50) nodes. The model was trained on 1500L epochs, learning rate (lr = 0.01), and loss = "mae". The DT regression model was constructed with tree() function from the tree library, with the recursive-partition method. The RR model was constructed using glmnet() from the glmnet package (Friedman et al., 2010), with ridge penalty (alpha=0). The optimal lambda was obtained by performing 10-fold cross-validation. The EGB model was built using xgboost() function in xgboost package (Chen et al., 2022). The optimal model was obtained at 524th boosting iteration with max depth =5 and other parameters set to default. The GLM regression model was constructed using glm() function (R core Team, 2022) with family Gaussian and log link to constrain the model of predicting positive outcomes. The KNN regression was constructed using knnreg() function from the caret package (Kuhn, 2022), with number of neighbors, *k*=17. The OLS model was constructed lm() from the stats package (R core Team, 2022). The RF model was run using randomForest() from the randomforest package (Liaw and Wiener, 2002) with default parameters and the optimal model was reached at 256th iteration. The SVM regression model with linear kernel was built using e1071 package (Meyer et al., 2021) and other parameters set to default.

The predictive performance of all tested models on the holdout dataset is depicted by the scatterplot (Fig. 5) of actual volume as recorded in the test set and predicted outcome values of each model. 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=0.2744, that is, 27% of variances in the model were explained by predictors. The second least performing was the RR with R^2= 0.3034, which is 3.6% improvement compared to OLS. Among all models, three out of nine, namely, OLS, SVM, and RR, performed below 50%; however, these models predicted well small values of volume (below 2000m^3). The MAE of these three models was higher than the remaining six models, namely DNN, DT, GLM, KNN, RF, and EGB. Among these lasts, the most performing was EGB with R^2= 0.88 of variance explained by predictors and MAE=146.6 m^3. The evaluation metrics for the training and tested models are summarized in Table 4. Considering the R^2, the three models, namely EGB, RF, and DNN, had a value of R^2 above 80% on the holdout set.

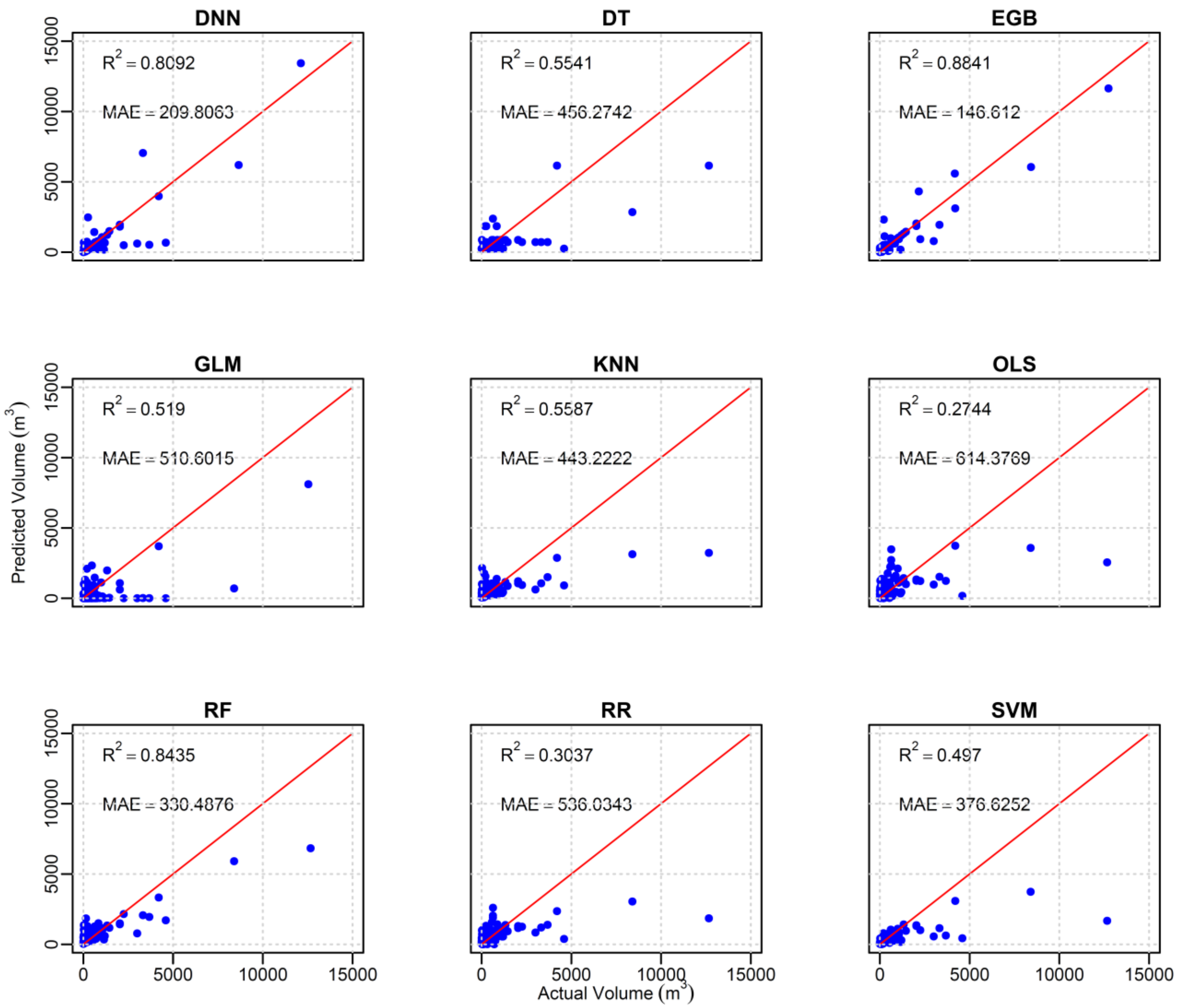

Figure 5: Scatterplot of actual and predicted values for the nine tested models.

Regarding the prediction on the training set, the GLM had an R^2 of 83%. Nevertheless, the prediction on the holdout set was 51.9%; this large variation in variance explained by predictors indicates that the GLM model did not catch all non-linear patterns in the holdout set. Notably, the prediction difference in R^2 on both training and test for the random forest exhibited a very small difference compared to EGB and DNN, that is, 1.75% compared to 12.17% and 7.72% for DNN and EGB, respectively. Despite the stable prediction of RF, the performance in terms of SMAPE, the DNN was the second lowest symmetric mean absolute percentage error, 43.83m^3 and 39.79 m^3 on training and test sets, respectively. According to Chicco et al. (2021), the R^2 is more informative in regression modeling; thus, RF had better predictions than the DNN.

Table 4: Summary of prediction metrics for tested models on the training and test set.

