1 **Prediction of landslide induced debris' severity using machine learning**

2 **algorithms: a case of South Korea**

4 Tuganishuri Jérémie[1], Chan-Young Yune[2], Gihong Kim[3], Seung Woo Lee[4],

5 Manik Das Adhikari[5], Sang-GukYum[6*]

7 [1] Department of Civil and Environmental Engineering, Gangneung-Wonju National

8 University, Gangneung, Gangwon-do, Republic of Korea, tugan.miya@gmail.com

9 [2] Department of Civil and Environmental Engineering, Gangneung-Wonju National

10 University, Gangneung, Gangwon-do, Republic of Korea, yune@gwnu.ac.kr

11 [3] Department of Civil and Environmental Engineering, Gangneung-Wonju National

12 University, Gangneung, Gangwon-do, Republic of Korea, ghkim@gwnu.ac.kr

13 [4] Department of Civil and Environmental Engineering, Gangneung-Wonju National

14 University, Gangneung, Gangwon-do, Republic of Korea, swl@gwnu.ac.kr

15 [5] Department of Civil and Environmental Engineering, Gangneung-Wonju National

16 University, Gangneung, Gangwon-do, Republic of Korea**,** rsgis.manik@gmail.com

17 [6] Department of Civil and Environmental Engineering, Gangneung-Wonju National

18 University, Gangneung, Gangwon-do, Republic of Korea, skyeom0401@gwnu.ac.kr

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

Prediction of landslide induced debris' severity using machine learning algorithms: a case of South Korea

Abstract

Rainfall-induced landslides frequently occur in the mountainous region of Korean peninsula. The resulting landslide induced debris cause extreme property damages, huge financial losses and human deaths. To mitigate their effect different landslide susceptibility mapping are frequently used. However, these methods do identify regions with potential landslides but they do not quantify their severity. In this paper multi-category ordered machine models, namely, proportional odd logistic regression (POLR), random forest (RF), support vector machine (SVM), and extreme gradient boosting (EGB) methods, are proposed to fill the specified gap. Moreover, the exploratory data analysis on landslide induced derris's dataset has been conducted on to examine patterns and relationship between landslide-induced debris severity(size), causal factors(rainfall) and influencing factors. Findings revealed that cumulative three days' rainfall and slope length were most responsible for the severity of landslide-originated debris severity and slopes between 20° to 40° was identified as most vulnerable region. Furthermore, the predictive accuracy statistics were compared to assess the suitable model for debris severity for Korean case. The RF and EGB ranked higher with an overall accuracy of 90.07% and 86.09% and *kappa* of 0.72 and 0.61 on the validation set, respectively. The findings of this research may be useful in the identification of high risk zones for extreme rainfall-induced debris to elaborate mitigation and resilience policies, post-disaster rehabilitation planning and land use management.

Keywords: Rainfall-induced debris Severity; Proportional odd logistic regression; Random forest; Support vector machine; Extreme gradient boosting; Machine learning , South Korea

1. Introduction

Rainfall-induced debris is a natural phenomenon that occurs when the slope fails due to the saturation of soil after the rainfall exceeds a certain threshold (Au 1998; Takara et al. 2010; Peruccacci et al. 2017; Segoni et al. 2018; Crawford et al. 2019; Coppola et al. 2022). Rainwater penetrates the soil through cracks or pores (Zeng et al. 2022) which destabilizes the slope and induces landslides (Franzluebbers 2002). Furthermore, the volume of landslide-induced debris depends on the geological condition of the terrain, rainfall intensity and duration (Chang et al. 2011). Extreme rainfall is the triggering factor for landslides which is one of the most damaging natural disasters with the expensive cost of repair and indemnification (Kockelman 1986; Gariano and Guzzetti 2016). In addition, heavy windstorms, typhoons, and extensive rainfall have destroyed many properties and taken many human lives yearly (Liu et al. 2018). Furthermore, landslides have caused enormous environmental degradation, infrastructure damage, casualties, and loss of life, which disturb the socio-economic aspect of the community (Li et al. 2012; Sarkar and Dorji 2019; Zhao et al. 2019; Taylor et al. 2020; Lacroix et al. 2020; Winter 2020; Negi et al. 2020; Ju et al. 2020; Van et al. 2021). Most rainfall-induced landslides were found to be shallow (de Jesús et al. 2019; Liu et al. 2021; Chang et al. 2021) however, some were very extreme and resulted in severe human and financial damage (Turner, 2018; Meena et al., 2021). Klose et al. (2016) found that from 1980 to 2013, landslides took thousands of lives and an annual average of about $20 bilion of economic losses, which was 17% of the total ($121 billion) annual mean of global disaster-induced losses.

The Korean peninsula is characterized by mountainous, which makes it more prone to rainfall-induced landslides (Lee et al. 2013). Lee et al. (2012) found that the triggering factor for landslides was short-duration heavy rainfall. Park et al. (2013) reported that the annual property damage caused by rainfall-induced landslides in South Korea averaged between US$500M to US$1000M and approximately 36 human deaths per year from 1997 to 2010.

76   Therefore, to mitigate the effect of landslides in South Korea, different studies have been

77   carried out on landslide assessment (Kadavi et al. 2019; Lee and Winter 2019; Sameen et al.

78   2020; Panahi et al. 2020; Hakim et al. 2022). Lee et al. (2020) applied the Naïve and Bayesian

79   Networks model for landslide susceptibility mapping in Umyeon Mountain. Lee et al. (2012)

80   used physical slope and probabilistic model, i.e., decision trees and logistic regression for

81   landslide susceptibility mapping in Gangwon-do. Lee et al. (2013) developed the binary

82   logistic regression model for predicting the occurrence of landslides. Woo et al. (2014)

83   constructed a landslide hazard map using binary logistic regression. Park and Kim (2019)

84   compared boosted trees and random forest model's performance in landslide susceptibility

85   mapping for Umyeon Mountain; the same methods were previously applied at Pyeong-Chang

86   by Kim et al. (2018). It was observed that the objectives of previous studies were to predict

87   landslide susceptibility; they did not specify how severe the occurring landslides would be.

88   Further, most of studies were performed on a small scale and only predicted the occurrence,

89   not the size. Therefore, in the present study, we analyzed landslide-induced debris severity

90   based on the causative variables and influencing factors. This study is novel in expressing the

91   relationship of debris' severity, causative and influencing factors. It is an extention on landslide

92   mapping which quantifies the magnitude of landslide-induced debris. The quantification of

93   debris severity may be useful in land management by highlighting regions prone to higher

94   rainfall-induced debris to know whether economic activities that may be carried out on the

95   given region may not be vulnerable to extreme landslide hazards. The severity of debris is

96   measured in unit of volume (m^3) and classified as shallow(below $500m^3$) small(500-$2000m^3$),

97   medium (2000-$5000m^3$) and critical, i.e., above $5000m^3$. Words severity of debris, debris

98   volume or size of debris express the same quantity in different ways and are used

99   interchangeably in this manuscript.

2. Study region

Korean peninsula is located in the northern hemisphere, between China and Japan in Northeast Asia. Its climate has continental and oceanic features with wide temperature differences. The yearly mean temperature ranges from 10°C to 16°C, that is, from -6°C to 7°C in winter and 23°C to 27°C in summer. In South Korea, the rainy season range from June to September, with 1000mm to 1800mm of precipitation in the southern part and 1100 to 1400mm in the central region (https://web.kma.go.kr/).

The altitude ranges from 0 to 1911 meters, with mount Halla (in Jeju Island) being the highest peak in South Korea. The Gangwon Province is the most mountainous region of about 64% of all tallest mountains in Korea, that is, 23 of 36 mountains. The surface geology of the Korean peninsula is mostly composed of igneous, sedimentary, and metamorphic rock (Chough et al., 2000). The arable soil depth varies between 1 to 2m (Lee and Winter, 2019). Due to the high intensity rainfall and weak geological formation in the mountainous region causes high frequency of landslides. Figure 1(a) illustrates the distribution of landslides by the volume of landslides, while subplots (b & c) exhibited the relationship of slope length, altitude, and rainfall with the landslide size.