Metrics		Models								
		DNN	DT	EGB	GLM	KNN	OLS	RF	RR	SVM
R^2	Train	0.9309	0.4514	0.9613	0.8380	0.3470	0.3775	0.8610	0.3382	0.5510
	Test	0.8092	0.5822	0.8841	0.5190	0.5587	0.2744	0.8435	0.3037	0.4970
MAE	Train	132.7429	407.0814	75.1250	308.9700	410.2945	502.0053	236.9516	470.1633	276.2000
	Test	209.8063	435.5836	146.6120	510.6015	443.2222	614.3769	330.4876	536.0343	376.6252
RMSE	Train	348.6190	940.4850	113.4940	570.0070	1027.3730	1001.7620	574.9720	1042.9110	916.5471
	Test	646.5438	1047.4880	501.8960	1055.9190	1115.5270	1234.1220	737.0857	1237.9420	1176.9410
MAPE	Train	0.5240	0.7930	0.1540	76.3530	0.6280	5.2310	0.3810	1.5330	1.1588
	Test	0.5623	0.8892	0.3132	1819.2220	0.6623	4.1277	0.4939	5.8428	1.0421
SMAPE	Train	43.8375	79.8680	13.1780	150.4262	67.4715	103.0555	52.3359	93.4002	67.3221
	Test	39.7998	81.4539	22.7237	152.4991	73.6498	106.9756	63.7582	93.9244	76.9794

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 slope length was the most contributing predictor in the performance of the EGB model, followed by maximum hourly rainfall and slope aspect. The altitude, three hours rainfall, slope angle and age of timber contributed moderately to the prediction of the outcome volumes with gain above 0.01 and less than 0.2. The antecedent rainfall from three days and above and continuous rainfall had a minor contribution, with a gain of less than 0.01 for each. The presence of rainwater drainage channels had a moderate contribution, with a gain close to 0.01. On the other hand, the contribution of soil depth and forest density in the models was insignificant and far below 0.01. Though Figure 2(a) depicted the association between larger volumes and fire history, the variable importance indicates that the relation was not significant. Even though some variables had minor contributions, depending on the case, the contribution of those variables may also increase depending on other regional settings. Therefore, all variables with GVIF below 10 were kept in the model. Figure 6 illustrates the variables importance for the EGB model. The vertical red line splits lanslides prediction features into two groups, the first containing features that contributed a gain above 0.01 and others with minor contributions.

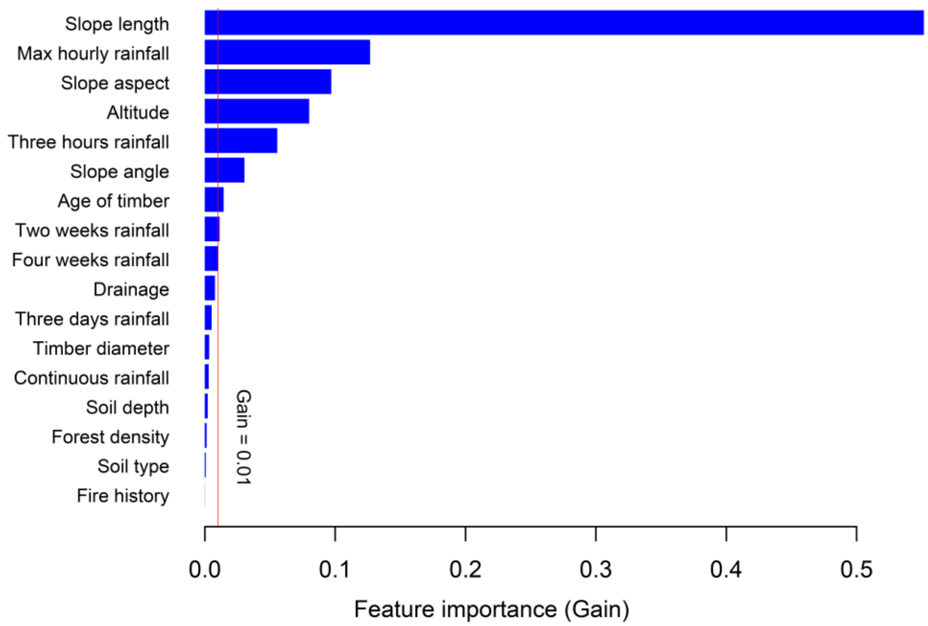

Figure 6: Variable importance for the EGB model.

The variable importance plot depicts the overall contribution of a given feature; 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, and Dunn's test were conducted. The Shapiro-Wilk's test (Dudley, 2023) results revealed that the distribution of volume was non-normal ($W = 0.40642$, $p\text{-value} < 0.001$). 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 7 illustrates that the slope length exhibited a higher value (dcor=0.56) followed by continuous rainfall altitude and three hours rainfall and kept decreasing up to timber diameter with a distance correlation of 0.08. Overall, the distance correlation between the volume of landslides shows a moderate strength of association between continuous predictors.

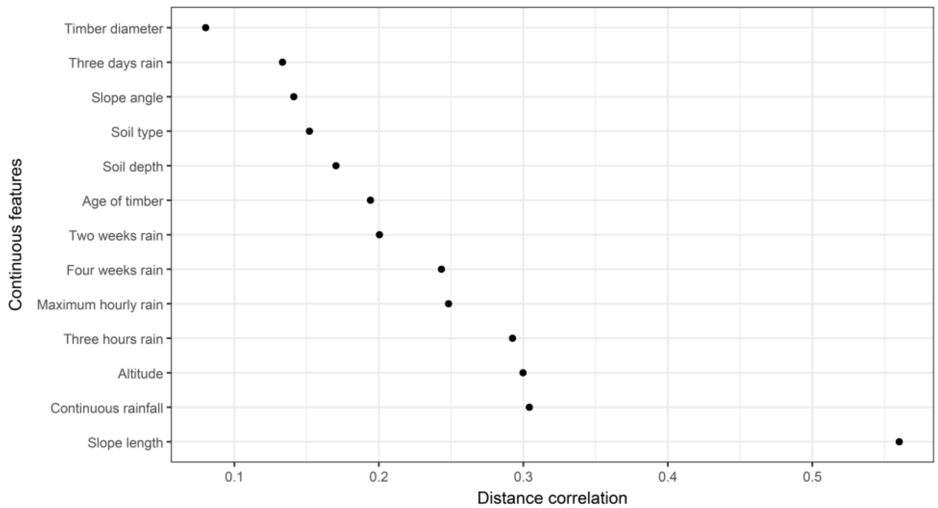

Figure 7: Distance correlation plot for the volume and continuous features.

Furthermore, to test for categorical features, Kruskal-Wallis test (McKight and Najab, 2010) was used to check whether the volume of the 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 different means of landslides' volume (with p-value <0.05). Figure 8 depicts that the southwest and southeast aspects had a higher frequency of landslides.

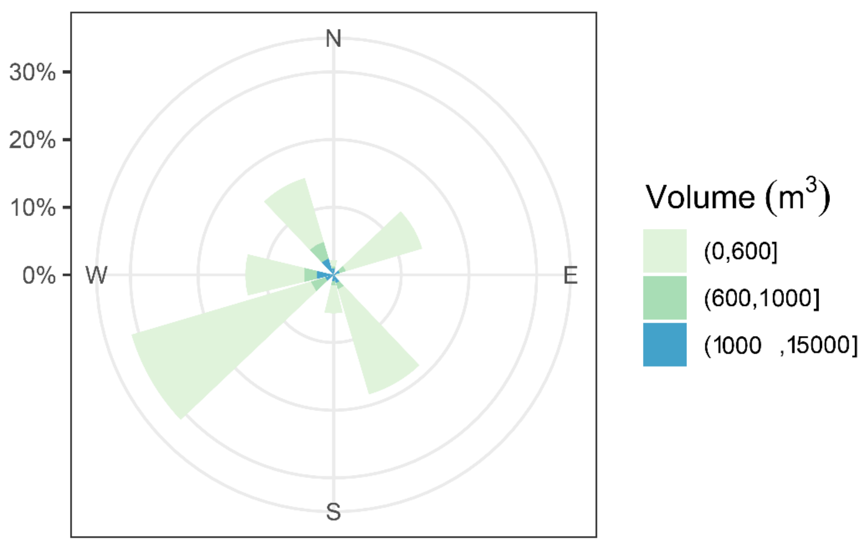

Figure 8: The distribution of the volume of landslides due to rainfall with respect to the slope aspect.

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

5. Discussion

Numerical models have traditionally been employed due to their foundation in physical principles such as slope stability and hydrological dynamics (Glade et al., 2005). These models are valuable for understanding the underlying mechanisms of landslide processes but often face limitations when applied to regions with complex or heterogeneous terrain, as they require detailed, high-quality input data that may not always be available (Caine, 1980). In the same way, statistical models, which use historical rainfall and landslide data to establish correlations, can offer useful predictions of VLDR in regions with extensive historical records (Chung and Fabbri, 2003). However, these models may struggle to account for local variations in topography or rapidly changing weather patterns, limiting their general applicability. Additionally, ML techniques have shown significant promise in improving predictive accuracy at the regional level due to the capability of processing large, diverse datasets and capturing complex, non-linear relationships that traditional models might fail to capture (Pourghasemi and Rahmati, 2018). Further, ML models can adapt to regional variations and continuously improve as new data is introduced, offering a more flexible and dynamic approach to predict VLDR on a regional scale (Liu et al., 2021b). Subsequently, the aim of this study was to construct a data-driven algorithm that accurately predicts the VLDR. 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, DT and KNN, other machine learning models (DNN, DT, RF, and EGB) exhibited high prediction capability with R^2 above 50% (Fig. 5). The DNN, EGB, and RF models achieved $R^2>0.8$ on both training and test set with accuracy reduced R^2 by 1.75, 7.72, and 12.17% for RF, EGB and DNN respectively, on the holdout set, indicating that the model could yield reliable volume estimates in adjacent areas with similar geological and environmental conditions. 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 compared to random forest; however, the model did not perform well on moderate volume values, resulting in reduced R^2. The EGB model tested on South Korean landslide inventory coupled with rainfall data at the time of landslide events and antecedent rainfall within one month of the event exhibited more accurate predictions compared to other constructed algorithms. The difference in performance may be due to the internal structure of each algorithm; the RF builds multiple decision trees and averages predictions to improve accuracy (Breiman, 2001), while the EGB builds sequential trees in a recursive order where the new built tree improves error occurred while building the previous decision tree and optimizes the loss function through a gradient descent (Chen et al., 2022).

The slope aspect played an important role in the prediction of the volume, and the landslide mostly occurred in locations oriented toward south-southwest and southeast. That may be due to the direction taken by typhoons, which hit the southwest versants of mountains upon landfall on the Korean peninsula toward the North East Pacific (Lee et al., 2013; Ha, 2022). The findings of this research are congruent with those of 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 a moderate rainwater drainage channel plays an important role in the prevention of landslides due to its stabilizing effect. The landslide location and pattern follow the rainfall climate scenario, which highlighted a higher intensity of rainfall in the northeastern region of South Korea (Lee, 2016). In addition, the findings of this study are congruent with Zhang et al. (2019) observations that highlighted the low influence of soil type in landslide modeling and the maximum rainfall and cumulative three hours of rainfall were the most contributing rainfall, which indicated that these shallow landslides may have been triggered by sudden rainfall concentrated in few hours before the occurrence of the event. The occurrence of landslides triggered by rainfall is a complex phenomenon that involves many interrelated environmental settings, 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 settings 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 in a few literatures related to the prediction of the volume of landslides using data-driven techniques. This model can serve as an effective tool for policy-makers to incorporate landslide volume risks into policies aimed at protecting infrastructure and residents dwelling in landslides high risks zones.

To understand the applicability of the developed models, the trained model was tested using unknown data (test data), with volume predictions generated solely based on the predictor variables; actual volume values were utilized only for evaluating model prediction accuracy. The outcome exhibited that the difference in R^2 on the training and holdout set of 7.72% for the optimal model (i.e., EGB) highlights that the model can be applied to another region of a similar setting. It was noted that without proper model calibration with the independent data set, it's difficult to determine whether these discrepancies in performance are due to model limitations or data differences in different regions (Huang et al., 2020). Therefore, future research will focus on developing an independent database containing recent landslide geometry data from various regions of the Korean Peninsula to enhance model accuracy, along with calibrating region-specific parameters to ensure the model's transferability to other regions.

The major limitation of this study is that the analysis is solely focused on shallow-seated landslides, specifically translational slope failures with volumes below 13,000m³. Thus, the analysis may not fully capture the variability in landslide characteristics across different geomorphological and geological contexts. Deep-seated landslides, for instance, often exhibit distinct failure mechanisms, material compositions, and depositional patterns that influence their volumetric characteristics, which were not considered in this investigation. Similarly, debris flows, known for their unique channelization and entrainment behaviors, were not included, potentially limiting the applicability of the optimized models to other landslide types. Further, this study was also performed using point-based landslide inventory data, which may not capture all variability of influencing factors and their exact state. The incorporation of high-resolution data from remote sensing and other sources may also improve the efficiency of the predictions. These limitations may impact the broader applicability of the proposed model; however, future studies will aim to address this by conducting separate analyses for deep-seated landslides and debris flows, allowing for a more comprehensive understanding of landslide volume predictions across diverse landslide types and geomorphological settings.

6. Conclusions

In this paper, the aim was to construct a data-driven model that predicts 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 the tested models, the EGB model produced the most accurate prediction. This is proven by the evaluation of the difference between actual and predicted values, such as R^2= 88.41% and MAE=146.6120m^3 on the holdout set. 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 landslides due to its high predictive performance. The model can be customized in 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 after re-training on the regional historical data to include regional variability. Therefore, this model can be a good 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.

Acknowledgments

This research was supported by the Korean government (MSIT) (2021R1C1C2003316) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by Ministry of Education (2021R1A6A1A03044326). The authors highly appreciate both anonymous reviewers and editor for their constructive suggestions that helped us improve the preprint version.

Code availability

The codes used forVDLR prediction are available from the corresponding author upon reasonable request.

Data availability

All data used in this study are available from the corresponding author upon request.

Author contributions

TJ: conceptualization, formal analysis, investigation, methodology, software/code, data curation, visualization, validation, and writing (original draft preparation and review and editing). CYN: data curation, supevision, and writing (review and editing). GK: data curation, supevision, and writing (review and editing). SWL: data curation, supevision, and writing (review and editing). MDA: conceptualization, formal analysis, investigation, methodology, software, data curation, visualization, validation, and writing (original draft preparation and review and editing). SGY:

conceptualization, investigation, supervision, methodology, project administration, and writing (review and editing).