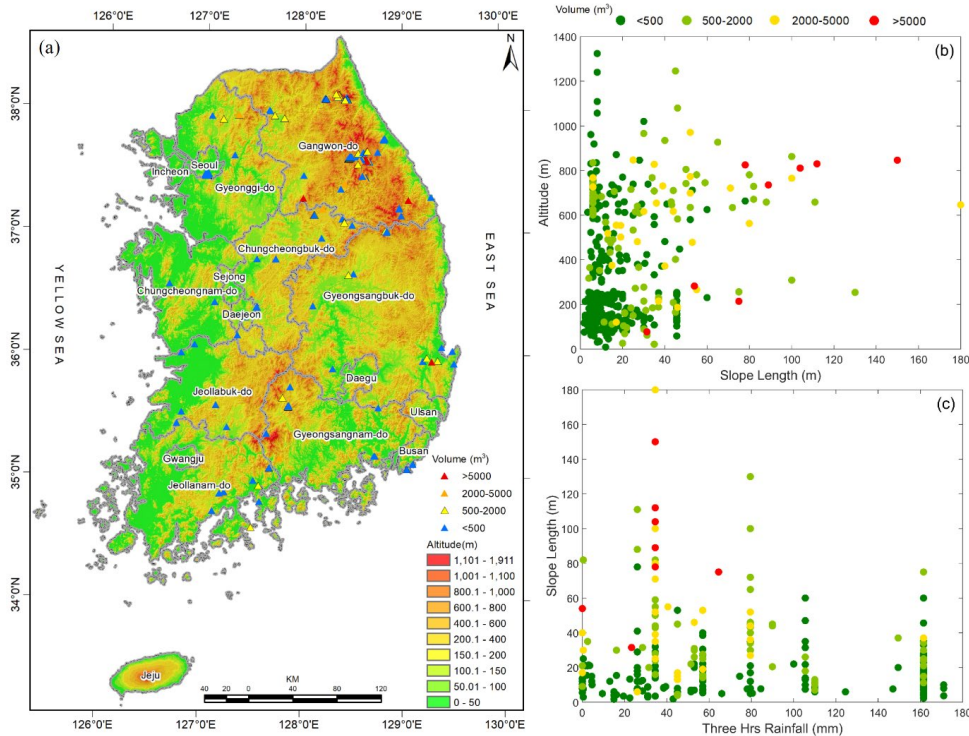



Figure 1. (a) Landslide location in South Korea, (b) Distribution of landslide volume per slope length and altitude, and (c) Distribution of landslide volume per slope length and rainfall (Data source: elevation data acquired from NGII, 2018).

3. Methodology

3.1. Problem formulation

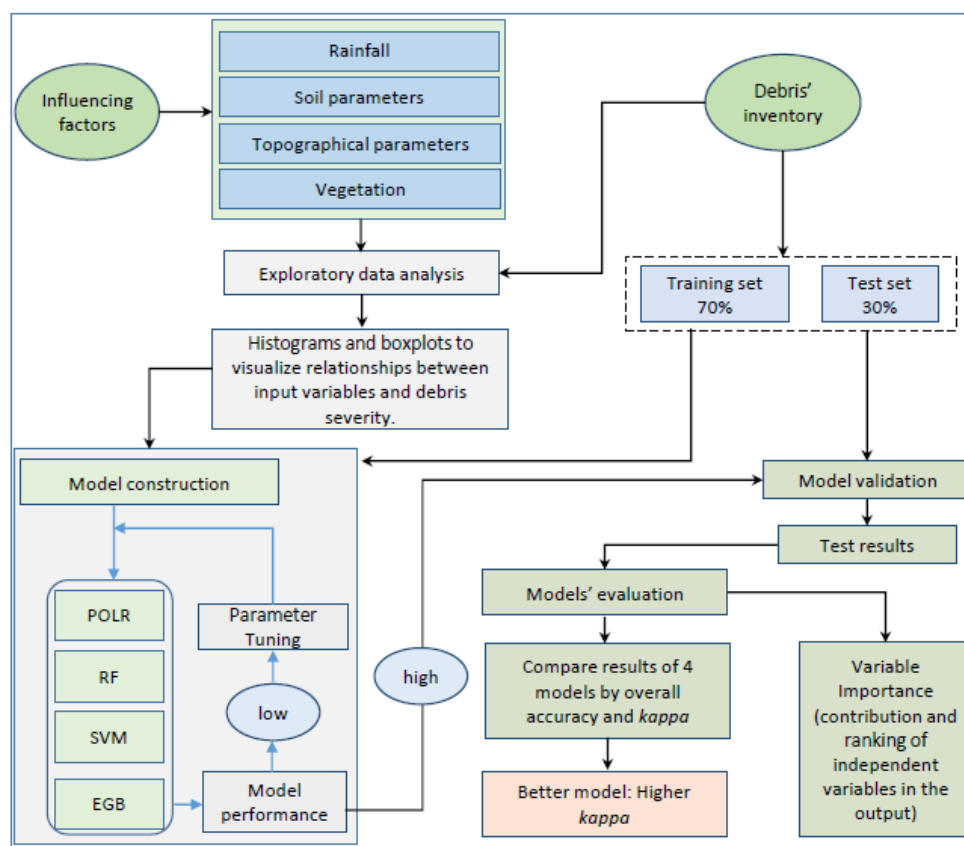
Predictive models that deal with multi-variate random variables were investigated to predict the severity of rainfall-induced landslides. Among those predictive models, proportional odd logistic regression and other machine learning-based algorithms such as extreme gradient boosting, random forest, and support vector machine are widely used to deal with classification problems (Marjanović et al. 2011; Lee et al. 2012; Woo et al. 2014; Wang et al. 2022a). The main steps for the modeling process were to analyze the interaction of variables

128   that influence the severity of landslides with results in a higher size of debris. For the purposes,

129   four machine learning models i.e., POLR, RF, SVM, and EGB, were used to assess the most

130   suitable predictive model for landslide-induced debris in south Korea (;Su et al.,2022). The

131   comparison was made using predictive accuracy and the value of *kappa*. Figure 2 shows the

132   steps followed in the construction of the model.

133

134   **Figure 2.** Modelling workflow process for the prediction of landslide induced debris' severity

135   using machine learning algorithms.

136   *3.2. Data description*

137   The dataset contains 455 debris inventory collected from field surveys with the help of

138   portable GPS, a laser ranger and a clinometer. Variable definitions and descriptions for each

139   feature in the dataset are presented in Table 1. The rainfall data were collected from Korea

140   meteorological Administration stations scattered around the country nearest each landslide-

141   induced debris site. It was revealed that the duration and quantity of rainfall directly affect

142   landslides (Berti et al. 2012; Kim et al. 2014; Sarkar et al. 2019; Liu et al. 2020; Ngo et al.

143   2021). Rainfall data were grouped into twelve variables: cumulative rainfall, continuous hourly

144   rainfall, three hours, six hours, nine hours, twelve hours, one day, three days, seven days, two

145   weeks, three weeks, and four weeks' rainfall. Different measures of rainfall were captured due

146   to the time-dependent cumulative effect of rainfall on the slope stability, and prolonged rainfall

147   has a more damaging effect in mountainous regions (Baum and Godt 2010; Hidalgo et al. 2017;

148   Meena et al. 2022). The conditioning factors, i.e., soil type, topsoil depth, altitude, slope, slope-

149   length, slope aspect, and vegetation (leafage, size of tree, age of trees, and fire history), were

150   collected. The soil type was classified as sandy loam, lithosols, silt loam, and clay. The soil

151   depth was grouped into below 20 centimeters, between 20-50, and 50-100 centimeters. The

152   rainfall infiltrating the topsoil causes saturation, which initiates the landslide and then results

153   in debris flow (Baum and Godt 2010;Vahedifard et al., 2017; Zhu and Zhang 2019). The anti-

154   erosive drainage presence and status were categorized into: very good, good, and bad. Dranage

155   channels reduces the concentration of water in soil and effect on water flow, saturation, soil

156   moisture, and valley landslides(Shahabi and Hashim 2015). The vegetation-covered and

157   necked lands have different affect of landslide (Lee et al. 2013; Ozioko and Igwe 2020;Huang

158   et al. 2021). The foliage information was classified into pines, broad-leaved, and mixture, while,

159   the size of trees was classified as large, small, and medium. The age of trees was grouped into

160   seven classes viz. 1-5 5-15, 15-25, 25-35, 35-45 and >45 years. The forest fire history was also

161   considered as a influencing factor due to its erosive nature(Huang et al. 2020). Geographical

162   features were found to contribute to the severity of landslides at different levels; steep slopes

163   were found to fail as the intensity of the rainfall increased (Brand et al. 1984; Au 1998; Charles

164   and Shi 1998; Nandi and Shakoor 2008; Pham et al. 2018; Chen et al. 2020). It was also noted

165   that plane areas beneath steep slopes face the damage caused by debris flowing from the top of

166   mountains (Raja et al. 2017,Wang et al. 2022b).

167

168   The output variable (volume) of landslides has been classified into four categories, i.e.,

169   below 500m^3, between 500-2000m^3, 2000-5000m^3, and above 5000m^3(Fig. 4c). There are

170   different types of landslides; which are classified based on the cause or shape after occurrence

171   (Causes 2001). Landslides may result from liquefaction, earthquakes, intense surface water

172   flow due to precipitation, underground water, ice melting, human activities, tectonic

173   movements etc.(Alexander 1992; Causes, 2001; van der Beek, 2021;McColl 2022). In this

174   paper, we only considered landslide-induced debris originating from rainfall. Table 1

175   summarizes the characteristics of debris: its types, frequency and size.