Competing interests

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

References

Alcantara, A. L., and Ahn, K. H.: Probability distribution and characterization of daily precipitation related to tropical cyclones over the Korean Peninsula, Water, 12(4), 1214, https://doi.org/10.3390/w12041214, 2020.

Alcántara-Ayala, I., and Sassa, K.: Landslide risk management: from hazard to disaster risk reduction, Landslides, 20(10), 2031-2037, https://doi.org/10.1007/s10346-023-02140-5, 2023.

Amesoeder, C., Hartig, F., and Pichler, M.: cito: An R package for training neural networks using torch, arXiv e-prints, arXiv-2303, https://doi.org/10.1111/ecog.07143, 2023.

Armstrong, J. S.: Combining forecasts (pp. 417-439), Springer US, https://doi.org/10.1007/978-0-306-47630-3_19, 2001.

Asada, H., and Minagawa, T.: Impact of vegetation differences on shallow landslides: a case study in Aso, Japan, Water, 15(18), 3193, https://doi.org/10.3390/w15183193, 2023.

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

Bonamutial, M., and Prasetyo, S. Y.: 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, https://doi.org/10.1109/ICIMTech59029.2023.10277860, 2023.

Borup, D., Christensen, B. J., Mühlbach, N. S., and Nielsen, M. S.: Targeting predictors in random forest regression, Int. J. Forecast., 39(2), 841-868, https://doi.org/10.1016/ j.ijforecast.2022.02.010, 2023.

Breiman, L.: Random forests, Machine Learning, 45, 5-32, https://doi.org/10.1023/ A:1010933404324, 2001.

Breiman, L.: Classification and regression trees, Routledge, https://doi.org /10.1201/9781315139470, 2017.

Caine, N.: The rainfall intensity-duration control of shallow landslides and debris flows, Geografiska annaler: series A, Phys. Geogr., 62(1-2), 23-27, https://doi.org/10.1080/04353676.1980.11879996, 1980.

Cellek, S.: The effect of aspect on landslide and its relationship with other parameters, In Landslides, IntechOpen., https://dx.doi.org/10.5772/intechopen.99389, 2021.

Chang, K. T., and Chiang, S. H.: An integrated model for predicting rainfall-induced landslides, Geomorphology, 105(3-4), 366-373, https://doi.org/10.1016/ j.geomorph.2008.10.012, 2009.

Chang, K. T., Chiang, S. H., and Lei, F.: Analysing the relationship between typhoon-triggered landslides and critical rainfall conditions, Earth Surf. Process. Landf.: J. British Geomor. Res. Group, 33(8), 1261-1271, https://doi.org/10.1002/esp.1611, 2008.

Chatra, A. S., Dodagoudar, G. R., and Maji, V. B.: Numerical modelling of rainfall effects on the stability of soil slopes, Int. J. Geotech. Eng., https://doi.org/10.1080/ 19386362.2017.1359912, 2019.

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., and Yuan, J.: xgboost: Extreme Gradient Boosting, R package version 1.6.0.1, https://CRAN.R-project.org/package=xgboost, last access: 25 January 2025, 2022.

Chen, C. W., Oguchi, T., Hayakawa, Y. S., Saito, H., and Chen, H.: Relationship between landslide size and rainfall conditions in Taiwan, Landslides, 14, 1235-1240, https://doi.org/10.1007/s10346-016-0790-7, 2017.

Chen, L., Guo, Z., Yin, K., Shrestha, D. P., and Jin, S.: The influence of land use and land cover change on landslide susceptibility: a case study in Zhushan Town, Xuan'en County (Hubei, China), Nat. Hazards Earth Syst. Sci., 19(10), 2207-2228, https://doi.org/10.5194/nhess-19-2207-2019, 2019.

Chen, X., Zhang, L., Zhang, L., Zhou, Y., Ye, G., and Guo, N.: Modelling rainfall-induced landslides from initiation of instability to post-failure, Comput. Geotech., 129, 103877, https://doi.org/10.1016/j.compgeo.2020.103877, 2021.

Chen, Z., Luo, R., Huang, Z., Tu, W., Chen, J., Li, W., Chen, S., Xiao, J. and Ai, Y.: Effects of different backfill soils on artificial soil quality for cut slope revegetation: Soil structure, soil erosion, moisture retention and soil C stock, Ecol. Eng., 83, 5-12, https://doi.org/10.1016/j.ecoleng.2015.05.048, 2015.

Cheung, R. W.: Landslide risk management in Hong Kong, Landslides, 18(10), 3457-3473, https://doi.org/10.1007/s10346-020-01587-0, 2021.

Chicco, D., Warrens, M. J., and Jurman, G.: The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, PeerJ Comput. Sci., 7, e623, https://doi.org/10.7717/peerj-cs.623, 2021.

Chowdhury, M. Z. I., Leung, A. A., Walker, R. L., Sikdar, K. C., O'Beirne, M., Quan, H., and Turin, T. C.: A comparison of machine learning algorithms and traditional regression-based statistical modeling for predicting hypertension incidence in a Canadian population, Sci. Rep., 13(1), 13, https://doi.org/10.1038/s41598-022-27264-x, 2023.

Chung, C. J. F., and Fabbri, A. G.: Validation of spatial prediction models for landslide hazard mapping, Nat. Hazards, 30, 451-472, https://doi.org/10.1023/ B:NHAZ.0000007172.62651.2b, 2003.

Cohen, D., and Schwarz, M.: Tree-root control of shallow landslides, Earth Surf. Dyn., 5(3), 451-477, https://doi.org/10.5194/esurf-5-451-2017 ,2017.

Culler, E. S., Livneh, B., Rajagopalan, B., and Tiampo, K. F.: A data-driven evaluation of post-fire landslide susceptibility, Nat. Hazards Earth Syst. Sci., 2021, 1-24, https://doi.org/10.5194/nhess-23-1631-2023, 2021.

Dahal, B. K., and Dahal, R. K.: Landslide hazard map: tool for optimization of low-cost mitigation, Geoenvironmental Disasters, 4, 1-9, https://doi.org/10.1186/s40677-017-0071-3, 2017.

Dai, F. C., and Lee, C. F.: Frequency–volume relation and prediction of rainfall-induced landslides, Eng. Geol., 59(3-4), 253-266, https://doi.org/10.1016/S0013-7952(00)00077-6, 2001.

Darlington, R. B.: Regression and linear models, Mcgraw-Hill, New York, USA, ISBN: 0070153728, 9780070153721, 1990.

Dinno, A.: Nonparametric pairwise multiple comparisons in independent groups using Dunn's test, Stata J., 15(1), 292-300, https://doi.org/10.1177 /1536867X1501500117, 2015.

Dobson, A. J., and Barnett, A. G.: An introduction to generalized linear models, CRC press, New York, USA, ISBN: 9781315182780, https://doi.org/10.1201/9781315182780, 2018.

Donnarumma, A., Revellino, P., Grelle, G., and Guadagno, F. M.: 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, Springer, Berlin , https://doi.org/10.1007/978-3-642-31325-7_56, 2013.

Duc, D. M.: Rainfall-triggered large landslides on 15 December 2005 in Van Canh district, Binh Dinh province, Vietnam, Landslides, 10(2), 219-230. https://doi.org/10.1007 /s10346-012-0362-4, 2013.

Dudley, R.: The Shapiro–Wilk test for normality, Available at https://math.mit.edu/~rmd/46512/shapiro.pdf , last access: 25 January 2025, 2023.

Evans, S., Mugnozza, G.S., Strom, A., Hermanns, R., Ischuk, A., Vinnichenko, S.: Landslides From Massive Rock Slope Failure And Associated Phenomena, In: Landslides from Massive Rock Slope Failure, NATO Science Series, vol 49, Springer, Dordrecht., https://doi.org/10.1007/978-1-4020-4037-5_1, 2006.

Friedman, J. H., Hastie, T., and Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw., 33, 1-22, Available at https://pmc.ncbi.nlm.nih.gov/articles/PMC2929880/, last access: 25 January 2025, 2010.

Gariano, S. L., Rianna, G., Petrucci, O., and Guzzetti, F.: Assessing future changes in the occurrence of rainfall-induced landslides at a regional scale, Sci. Total Environ., 596, 417-426, https://doi.org/10.1016/j.scitotenv.2017.03.103, 2017.

Gelman, A., and Hill, J.: Data analysis using regression and multilevel/hierarchical models, Cambridge University Press, New York, ISBN: 0-521-86706-1, 2007.

Glade, T., Anderson, M. G., and Crozier, M. J.: Landslide hazard and risk (Vol. 807), John Wiley & Sons, ISBN: 9780470012659, https://doi.org/10.1002/9780470012659, 2005.

Gong, Q., Wang, J., Zhou, P., and Guo, M.: A regional landslide stability analysis method under the combined impact of rainfall and vegetation roots in south China, Adv. Civ. Eng., 1-12, https://doi.org/10.1155/2021/5512281, 2021.

Gonzalez-Ollauri, A., and Mickovski, S. B.: Hydrological effect of vegetation against rainfall-induced landslides, J. Hydrol., 549, 374-387. https://doi.org/10.1016/ j.jhydrol.2017.04.014, 2017.

Greenwood, J. R., Norris, J. E., and Wint, J.: Assessing the contribution of vegetation to slope stability, Proc. Inst. Civil Eng. Geotech. Eng., 157(4), 199-207, https://doi.org/10.1680/geng.2004.157.4.199, 2004.

Gutierrez-Martin, A.: A GIS-physically-based emergency methodology for predicting rainfall-induced shallow landslide zonation, Geomorphology, 359, 107121, https://doi.org/10.1016/j.geomorph.2020.107121, 2020.

Guzzetti, F., Peruccacci, S., Rossi, M., and Stark, C. P.: The rainfall intensity–duration control of shallow landslides and debris flows: an update, Landslides, 5, 3-17, https://doi.org/10.1007/s10346-007-0112-1, 2008.

Ha, K. M.: Predicting typhoon tracks around Korea, Nat. Hazards, 113(2), 1385-1390, https://doi.org/10.1007/s11069-022-05335-6, 2022.

Hastie, T.: The elements of statistical learning: data mining, inference, and prediction, 2nd edition, Springer, New York, https://doi.org/10.1111/j.1541-0420.2010.01516.x, ISBN: 9780387848570, 2009.

Highland, L. and Bobrowsky, P.: The Landslide Handbook: A Guide to Understanding Landslides, USGS, Reston, VA, Circular 1325, Available at https://pubs.usgs.gov/circ/1325/, last access: 25 January 2025, 2008.

Holcombe, E. A., Beesley, M. E., Vardanega, P. J., and Sorbie, R.: Urbanisation and landslides: hazard drivers and better practices, In Proc. Inst. Civ. Eng. Civ. Eng. (Vol. 169(3), pp. 137-144), Thomas Telford Ltd, https://doi.org/10.1680/jcien.15.00044, 2016.

Hovius, N., Stark, C. P., and Allen, P. A.: Sediment flux from a mountain belt derived by landslide mapping, Geology, 25(3), 231-234, https://doi.org/10.1130/0091-7613(1997)025<0231:SFFAMB>2.3.CO;2, 1997.

Huang, J., Hales, T. C., Huang, R., Ju, N., Li, Q., and Huang, Y.: A hybrid machine-learning model to estimate potential debris-flow volumes, Geomorphology, 367, 107333, https://doi.org/10.1016/j.geomorph.2020.107333, 2020.

Hyde, K. D., Riley, K., and Stoof, C.: Uncertainties in predicting debris flow hazards following wildfire, Nat. Hazards, https://doi.org/10.1002/9781119028116.ch19, 2016.

Hyndman, R. J., and Koehler, A. B.: Another look at measures of forecast accuracy, Int. J. Forecast., 22(4), 679-688, https://doi.org/10.1016/ j.ijforecast.2006.03.001, 2006.

Hyun, Y. K., Kar, S. K., Ha, K. J., and Lee, J. H.: Diurnal and spatial variabilities of monsoonal CG lightning and precipitation and their association with the synoptic weather conditions over South Korea, Theor. Appl. Climatol., 102, 43-60, https://doi.org/10.1007/s00704-009-0235-5, 2010.

Intrieri, E., Carlà, T., and Gigli, G.: Forecasting the time of failure of landslides at slope-scale: A literature review, Earth-Sci. Rev., 193, 333-349, https://doi.org/10.1016/ j.earscirev.2019.03.019, 2019.

Jaboyedoff, M., Choffet, M., Derron, M. H., Horton, P., Loye, A., Longchamp, C., Mazotti, B., Michoud, C., and Pedrazzini, A.: Preliminary slope mass movement susceptibility mapping using DEM and LiDAR DEM, In Terrigenous mass movements: Detection, modelling, early warning and mitigation using geoinformation technology, 109-170, Springer, Berlin Heidelberg, https://doi.org/10.1007/978-3-642-25495-6_5, 2012.

Jin, H. G., Lee, H., and Baik, J. J.: Characteristics and possible mechanisms of diurnal variation of summertime precipitation in South Korea, Theor. Appl. Climatol., 148(1), 551-568, https://doi.org/10.1007/s00704-022-03965-1, 2022.

Ju, L. Y., Zhang, L. M., and Xiao, T.: Power laws for accurate determination of landslide volume based on high-resolution LiDAR data, Eng. Geol., 312, 106935, https://doi.org/10.1016/j.enggeo.2022.106935, 2023.

Jung, M. J., Jeong, Y. J., Shin, W. J., and Cheong, A. C. S.: Isotopic distribution of bioavailable Sr, Nd, and Pb in Chungcheongbuk-do Province, Korea, J. anal. sci. technol., 15(1), 46,https://doi.org/10.1186/s40543-024-00460-2, 2024.

Jung, Y., Shin, J. Y., Ahn, H., and Heo, J. H.: The spatial and temporal structure of extreme rainfall trends in South Korea, Water, 9(10), 809, https://doi.org/10.3390/w9100809, 2017.