176   Table 1: Landslide-induced debris types and frequency

Volume (m^3) / Destroyed area	<500	500-2000	2000-5000	>5000	Total
valley erosion	1	1	1		3
falling rocks	1				1
mixed/ complexes	3	2	1		6
slope	1	1		1	3
scour	1				1
curved wedges	4	1			5
a circular arc	205	45	14	2	266
Plane	120	35	10	5	170
Total	336	85	26	8	455

177   ***3.3 Exploratory data analysis***

178   The relationship between the influencing factors and the debris size were analyzed. We

179   consider categorical variables, also known as qualitative set of information, that is divided into

180   groups. It describes data which are non-numerical and serve qualitative purposes, such as

181 representing a certain value or distinctive character of objects (nominal), when there is no order

182 in levels and ordinal when there is an order in the dataset (Bilder and Loughin 2014). In addition,

183 continuous features were presented in form of histograms and boxplots to analyze the

184 dispersion and influence on the size of debris. Figure 3 illustrates that most of the occurred

185 debris was shallow (below 500m^3, 73.85%), followed by 500-2000m^3 (18.68%), and the

186 extreme debris (above 5000m^3) was the least frequent (1.76%). Debris with a size above

187 5000m^3 was mainly associated with sandy loam soil of depth above 20cm and a non-perfected

188 drainage system, where the forest density was medium and in a place that experienced wildfire.

189 The region with pines leafage experienced shallower debris compared to other types of leafage.

190 The region with older trees, above 45 years of age, experienced more severe debris than

191 younger forests. Šilhán and Stoffel (2015) highlighted that the area with timber age of above

192 45years were more sensitivity to landslide occurrences. The size of debris for wildfire-

193 experienced regions was observable; compared to the frequency of cases with no wildfire

194 history, the severity was quite higher. The wildfire influence on the rainfall-induced debris is

195 due to the reduced infiltration of water into the soil, which increases the erodibility of soil

196 (Ranger et al. 2020; Tiwari et al. 2020). The inadequate drainage system resulted in severe

197 debris(Popescu 2002), where the system was very good, no severe debris occurred. The

198 resulting huge number of shallow debris for a very good drainage system is due to the fact

199 those systems are usually created in the most vulnerable regions and the shallow debris is an

200 indication of improvement in landslide mitigation.

201

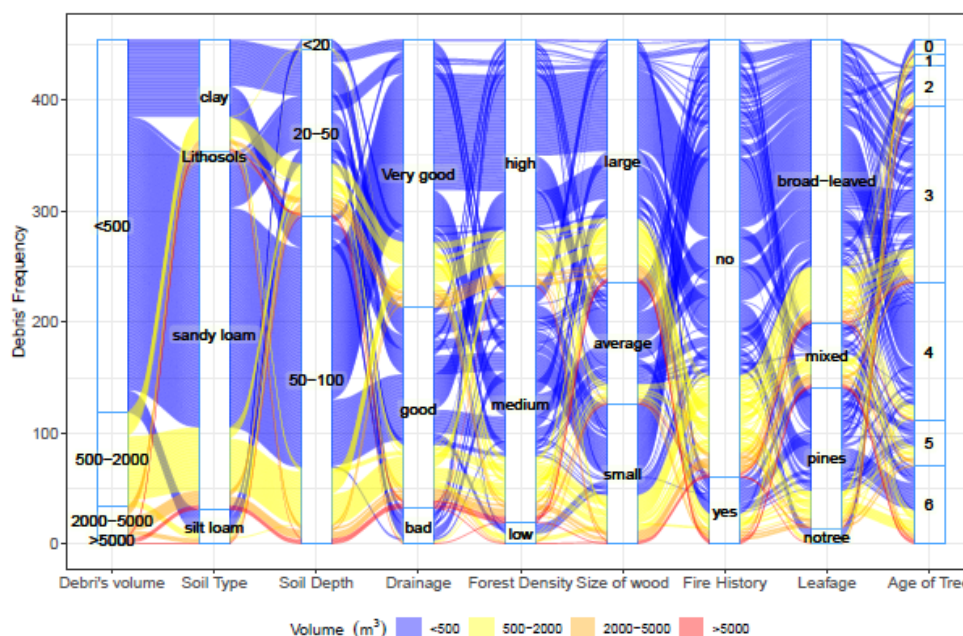



Figure 3. Interlinkage between the size of debris and categorical influencing factors.

203   Figure 4a illustrates that sandy soil, which totalizes about 70% of all debris, was the

204   most vulnerable soil, followed by clay (22%), silt loam (7%) and lithosols (1%). Similarly,

205   sandy soil not only ranked higher in terms of frequency but also in terms of size. The high

206   sensitivity of sandy soil to rainfall-induced debris may be due to its high coefficient of

207   permeability which facilitates fast saturation of topsoil during the rainfall period that induces

208   shallow debris (Lee et al. 2013). Overall, about 73.85% of all debris was shallow, that is, below

209   500m^3, 18.68 % (500-2000m^3) was small, 5.71% (2000-5000m^3) was medium, and only 1.76%

210   were critical, i.e., above 5000m^3, as depicted in Fig. 4b. Figure 4c, illustrates that about 74%

211   of landslide-induced debris occurred in Gangwon and Seoul; this made the two provinces more

212   vulnerable than other regions. South and North Gyeongsang provinces ranked third and fourth

213   with 7% (25 cases) and 5% (18 cases) of all debris inventory, respectively.

214

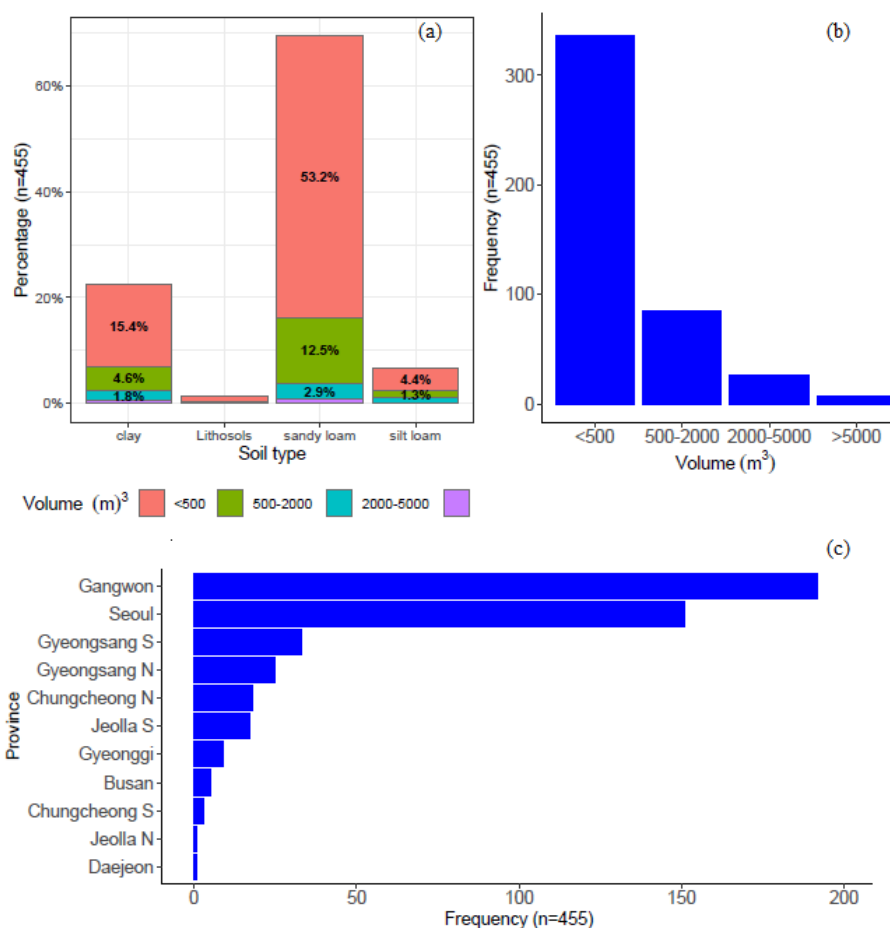



Figure 4. Distribution of debris': (a) debris' volume (in m³) per soil types, (b) debris'

frequency per volume, and (c) debris' frequency per provinces.

217

218   To examine the relationship between continuous explanatory variables and their effect

219 on the size of debris, 3D scatter plots and boxplots were presented (Fig. 5). It was observed on

220 the 3D scatterplot that most debris occurred at slope between 20° and 40° (Fig. 5a), this was

221 also confirmed by the boxplot (Fig. 5b), where few exceptions were observed for shallow

222 debris for clay and sandy soil where debris occurred at small or at very large slope angles as

223 outliers. Shallow debris was independent of the slope angle as depicted in Fig.5c; outliers were

224   scattered on all slope angles, and as the size of debris increased, the occurrence converged at