Kafle, L., Xu, W. J., Zeng, S. Y., and Nagel, T.: 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, Eng. Geol., 297, 106508, https://doi.org/10.1016/j.enggeo.2021.106508, 2022.

Kang, M. W., Yibeltal, M., Kim, Y. H., Oh, S. J., Lee, J. C., Kwon, E. E., and Lee, S. S.: Enhancement of soil physical properties and soil water retention with biochar-based soil amendments, Sci. Total Environ., 836, 155746, https://doi.org/10.1016/ j.scitotenv.2022.155746, 2022.

Keefer, R. F.: Handbook of soils for landscape architects, Oxford University Press, ISBN: 0-19-51202-3, 2000.

Khan, M. A., Basharat, M., Riaz, M. T., Sarfraz, Y., Farooq, M., Khan, A. Y., Pham, Q. B., Ahmed, K. S., and Shahzad, A.: An integrated geotechnical and geophysical investigation of a catastrophic landslide in the Northeast Himalayas of Pakistan, Geol. J., 56(9), 4760-4778, https://doi.org/10.1002/gj.4209, 2021.

Khan, Y. A., Lateh, H., Baten, M. A., and Kamil, A. A.: Critical antecedent rainfall conditions for shallow landslides in Chittagong City of Bangladesh, Environmental Earth Sciences, 67, 97-106. https://doi.org/10.1007/s12665-011-1483-0, 2012.

Kim, D. E., Seong, Y. B., Weber, J., and Yu, B. Y.: Unsteady migration of Taebaek Mountain drainage divide, Cenozoic extensional basin margin, Korean Peninsula, Geomorphology, 352, 107012, https://doi.org/10.1016/j.geomorph.2019.107012, 2020.

Kim, H. G., and Park, C. Y.: Landslide susceptibility analysis of photovoltaic power stations in Gangwon-do, Republic of Korea, Geomat. Nat. Hazards Risk., 12(1), 2328-2351, https://doi.org/10.1080/19475705.2021.1950219, 2021.

Kim, J., Lee, K., Jeong, S., and Kim, G.: GIS-based prediction method of landslide susceptibility using a rainfall infiltration-groundwater flow model, Eng. Geol., 182, 63-78, https://doi.org/10.1016/j.enggeo.2014.09.001, 2014.

Kim, M. S., Onda, Y., Kim, J. K., and Kim, S. W.: Effect of topography and soil parameterisation representing soil thicknesses on shallow landslide modelling, Quat. Int, 384, 91-106, https://doi.org/10.1016/j.quaint.2015.03.057, 2015.

Kim, S. W., Chun, K. W., Kim, M., Catani, F., Choi, B., and Seo, J. I.: Effect of antecedent rainfall conditions and their variations on shallow landslide-triggering rainfall thresholds in South Korea, Landslides, 18, 569-582, https://doi.org/10.1007/s10346-020-01505-4, 2021.

Kitutu, M. G., Muwanga, A., Poesen, J., and Deckers, J. A.: Influence of soil properties on landslide occurrences in Bududa district, Eastern Uganda, Afr. J. Agric. Res., 4(7), 611-620, Available at https://lirias.kuleuven.be/retrieve/78489, last access: 25 January 2025, 2009.

Korup, O.: Geomorphometric characteristics of New Zealand landslide dams, Eng. Geol., 73(1-2), 13-35. https://doi.org/10.1016/j.enggeo.2003.11.003, 2004.

Korup, O., Clague, J. J., Hermanns, R. L., Hewitt, K., Strom, A. L., and Weidinger, J. T.: Giant landslides, topography, and erosion, Earth Planet. Sci. Lett., 261(3-4), 578-589, https://doi.org/10.1016/j.epsl.2007.07.025, 2007.

Kotsakis, C.: Ordinary Least Squares, In Encyclopedia of Mathematical Geosciences (pp. 1032-1038), Cham: Springer, https://doi.org/10.1007/978-3-030-85040-1_237, 2023.

Kramer, O., and Kramer, O.: K-nearest neighbors, Dimensionality reduction with unsupervised nearest neighbors, 13-23, https://doi.org/10.1007/978-3-642-38652-7_2, 2013.

Krizhevsky, A., Sutskever, I., and Hinton, G. E.: Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst., 25, Available at https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf, last access: 25 January 2025, 2012.