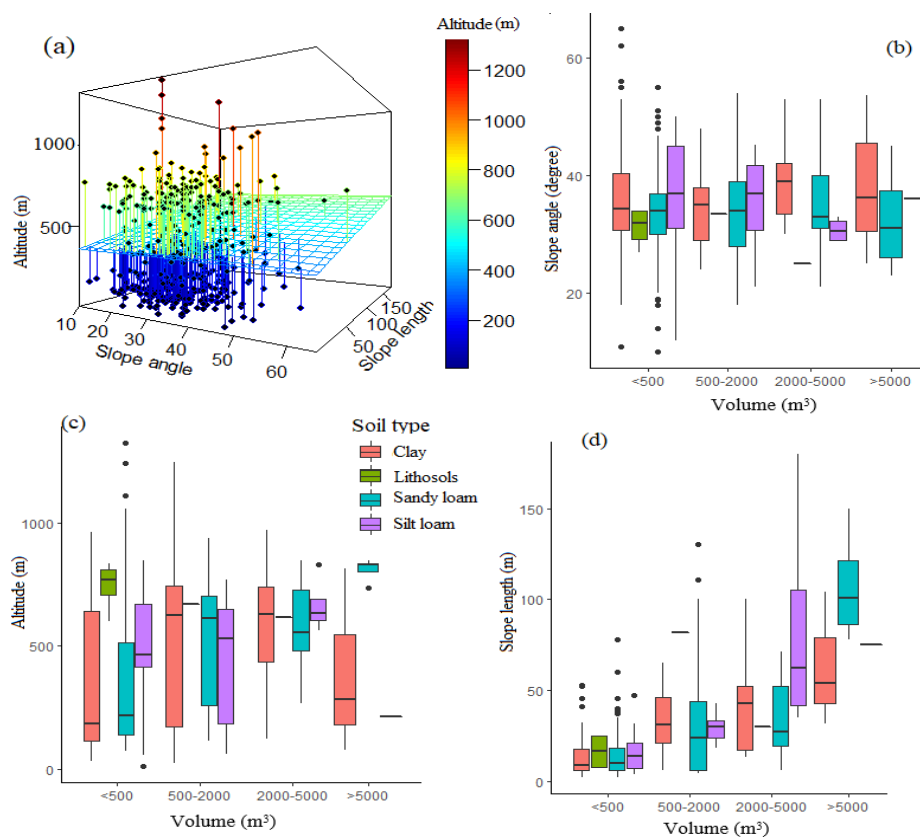
225   slope between 20° and 40°. There was a pseudo-decreasing trend between altitude and size of

226   debris (Fig. 5c), the highest altitudes (above 600m) were associated with shallow debris and

227   critical debris occurred at altitude between 200 and 900m. On the other hand, a quasi-increasing

228   relationship between size of debris and slope length was observed in all type of soil (Fig. 5d).

229   Debris below 500m^3 was associated with slope length below 80m, the highest quartile was

230   about 140m associated with clay soil of above 5000m^3 of debris size.

231

232   Figure 5. Distribution of debris and continuous predictors: (a) 3D Scatter plot between altitude,

233   topographic slope and slope length, (b) boxplot of volume of debris per slope angle and

234   soil types, (c) boxplot of volume of debris per altitude and soil types, and (d) boxplot of

235   volume of debris per slope length and soil types.

236   The analysis of the relationship between the size of debris and rainfall over different

237 time intervals has shown that short time rainfall was associated with shallow debris(fig.6 a-f),

238 as the cumulative time interval of rainfall increased, the debris size increased and stabilized at

239 three days cumulative rainfall(Fig. 6g). we observed that the increase in severity of debris was

240 associated with lower precipitation as reflected in the yellow boxplot for debris above

241 $5000m^3$(Fig. 6. g-f), the precipitation of occurrence was below the corresponding median

242 rainfall. Therefore the lower precipitation on prolonged time greater or equal to three days was

243 responsible for severe debris. On the other hand, short-duration heavy rains were responsible

244 for shallow debris(Polemio and Petrucci 2000).The relationship between shallow debris and

245 heavy rain was reflected in the boxplot representing debris below $500m^3$(Fig. 6 a-f), where the

246 interquartile range of precipitation was large.

247

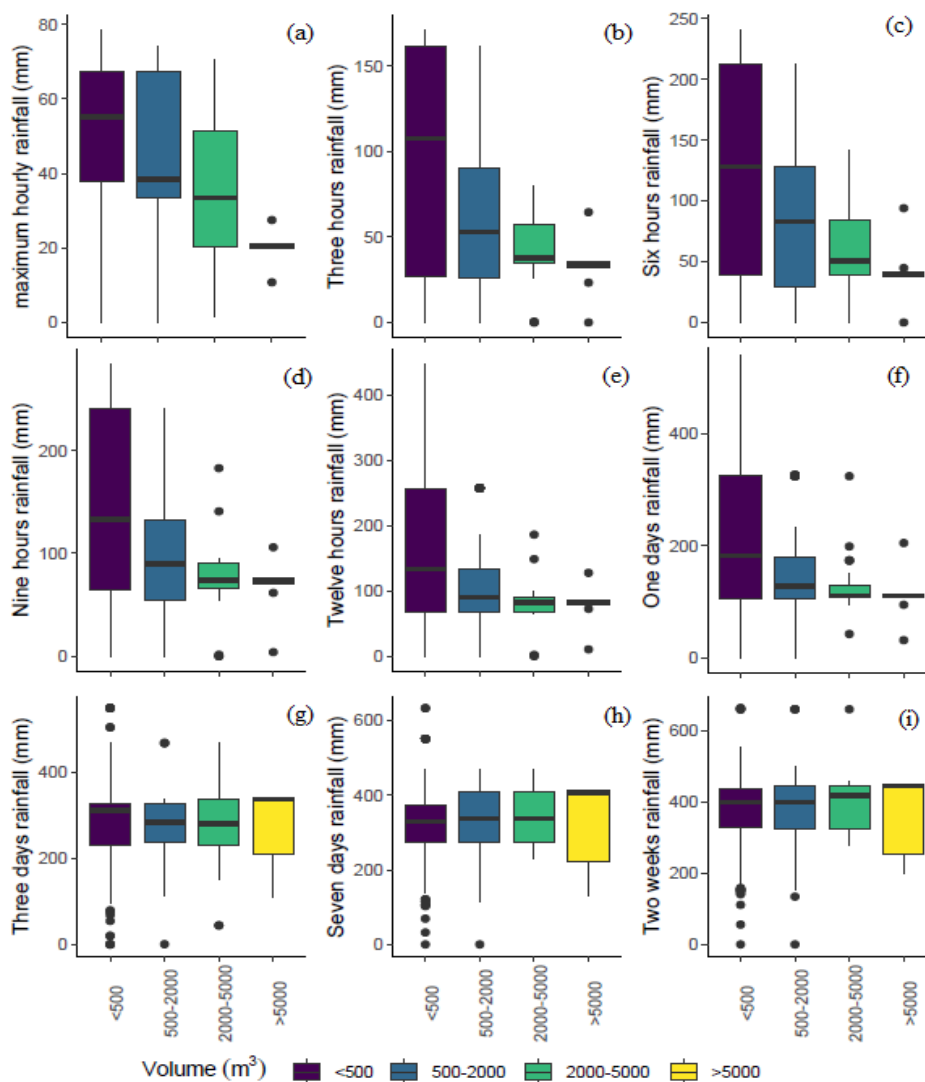



Figure 6. Distribution of severity of debris per antecedent rainfall groupped in different time
intervals

250

251   The precipitation at the time of the incident exhibited an inverse relationship between

252   the size of debris and rainfall intensity (Fig. 7 a-f). From three days' cumulative of antecedent

253   rainfall, the relationship became almost constant in all time-based cumulative rainfall (Fig. 7

254   g-h). The rainfall of lower intensity falling over a prolonged period was observed to trigger the

255 large size of rainfall-induced debris (Tang et al. 2017; Rivas et al. 2022). The threshold for the

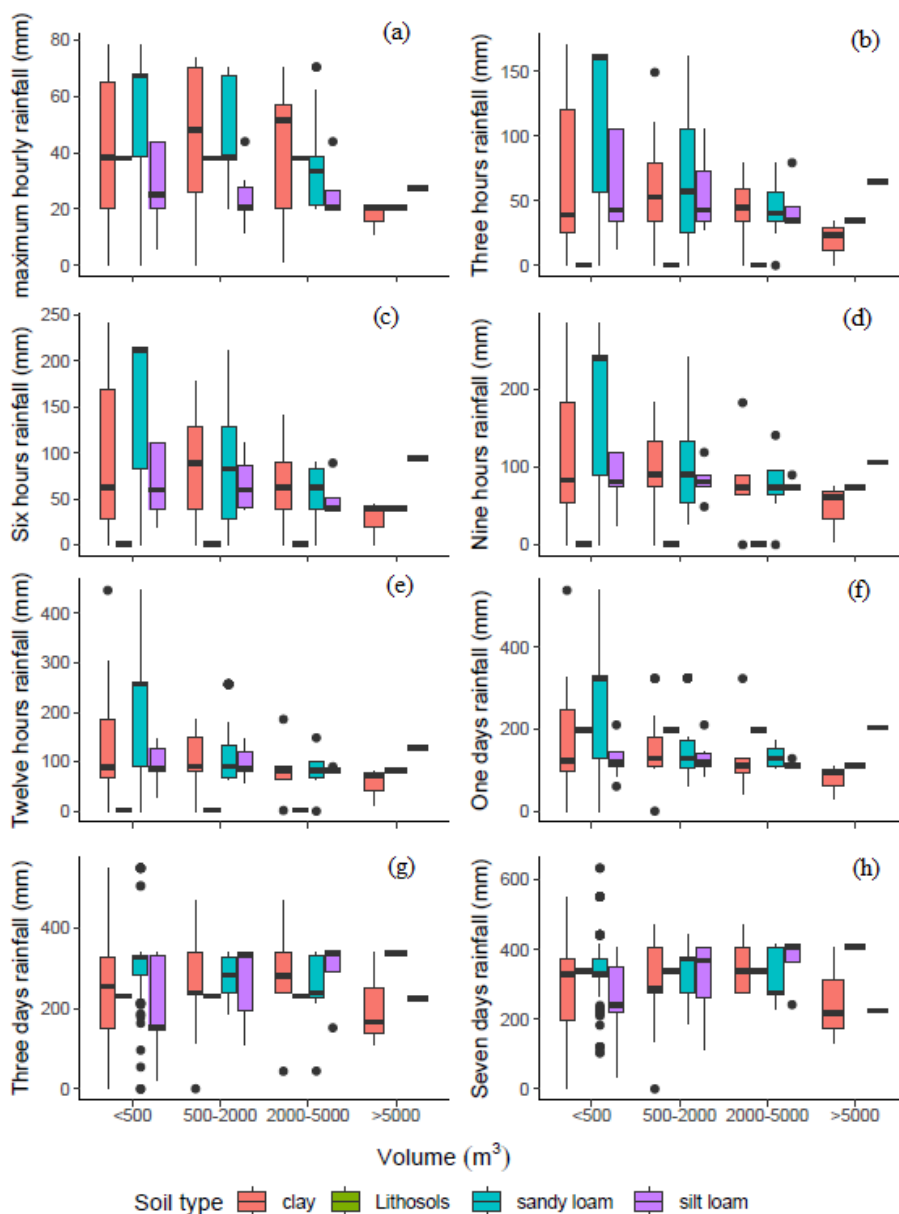
256 triggering factor was the rainfall duration of atleast 3days (Chinkulkijniwat et al., 2020;