Kuhn, M.: caret: Classification and Regression Training R package version 6.0-92, Available at https://CRAN.R-project.org/package=caret, last access: 25 January 2025, 2022.
Kunz, M., and Kottmeier, C.: Orographic enhancement of precipitation over low mountain ranges, Part II: Simulations of heavy precipitation events over southwest Germany, J. Appl. Meteorol. Clim., 45(8), 1041-1055, https://doi.org/10.1175/JAM2390.1, 2006.
Lacerda, W. A., Palmeira, E. M., Netto, A. L. C., and Ehrlich, M. (Eds.).: Extreme rainfall induced landslides: an international perspective, Oficina de Textos, ISBN: 978-85-7975-150-9, 2014.
Lann, T., Bao, H., Lan, H., Zheng, H., and Yan, C.: Hydro-mechanical effects of vegetation on slope stability: A review, Sci. Total Environ., 171691, https://doi.org/10.1016/j.scitotenv.2024.171691, 2024.
LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, Nature, 521(7553), 436-444, https://doi.org/10.1038/nature14539, 2015.
Lee, D. B., Kim, Y. N., Sonn, Y. K., and Kim, K. H.: Comparison of Soil Taxonomy (2022) and WRB (2022) Systems for classifying Paddy Soils with different drainage grades in South Korea, Land, 12(6), 1204, https://doi.org/10.3390/land12061204, 2023.
Lee, D. H., Cheon, E., Lim, H. H., Choi, S. K., Kim, Y. T., and Lee, S. R.: An artificial neural network model to predict debris-flow volumes caused by extreme rainfall in the central region of South Korea, Eng. Geol., 281, 105979, https://doi.org/10.1016/j.enggeo.2020.105979, 2021.
Lee, D. H., Kim, Y. T., and Lee, S. R.: Shallow landslide susceptibility models based on artificial neural networks considering the factor selection method and various non-linear activation functions, J. Remote Sens., 12(7), 1194, https://doi.org/10.3390/rs12071194, 2020.
Lee, J. U., Cho, Y. C., Kim, M., Jang, S. J., Lee, J., and Kim, S.: The effects of different geological conditions on landslide-triggering rainfall conditions in South Korea, Water, 14(13), 2051, https://doi.org/10.3390/w14132051, 2022.
Lee, M. J.: Rainfall and landslide correlation analysis and prediction of future rainfall base on climate change, In Geohazards Caused by Human Activity, IntechOpen, https://dx.doi.org/10.5772/64694, 2016.
Lee, S. W., Kim, G., Yune, C. Y., and Ryu, H. J.: Development of landslide-risk assessment model for mountainous regions in eastern Korea, Disaster Adv., 6(6), 70-79, 2013.
Li, C. J., Guo, C. X., Yang, X. G., Li, H. B., and Zhou, J. W.: A GIS-based probabilistic analysis model for rainfall-induced shallow landslides in mountainous areas, Environ. Earth Sci., 81(17), 432, https://doi.org/10.1007/s12665-022-10562-y, 2022.
Liaw, A., and Wiener, M.: Classification and regression by randomForest, R News 2(3), 18-22, Available at https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf, last access: 24 January 2025, 2002.
Liu, Y., Deng, Z., and Wang, X.: The effects of rainfall, soil type and slope on the processes and mechanisms of rainfall-induced shallow landslides, Appl. Sci., 11(24), 11652, https://doi.org/10.3390/app112411652, 2021a.
Liu, Z., Gilbert, G., Cepeda, J. M., Lysdahl, A. O. K., Piciullo, L., Hefre, H., and Lacasse, S.: Modelling of shallow landslides with machine learning algorithms, Geosci. Front., 12(1), 385-393, https://doi.org/10.1016/j.gsf.2020.04.014, 2021b.
Luino, F., De Graff, J., Biddoccu, M., Faccini, F., Freppaz, M., Roccati, A., Ungaro, F., D'Amico, M., and Turconi, L.: The Role of soil type in triggering shallow landslides in the alps (Lombardy, Northern Italy), Land, 11(8), https://doi.org/1125. 10.3390/land11081125, 2022.
Martinović, K., Gavin, K., Reale, C., and Mangan, C.: Rainfall thresholds as a landslide indicator for engineered slopes on the Irish Rail network, Geomorphology, 306, 40-50, https://doi.org/10.1016/j.geomorph.2018.01.006, 2018.
McKenna, J. P., Santi, P. M., Amblard, X., and Negri, J.: Effects of soil-engineering properties on the failure mode of shallow landslides, Landslides, 9, 215-228, https://doi.org/10.1007/s10346-011-0295-3, 2012.
McKight, P. E., and Najab, J.: Kruskal-wallis test, The corsini encyclopedia of psychology, 1-1, https://doi.org/10.1002/9780470479216.corpsy0491, 2010.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F.: e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien R package version 1.7-9, https://doi.org/10.32614/CRAN.package.e1071, 2021.
Miao, F., Wu, Y., Xie, Y., and Li, Y.: Prediction of landslide displacement with step-like behavior based on multialgorithm optimization and a support vector regression model, Landslides, 15, 475-488, https://doi.org/10.1007/s10346-017-0883-y, 2018.
Montgomery, D. R., Schmidt, K. M., Dietrich, W. E., and McKean, J.: Instrumental record of debris flow initiation during natural rainfall: Implications for modeling slope stability, J. Geophys. Res. Earth Surf., 114(F1), https://doi.org/10.1029/2008JF001078, 2009.
Nguyen, Q. H., Ly, H. B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., Prakash, I., and Pham, B. T.: Influence of data splitting on performance of machine learning models in prediction of shear strength of soil, Math. Probl. Eng., 2021(1), 4832864, https://doi.org/10.1155/2021/4832864, 2021.
O'brien, R. M.: A caution regarding rules of thumb for variance inflation factors, Qual. Quant., 41, 673-690, https://doi.org/10.1007/s11135-006-9018-6, 2007.
Omwega, A. K.: Crop cover, rainfall energy and soil erosion in Githunguri (Kiambu District), Kenya, The University of Manchester (United Kingdom), Available at https://www.proquest.com/openview/dd7c169f804775d18041ec262d03e4c1/1?cbl=2026366&diss=y&pq-origsite=gscholar, last access: 24 Janurary 2025, 1989.
Panday, S., and Dong, J. J.: Topographical features of rainfall-triggered landslides in Mon State, Myanmar, August 2019: Spatial distribution heterogeneity and uncommon large relative heights, Landslides, 18(12), 3875-3889, https://doi.org/10.1007/s10346-021-01758-7, 2021.
Park, C. Y.: The classification of extreme climate events in the Republic of Korea, J. Korean Assoc. Regional Geograp., 21(2), 394-410, Available at https://koreascience.kr/article/JAKO201507740043627.page, last access: 25 January 2025, 2015.
Park, S. J., and Lee, D. K.: Predicting susceptibility to landslides under climate change impacts in metropolitan areas of South Korea using machine learning, Geomat. Nat. Hazards Risk. and Risk, 12(1), 2462-2476, https://doi.org/10.1080/19475705.2021.1963328, 2021.
Pham, B. T., Tien Bui, D., and Prakash, I.: Bagging based support vector machines for spatial prediction of landslides, Environ. Earth Sci., 77, 1-17, https://doi.org/10.1007/s12665-018-7268-y, 2018.
Phillips, C., Hales, T., Smith, H., and Basher, L.: Shallow landslides and vegetation at the catchment scale: A perspective, Ecol. Eng., 173, 106436. https://doi.org/10.1016/ j.ecoleng.2021.106436, 2021.
Pisner, D. A., and Schnyer, D. M.: Support vector machine, In Machine learning (pp. 101-121), Academic Press, https://doi.org/10.1016/B978-0-12-815739-8.00006-7, 2020.
Pourghasemi, H. R., and Rahmati, O.: Prediction of the landslide susceptibility: Which algorithm, which precision?, Catena, 162, 177-192, https://doi.org/10.1016/j.catena. 2017.11.022, 2018.
Qiu, H., Regmi, A. D., Cui, P., Cao, M., Lee, J., and Zhu, X.: Size distribution of loess slides in relation to local slope height within different slope morphologies, Catena, 145, 155-163, https://doi.org/10.1016/j.catena.2016.06.005, 2016.
R Core Team: R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, Available at https://www.R-project.org/, last access: 24 January 2025, 2022.
Rahman, M. S., Ahmed, B., and Di, L.: Landslide initiation and runout susceptibility modeling in the context of hill cutting and rapid urbanization: a combined approach of weights of evidence and spatial multi-criteria, J. Mt. Sci., 14(10), 1919-1937, https://doi.org/10.1007/s11629-016-4220-z, 2017.
Ran, Q., Wang, J., Chen, X., Liu, L., Li, J., and Ye, S.: The relative importance of antecedent soil moisture and precipitation in flood generation in the middle and lower Yangtze River basin, Hydrol. Earth Syst. Sci., 26(19), 4919-4931, https://doi.org/10.5194/hess-26-4919-2022, 2022.
Rathore, S. S., and Kumar, S.: A decision tree regression-based approach for the number of software faults prediction, ACM SIGSOFT Software Engineering Notes, 41(1), 1-6. https://doi.org/10.1145/2853073.2853083, 2016.

Razakova, M., Kuzmin, A., Fedorov, I., Yergaliev, R., and Ainakulov, Z.: Methods of calculating landslide volume using remote sensing data, In E3S Web of Conferences (Vol. 149, p. 02009), EDP Sciences, https://doi.org/10.1051/e3sconf/202014902009, 2020.

Rosi, A., Peternel, T., Jemec-Auflič, M., Komac, M., Segoni, S., and Casagli, N.: Rainfall thresholds for rainfall-induced landslides in Slovenia, Landslides, 13, 1571-1577, https://doi.org/10.1007/s10346-016-0733-3, 2016.