257 Rahardjo et al., 2020). In terms of soiltype, clay exhibited shallow debris at lowest

258 precipitation followed by sandy soil. The median of the occurrence of shallow debris was the

259 lowest for silt loam compared with other soil types. From three days cumulative rainfall(Fig.

260 7g), the median of occurrence of all size of debris stabilized around 300mm of precipitation

261 and clay was more likely to produce severe debris.

263   Figure 7. Relationship between rainfall and the size of debris.

3.3. Method Description

The analysis, construction, and evaluation of models were done in the following chronological order as depicted in Fig. 2. First, the data set was curated: formatting the variables to match their types, that is, numerical variables into a numerical format, and categorical variables into factor and ordered factors according to their natural characteristics (Table 1). Second, the dataset was split into training and test set. Third, the four machine learning algorithms (Kainthura and Sharma 2022; Su et al.,2022), were applied to the training and test set on all variables consecutively. Finally, confusion matrices for each method were generated to compare the performance based on the overall accuracy and *kappa*. The variable ranking plot was generated to identify the cause of differences in the predictive power of the four methods.

All analyses were done using the following packages in R software: Caret (Max 2022) for the creation of confusion matrix, dplyr (Wickham et al. 2022) for data manipulation and formatting, MASS for running proportional odd logistic model, Random Forest for running the random forest algorithm, Xgboost for running extreme gradient boost, SVM for supporting vector machine, Ggplot2 (Wickham 2016), alluvial (Brunson 2020) for plotting, and Matrix for creation of sparse matrix which is used in training extreme gradient boosting.

Variable importance is a systematic approach for identifying the contribution of input variable in the prediction of the outcome variable in the predicative model. For graphical representation of variable importance (Biecek and Burzykowski 2021), the plot was made using DALEX libraries (Biecek 2018) and the e1071 package (Meyer et al. 2021) in R (Team 2021). This plot ranks variables according to its influence on the predictive power of the model.

The selected methods for modeling were chosen based on low parsimony and are frequently applied to ordered outcome problems. The predictive performance for each model was evaluated using confusion matrix information accuracy and *kappa* (Caelen 2017). The

290   kappa statistic k measures the agreement between the observed and predicted values to quantify

291   the ability of the model to classify the output variable into their appropriate classes or categories.

292   Let both the observed Y variable and predicted Z variable have g categories or levels. Let f_{ij}

293   be the frequencies of observations in the ith categorical output variable Y and the jth category

294   of the predicted values, then the frequency table known as the confusion matrix can be arranged

295   as follows:

	Z=1	Z=2	...	Z=g
Y=1	f_{11}	f_{21}	...	f_{1g}
Y=2	f_{21}	f_{22}	...	f_{2g}
⋮	⋮	⋮	...	⋮
Y=g	f_{g1}	f_{2g}	...	f_{gg}

296

297   The observed (actual values) ratio of agreement between Y and Z is expressed as:

298
$$p_0 = \frac{1}{n}\sum_{i=1}^{g} f_{ii} \qquad (1)$$

299

300   and the expected agreement by chance is defined as:

301
$$p_e = \frac{1}{n^2}\sum_{i=1}^{g} f_{i+}f_{+i} \qquad (2)$$

302

303   where f_{i+} is the total for the ith row f_{i+} is the total for the ith column. The value of kappa is

304   the estimate of the population coefficient calculated using the following formula:

305
$$k = \frac{\Pr[Y=Z] - \Pr[y=z|Y \text{ and } Z \text{ are independent}]}{1 - \Pr[y=z|Y \text{ and } Z \text{ are independent}]} \qquad (3)$$

306

307   The confusion matrix   is useful for analyses, control tunes of different classifiers, and

308   identification of a combination of classes with its recognition values or rates (Susmaga 2004).

309   The confusion matrix is a (g × g) dimension table (matrix) which matches predicted values

310   from the model vs. actual values from the dataset, where g stands for levels of the outcome

311   variable, entries on the diagonal represents the correct classification, and non-diagonal

312   elements represent misclassification. The accuracy is defined as the ratio of correctly classified

313   entries and the sum of correctly classified and misclassification. The accuracy of the model is

314   compared to the value of the No information rate (NIR). The NIR is the baseline for assessing

315   performance, not 0.5. For the model to be useful (better than random guess), the lower bound

316   for a 95% confidence interval (CI) of prediction must be greater than NIR (Garson 2021). The

317   next paragraphs describe each method in detail.

318   The proportional odds logistic model (POLR) (McCullagh 1980) is one of the usefull

319   methods designed to handle ordered or ranked outcome variables when the outcome categories

320   (levels) are more than two. This model is constructed based on cumulative probability

321   distribution (Brant 1990), $y_j = \Pr(y \leq j)$ and is expressed in the form:

$$\text{logit}\ [y_j/(1-y_j)] = \theta_j - \beta^t \mathrm{X} \qquad (4\ )$$

323   where y is a set of N and independent observation taking values j= 1, 2, …, k, X is a

324   vector of independent variables, $\theta_1 < \theta_2 < \cdots < \theta_{k-1}$ and β^t are unknown parameters. To use

325   ordinal regression, assumptions must be satisfied.

326   To use the proportional odds logistic regression, the proportional odd assumption or the

327   parallel regression conditions must be satisfied. The first states that no independent variable

328   has a disproportionate effect on any level of the dependent variable (McNulty 2021). If this

329   condition is not satisfied, other methods such as adjacent logit models may be used (Agresti,

330   2010; Harrell, 2015). To test the parallel regression, the Brant-Wald test is used and this test

331 compares the general ordinal logistic regression (with no assumption of proportional odd) with

332 POLR. It is the test of significance in the difference between the two models, which produces

333 chi-square statistics. If the p-value of each of the different coefficients of variables in the model

334 is greater than 0.05, then the parallel regression assumption holds (Brant 1990) or the

335 assumption is violated otherwise. Binary logistic regression is frequentry used for landslide

336 susceptiblity mapping(Yesilnacar and Topal 2005; Yilmaz 2009; Lin et al. 2017; Lombardo and

337 Mai 2018; Sun et al. 2021), the ordered logistic regression is an extension for binary logistic

338 applied to solve problems with multilevel ordered outcomes(Brant 1990).

339   The second used method was the Random Forest which is a classification and regression

340 methods. The RF algorithm (Breiman 2001) is a combination of tree predictors and every tree

341 is made based on values of random vectors, which are sampled independently using the same

342 distribution to create all trees of the forest. In this study, the RF classification method (Biau

343 and Scornet 2016; Lechner and Okasa 2020;) is appropriate due to its capability to handle

344 multiple outcome-related problems (Diaz-Uriarte and de Andrés 2005). It was applied in

345 different regions for landslide susceptibility mapping(He et al. 2021; Sun et al. 2021; Huang et

346 al. 2022).

347   The third used method is SVM, which was widely used for mapping the likelihood of

348 landslides (Lee et al., 2017). Among the multiple class prediction methods, the SVM method

349 performed better for protein fold recognition (Ding and Dubchak 2001;Huang and Zhao 2018),

350 landslide hazard (Hong et al. 2015), landslide spatial prediction (Pham et al. 2018), etc. SVM

351 performed not only in multiclass problems but also better in ordered multilevel problems and

352 it worked better than traditional regression methods (Li et al. 2012). More details about the

353 SVM algorithm is described by Noble (2006).

354   The last but not least to choose was the EGB. Extreme gradient boosting is also a machine

355 learning algorithm known for its high-speed performance and efficient prediction accuracy

356 (Chang et al. 2018). The algorithm is based on Friedman's work (Friedman 2001) and

357 implemented in R under the package xgboost (Chen et al. 2015). This method is known for its

358 high performance on larger sample sizes (Georganos et al. 2018).

359 **4. Analysis & discussion**

360 To identify the suitable model for the prediction of the size of debris, previously discussed

361 methods were explored, and results were summarized and compared in this section.

362 *4.1 Debris prediction using POLR*

363 The significance of the observed relationship between the size of the debris and the

364 associated explanatory variables was measured using proportional odd logistic regression.

365 Variable selection was made using backward selection (Andersen 2010). The above coefficients

366 can be interpreted as follows: taking into consideration of p-values, and for each case supposing

367 that all other coefficients were held constant, the decrease in three hours' rainfall is associated

368 with a 94% lower odd the size of debris, an increase is six hours' rainfall was associated with

369 4% increase in higher odd of the size of the debris (Table 2).

370 Table 2. POLR model coefficients.

Variables	Coefficient	P value	Odds ratio
Three hours rain	-0.051	<0.01	0.949
Six hours rain	0.039	<0.01	1.040
One day rain	-0.009	0.01	0.990
Slope length	0.039	<0.01	1.040
altitude	0.001	0.02	1.001
Drainage: Good	1.246	0.13	3.477
Drainage: Very good	1.818	0.04	6.164
Intercepts:			
<500|500-2000	2.313	<0.01	10.105
500-2000|2000-5000	4.455	<0.01	86.102
2000-5000|>5000	6.428	<0.01	619.397

371

372

373   The goodness of fit and model diagnostic result shows that McFadden pseudo-R^2 was

374   0.23 where the values between 0.2 and 0.4 indicated an excellent fit (Louviere et al. 2000). To

375   test proportional odd assumptions, the Brant-Wald test (Brant, 1990) was used, and the result

376   was summarized in Table 3. It shows that all probabilities for all variables including the

377   omnibus are greater than the 0.05 threshold, except for slope length, which proves that the

378   parallel regression assumption is satisfied.

379   Table 3. Parallel regression assumption test.

Test for	X^2	df	Probability
Omnibus	12.6	14	0.56
Three hours rain	1.76	2	0.41
Six hours rain	1.18	2	0.56
One day rain	1.18	2	0.55
Slope length	7.1	2	0.03
altitude	1.31	2	0.52
Drainage	0.19	4	1

380

381   The effect plot was used to demonstrate the change, in the likelihood of occurrence of

382   landslides of a given volume, associated with the change in selected predictors. Figure 7 depicts

383   the variation of probability of landslides of a given volume per three days' rainfall and slope

384   length. The likelihood of occurrence of debris below 500m^3 was directly proportional to an

385   increase in the rainfall for the slope length below 30m. There was a decrease in the occurrence

386   of debris larger than 500m^3 as the rainfall increased. This decrease is an indication that shorter

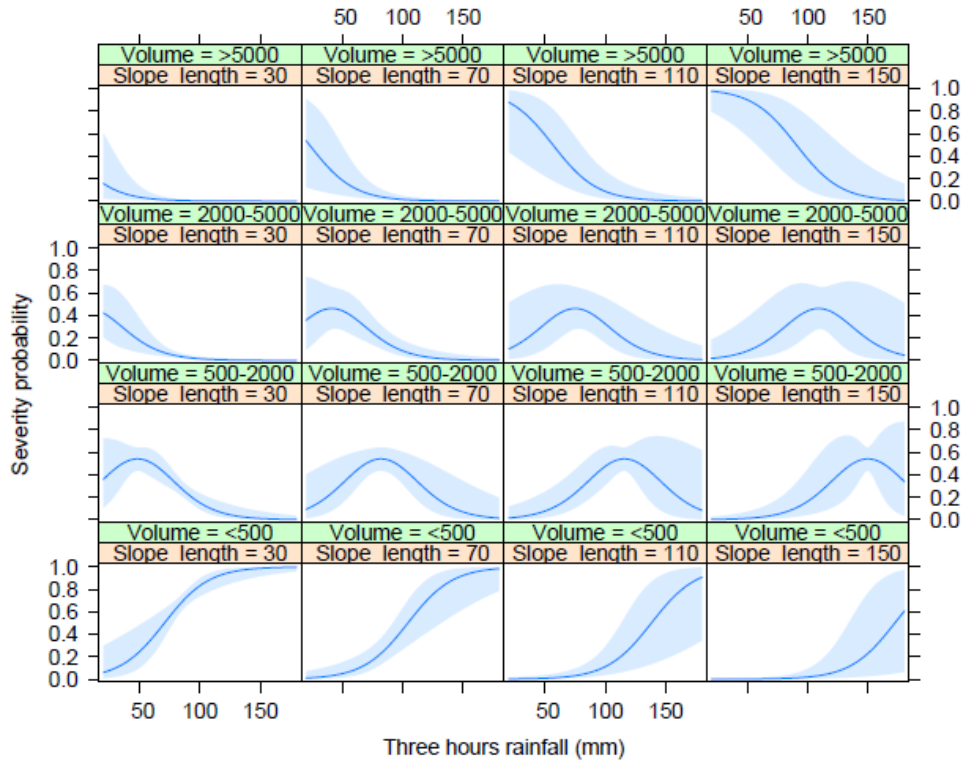
387   slope length was associated with a shallow debris (First column of Fig. 7). Moving from the

388   first to the 4th column, the probability of shallow debris (below 500m^3), the probability shifted

389   from 0.9 to 0.6 as slope length increased from 30m to 170m, and the long tail of probability

390   plots for debris of size above 500m^3 disappeared as the slope-length and rainfall increased. This

391   shifted up the probability curve for critical debris from 0.2 to 0.9 for slope length below 30m

392   to 0.9 for slope length above 110m, respectively.

393

394   Figure 8. Effect of three hours of rainfall and slope length on size of debris.

395

396   Figure 9 illustrates that the maximum size of debris occurred at slope lengths above 50m

397   and with rainfall between 80mm to 110mm. Looking at the two upper light hand corners of Fig.

398   9, the probability of occurrence of debris above 5000m^3 peaked at slope length above 100m

399   and fade as the rain value increased, this fading associated with rain does not mean that there

400   is an inverse relationship with rainfall but it associated with the rarity of heavy rain in the

401   dataset. The probability of occurrence of shallow debris decreased as both altitude and slope

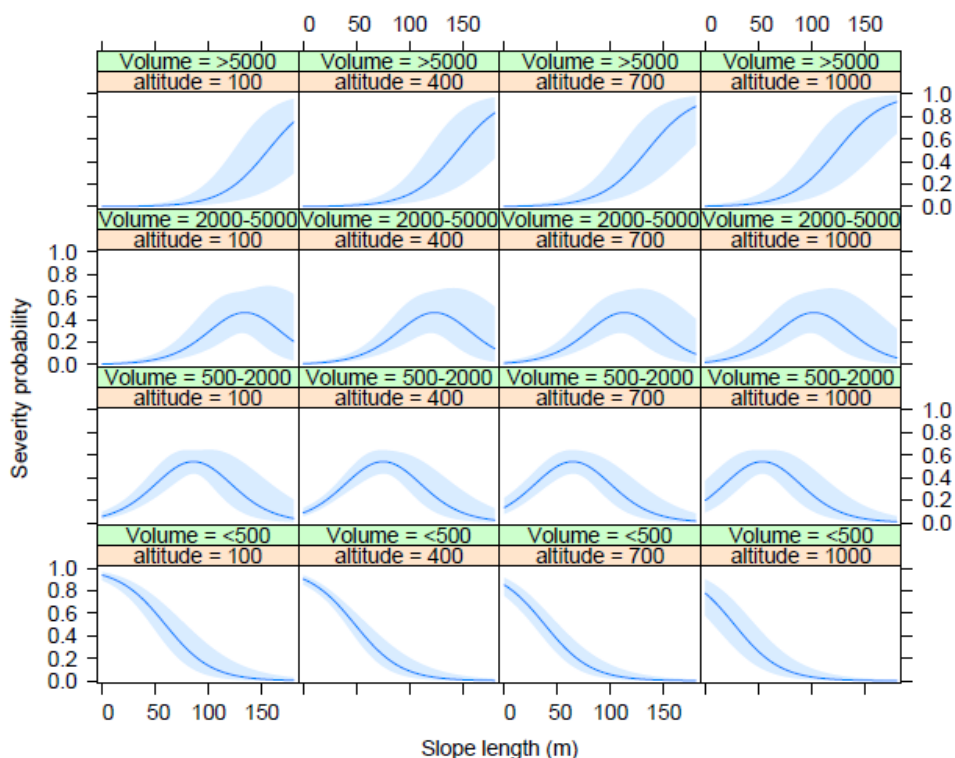
402   length increased. On the other hand, the chance of occurrence of debris of size between 500m^3

403   and 2000m^3 increased with slope length and attained the maximum length between 70 and 80m

404   with the maximum probability of occurrence of 0.6, and the last one decreased for slope length

405   above 100m. For the medeium sized debris, the maximum probability of occurrence was 0.7

406   and associated with slope lengths between 120m and 150m. The probability of occurrence of

407   critical debris increased exponentially with slope length and altitude.