Rotaru, A., Oajdea, D., and Răileanu, P.: Analysis of the landslide movements, Int. J. Coal Geol., 1(3), 70-79, Available at https://naun.org/multimedia/NAUN/ geology/ijgeo-10.pdf, last access: 24 January 2025, 2007.

Saito, H., Korup, O., Uchida, T., Hayashi, S., and Oguchi, T.: Rainfall conditions, typhoon frequency, and contemporary landslide erosion in Japan, Geology, 42(11), 999-1002, https://doi.org/10.1130/G35680.1, 2014.

Saleh, A. M. E., Arashi, M., and Kibria, B. G.: Theory of ridge regression estimation with applications, John Wiley and Sons, ISBN: 9781118644614, 2019.

Sato, T., Katsuki, Y., ans Shuin, Y.: Evaluation of influences of forest cover change on landslides by comparing rainfall-induced landslides in Japanese artificial forests with different ages, Sci. Rep., 13(1), 14258, https://doi.org/10.1038/s41598-023-41539-x, 2023.

Scheidl, C., Heiser, M., Kamper, S., Thaler, T., Klebinder, K., Nagl, F., Lechner, L., Markart, G., Rammer, W., and Seidl, R.: The influence of climate change and canopy disturbances on landslide susceptibility in headwater catchments, Sci. Total Environ., 742, 140588, https://doi.org/10.1016/j.scitotenv.2020.140588, 2020.

Seger, C.: An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing, Available at https://www.diva-portal.org/smash/get/diva2:1259073/FULLTEXT01.pdf, last access: 24 January 2025, 2018.

Shirzadi, A., Shahabi, H., Chapi, K., Bui, D. T., Pham, B. T., Shahedi, K., and Ahmad, B. B.: A comparative study between popular statistical and machine learning methods for simulating volume of landslides, Catena, 157, 213-226, https://doi.org/10.1016/ j.catena.2017.05.016, 2017.

Singh, D., and Singh, B.: Feature wise normalization: An effective way of normalizing data, Pattern Recognition, 122, 108307, https://doi.org/10.1016/j.patcog.2021.108307, 2022.

Smith, H. G., Neverman, A. J., Betts, H., and Spiekermann, R.: The influence of spatial patterns in rainfall on shallow landslides, Geomorphology, 437, 108795, https://doi.org/10.1016/j.geomorph.2023.108795, 2023.

Stoof, C. R., Vervoort, R. W., Iwema, J., Van Den Elsen, E., Ferreira, A. J. D., and Ritsema, C. J.: Hydrological response of a small catchment burned by experimental fire, Hydrol. Earth Syst. Sci., 16(2), 267-285, https://doi.org/10.5194/hess-16-267-2012, 2012.

Sun, H. Y., Wong, L. N. Y., Shang, Y. Q., Shen, Y. J., and Lü, Q.: Evaluation of drainage tunnel effectiveness in landslide control, Landslides, 7, 445-454, https://doi.org/10.1007/s10346-010-0210-3, 2010.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K.: Measuring and testing dependence by correlation of distances, https://doi.org/10.1214/009053607000000505, 2007.

Tacconi Stefanelli, C., Casagli, N., and Catani, F.: Landslide damming hazard susceptibility maps: a new GIS-based procedure for risk management, Landslides, 17, 1635-1648, https://doi.org/10.1007/s10346-020-01395-6, 2020.

Tsai, T. L., and Chen, H. F.: Effects of degree of saturation on shallow landslides triggered by rainfall, Environ. Earth Sci., 59, 1285-1295, https://doi.org/10.1007/ s12665-009-0116-3, 2010.

Turner, T. R., Duke, S. D., Fransen, B. R., Reiter, M. L., Kroll, A. J., Ward, J. W., Bach, J. L., Justice, T. E., and Bilby, R. E.: Landslide densities associated with rainfall, stand age, and topography on forested landscapes, southwestern Washington, USA, For. Ecol. Manag., 259(12), 2233-2247, https://doi.org/10.1016/ j.foreco.2010.01.051, 2010.

Um, M. J., Yun, H., Cho, W., and Heo, J. H.: Analysis of orographic precipitation on Jeju-Island using regional frequency analysis and regression, Water Resour. Manag., 24, 1461-1487, https://doi.org/10.1007/s11269-009-9509-z, 2010.

Van Westen, C. J.: The modelling of landslide hazards using GIS, Surv. Geophys., 21(2), 241-255, https://doi.org/10.1023/A:1006794127521, 2000.

Wang, D., Hollaus, M., Schmaltz, E., Wieser, M., Reifeltshammer, D., and Pfeifer, N.: Tree stem shapes derived from TLS data as an indicator for shallow landslides, Procedia Earth Planet. Sci., 16, 185-194, https://doi.org/10.1016/j.proeps.2016.10.020, 2016.

Wei, Z. L., Shang, Y. Q., Sun, H. Y., Xu, H. D., and Wang, D. F.: The effectiveness of a drainage tunnel in increasing the rainfall threshold of a deep-seated landslide, Landslides, 16, 1731-1744, https://doi.org/10.1007/s10346-019-01241-4, 2019.

Wieczorek, G.: Debris flows/avalanches: process, recognition, and mitigation, Volume VII, GSA, Boulder, Colorado, ISBN:0-8137-4107-6, 1987.

Willmott, C. J., and Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, Climate Res., 30(1), 79-82, https://doi.org/10.3354/cr030079, 2005.

Yan, L., Xu, W., Wang, H., Wang, R., Meng, Q., Yu, J., and Xie, W. C.: Drainage controls on the Donglingxing landslide (China) induced by rainfall and fluctuation in reservoir water levels, Landslides, 16, 1583-1593, https://doi.org/10.1007/s10346-019-01202-x, 2019.

Yoon, S. S., and Bae, D. H.: Optimal rainfall estimation by considering elevation in the Han River Basin, South Korea, J. Appl. Meteorol. Climatol., 52(4), 802-818, https://doi.org/10.1175/JAMC-D-11-0147.1, 2013.

Yun, H. S., Um, M. J., Cho, W. C., and Heo, J. H.: Orographic precipitation analysis with regional frequency analysis and multiple linear regression, Korea Water Resour. Assoc., 42(6), 465-480, https://doi.org/10.3741/JKWRA.2009.42.6.465, 2009.

Yune, C. Y., Jun, K. J., Kim, K. S., Kim, G. H., and Lee, S. W.: Analysis of slope hazard-triggering rainfall characteristics in Gangwon Province by database construction, J. Korean Geotech. Soc., 26(10), 27-38. https://doi.org/10.7843/kgs. 2010.26.10.27, 2010.

Zaruba, Q., and Mencl, V.: Landslides and their control, Elsevier, ISBN: 0444600760, 9780444600769, 2014.

Zhang, K., Wang, S., Bao, H., and Zhao, X.: Characteristics and influencing factors of rainfall-induced landslide and debris flow hazards in Shaanxi Province, China, Nat. Hazards Earth Syst. Sci., 19(1), 93-105, https://doi.org/10.5194/nhess-19-93-2019, 2019.