408

409   Figure 9. Effect of altitude and slope length on size of debris.

410   The intuitive explanation of the continuous decrease with shallow debris is associated

411   with the cumulative character of the model, which means that, what looks like a decrease is not

412   a real decrease, it is a shift from a lower level to a higher level associated with an increase in

413   variables into consideration. For altitude below 300m, shifting from a size below 500m^3 to 500-

414   2000m^3 happened at 20m of slope length, while shifting from 500-2000m^3 to 2000-5000m^3 and

415   above occurred at slope length between 60 to 80m.

4.2 Debris prediction using RF

The random forest model was run on the training set of 304 observations. The number of grown trees was 1000, the variable tried at each split was 4, and the out-of-bag error estimates (OOB) error rate was 24.42%. Figure 10a depicts the predictions of a class below 500m^3 that had the least errors. Since the model is for classification, to calculate the prediction error, the model was run on the test set. The variable importance graph illustrates the slope length was the most contributing variable to the model accuracy (Fig. 10b). The mean decrease Gini (cross-entropy loss after permutation) value is the measure of the contribution of each variable to the homogeneity of nodes and leaves in the random forest (Martinez-Taboada and Redondo 2020). The higher the value, the more important the variable is, which exhibited that slope length is more influencing factor for the size of debris.

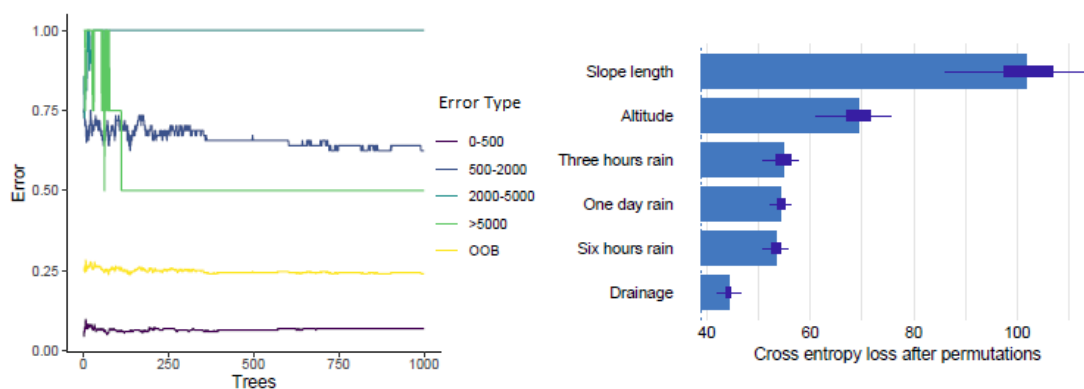



Figure 10. RF model: (a) training error rate, and (b) variable importance.

4.3 Debris prediction using EGB

To run the EGB, the train and test set were transformed into a sparse matrix, as it is run on numerical data matrices (Chen et al. 2022). The optimum model was obtained at the 881 number of iterations and the learning rate of 0.3. Since the task was a multi-classification, the multi: softmax objective was used (Chen et al. 2022). The evaluation metric was mlogloss

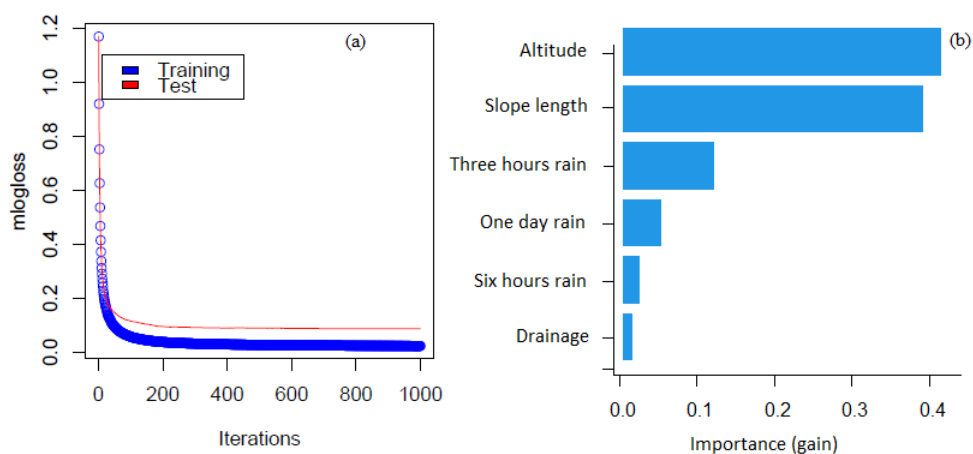
434   (Multi-class log loss) (Kabani and El-Sakka 2016). While training the model, the minimum

435   mlogloss were 0.025 and 0.088 for the training and test sets respectively (Fig. 11a). The

436   difference between training and the test error is due to the larger variance associated with the

437   fewer number of observations in the test set. The extreme gradient boosting associated with

438   higher importance to slope length, altitude, slope and three hours of rainfall, respectively (Fig.

439   11b).

440

441   Figure 11. EGB model: (a) error rate and (b) variable importance.

442   ***4.4 Debris prediction using SVM***

443   The support vector classification model with linear kernel was applied, and the number

444   of support vectors was 142. The outcome of SVM model shown a higher performance rate on

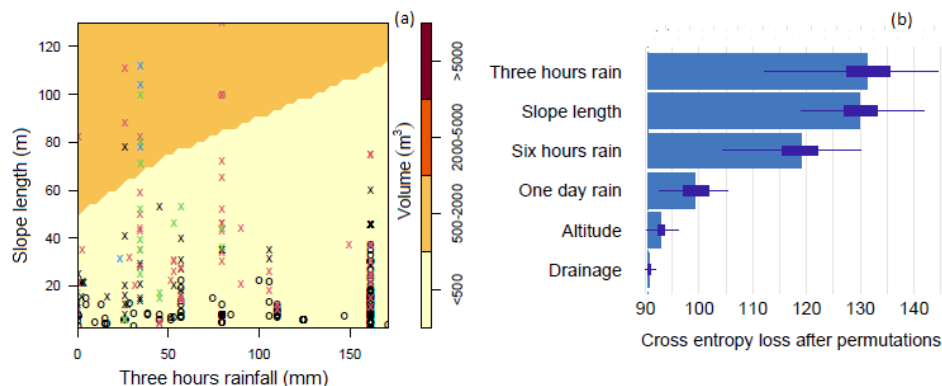
445   the test than the training set. Figure 11a depicts a two-dimensional projection of train data using

446   slope length and three hours of rainfall showing different shading and support vectors. One of

447   the weaknesses of the SVM, it classified predictions into two lower classes as indicated in Fig.

448   12a, it couldn't distinguish the moderates from extreme debris. The SVM assigned higher

449   importance to slope length and three hours of rainfall as the second high ranking variable as

450   shown in Fig. 11b.

451

452   Figure 12. SVM model: (a) SVM classification plot, and (b) variable importance.

453   *4.4 Discussion and model suitability assessment for landslide-induced debris severity*

454   *prediction*

455   This study analyzed the relationship between the severity of landslide-induced debris. The

456   exploratory data analysis revealed that 93% of occurred debris was below 2000m^3. The

457   Gangwon province and Seoul were more vulnerable regions in terms of the frequency of

458   incidents. Despite a higher frequency of debris in Seoul, their size was small compared to

459   Gangwon province with higher number of cases and more large sized-debris. To analyse the

460   significance of the relationship between the size of debris with different models ranked the

461   slope length as the most influential variable for the size of the debris. To visualize the

462   association of the slope lengh with the size of debris, the scatter plot (Fig.13) revealed that the

463   pattern of increasing trend of slope length and size of debris was more remarkable across all

464   provinces, shallow debris were associatted with slope length blow 50m. The debris cases were

465   clustered between 20° and 40°, as depicted in Fig.13, and critical debris tended to be clustered

466   around 30 degrees; the association of slope angle and size of debris was not statistically

467   significant. For the soil-debris size relationship, sandy soil and clay were associated with a

468   higher frequency of debris; they exhibited shallow debris in all provinces except in Gangwon

469   where severe debris occurred. It was observed that even though the silt loam soil was not highly

470   frequent, it was vulnerable to severe debris as the slope length increased. Gangwon province

471   was the region where the increasing relationship between the size of debris was observed, other

472   provinces were not prone to severe debris.

473

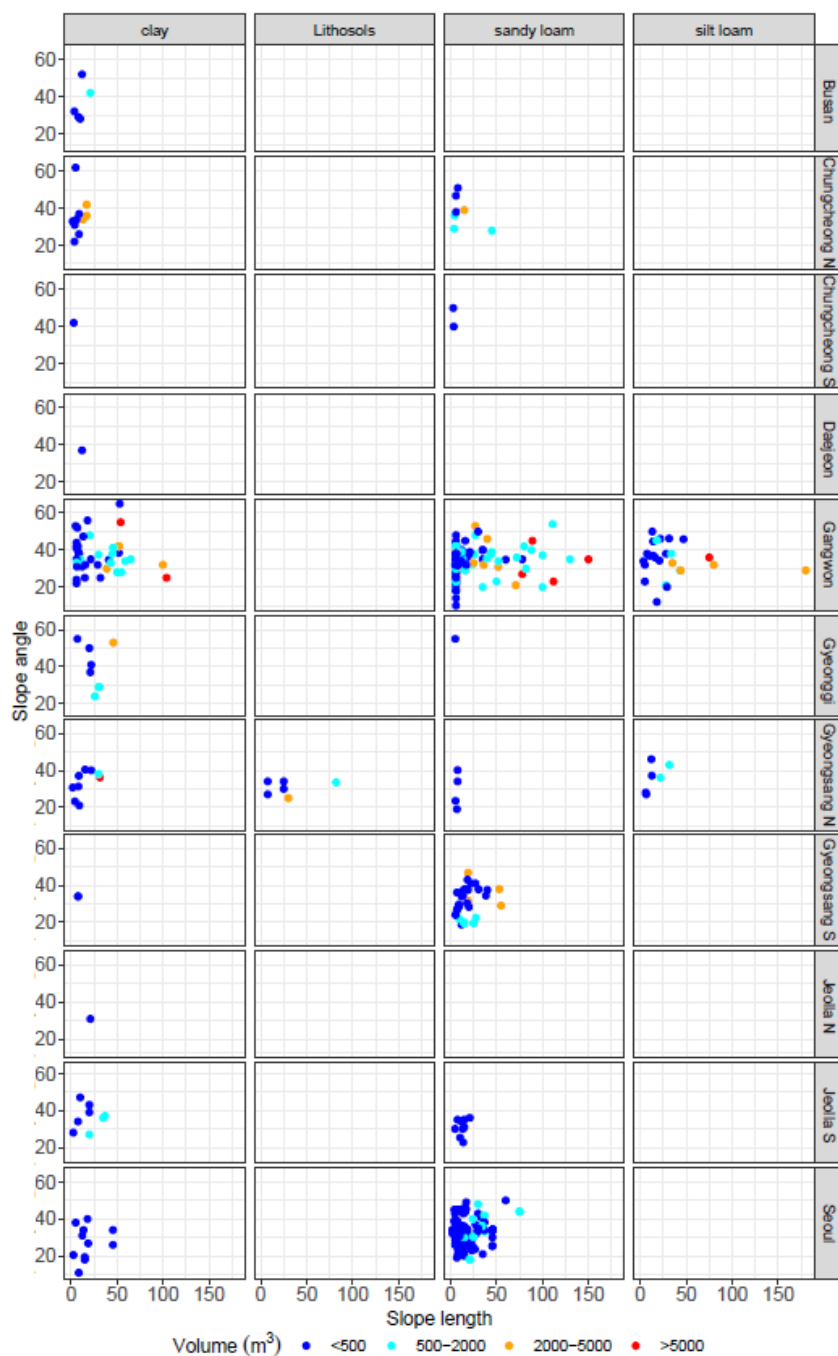



Figure 13. Province-wise scatter plot of debris' size per slope length, slope angle and soil

475   types.

476

477   To assess the suitability of utilized models, it was observed that the ordinal regression

478   prediction power was low due to its inherent weakness; it does not perform well on the

479   imbalanced dataset (Agresti 2010). That is, the outcome variable has some more prevalent

480   levels than others. It also estimates the coefficient using maximum likelihood techniques,

481   which require a large sample size, the used dataset was highly imbal

482   anced. Figure 7 revealed that almost 80% of all occurred debris was shallow, as result, the

483   model coefficients became unstable with larger prediction intervals. Despite the weakness, the

484   ordinal model has the effect display (Fox and Weisberg 2019), which clearly shows the effect

485   of each variable to each extent and the associated probability. The predictive performance for

486   each of the four discussed methods was summarized in Table 4 to facilitate their comparison.

487   The random forest model performed well in all cases on the training set and validation as well.

488   This model associated the influencing factor with higher importance and lower importance to

489   rain-associated variables. RF prediction accuracy was very high on the training set (0.93), and

490   0.90 on the test set. The prediction at 95% of confidence interval width ranged from 0.84 to

491   0.94 on training and test sets, respectively. The NIR shifted from 72.28% to 74.17%, and this

492   small increase is due to a small sample of the validation set.

493   The model accuracy for POLR was quite moderate based on the *kappa* value of 0.38. The

494   performance accuracy was better, the no information rate, NIR=0.7228<0.7314 lower

495   boundaries of CI on the training set. This last condition was not satisfied for the test set, that is

496   NIR=0.74 was higher than 0.69, which was the lower CI for the prediction interval on the test

497   set. Based on the p-value for the prediction on test 0.26 > 0.05, the performance was not reliable.

498   This is confirmed by the overall performance metric *kappa* =0.30, which was quite moderate.

499

500

501   The EGB model ranked the second best model after the RF, it satisfied all prediction

502   conditions the overall performance (*kappa* = 0.6121) was slightly below the random forest

503   0.7273. The SVM model result was satisfactory for the training set. The lower bound of very

504   close to the information rate and the overall performance was moderate (*kappa*=0.32), the

505   weakness of the model was its incapacity to distinguish the moderate debris from the extreme

506   ones; as result, it predicted all debris into two lower categories in Fig. 8a. The NIR fell into the

507   prediction interval on the validation set, and the p-value was 0.11>0.05, which is an indication

508   of moderate prediction accuracy.

509   Table 4. Model accuracy statistics for the four methods.

		Model accuracy statistics				
Method	**Data**	**Accuracy**	**95% CI**	**NIR**	**P-Value**	***kappa***
	Train	0.93	(0.89, 0.95)	0.72	<0.001	0.82
RF	Test	0.90	(0.84, 0.94)	0.74	<0.001	0.72
	Train	0.78	(0.73, 0.82)	0.72	0.011	0.38
POLR	Test	0.76	(0.69, 0.83)	0.74	0.26	0.3
	Train	0.77	(0.72, 0.82)	0.72	0.015	0.32
SVM	Test	0.78	(0.71, 0.85)	0.74	0.11	0.3
	Train	0.86	(0.81, 0.89)	0.72	<0.001	0.63
EGB	Test	0.86	(0.79, 0.91)	0.74	0	0.6

510

511   The landslide-induced debris prediction is an extension of landslide susceptibility mapping and

512   may be useful in the quantification and prediction of debris resulting from a rainfall-induced

513   landslide. This quantification can facilitate risk management (Ho and Ko 2009), in the

514   identification of regions prone to severe debris and the making of policies for mitigation

515   (Carmela and Mario Parise 2022). For example the decision of planting more vegetation that

516   fits the conditions of the region to strengthen the soil or deciding an appropriate activity to be

517   done in a given region to improve stability ,safety and efficient land use (Mayer et al. 2008) .

518   Furthermore, some activities in regions prone to severe debris may be prohibited for the safety

519   and well-being of the public(Frattini et al. 2010; DeGraff and Romesburg 2020; Di Napoli et

520   al. 2020).   In addition, the model may serve the disaster manager to create appropriate funds

521   for post-disaster recovery. For example, if a region is expected to have shallow debris, the

522   manager may establish a small fund for paying minor labour to repair the damaged environment.

523   In extreme cases, a big fund may be created to pay for machinery and construction of preventive

524   walls, plantation, and cost of machinery to remove debris in the affected region, to rehabilitate

525   the impacted economic activities in the neighbourhood (Kachi et al. 2016). Due to the lack of

526   financial data associated with the inventories the cost of post-disaster recovery was not

527   estimated, more studies in the future may be carried out to fill this gap. The approach in this

528   paper is valid for the studied area based on the user input data; more research in the future may

529   be conducted to know whether the findings in this paper are general for regions with different

530   characteristics or settings.

531   **6. Conclusions**

532   The study analyzed the relationship between the size of rainfall-induced debris and causal

533   factors, i.e., time-based cumulative rainfall and influencing factors: soil types, vegetation, and

534   geomorphology features. The exploratory data analysis revealed that the Gangwon province is

535   prone to more frequent and more severe landslide-induced debris. Soil-related information

536   revealed that the landslide-induced debris was more frequent in sandy soil and more severe,

537   but its influence was not statistically significant in the predictive model. The region with non-

538   perfected drainage systems also experienced severe debris. The regions with old timber that

539   experienced fire had a higher debris likelihood. To examine the significance and to identify the

540   suitable model for landslide-induced debris severity, four predictive modeling techniques i.e.,

541   POLR, RF, EGB and SVM, were applied to examine the causal and influencing factors of the

542   severity of rainfall-induced debris in South Korea. The performance metrics, accuracy and

543   *kappa* were applied to compare the predictive power of each of the four methods. The findings

544   of this research revealed that three hours' rainfall, one-day rainfall, and slope length were the

545   most influencing factor and altitude took the second place. This finding was consistent with the

546   results of Lee et al. (2012), stating that short-duration rainfalls were responsible for landslides

547   and are the cause of their severity. The comparative analysis has shown that random forest had

548   better predictive power with an accuracy of 90% and *kappa* = 0.72, and extreme gradient

549   boosting followed with an accuracy of 86% (*kappa* = 0.6). The last two methods SVM with an

550   accuracy of 78% (4% above NIR), and POLR performance was moderate at 76%, which is

551   only 2% above 74% performance decision basis (No information rate NIR), but we did not

552   have enough information to confirm their use as a basis for creation early warning system for

553   rainfall-induced extreme debris. This is because POLR does not perform well on limited and

554   imbalanced data (Rahman et al. 2021), which is the root cause of a wide range of prediction

555   intervals. Thus, RF and EGB may be used as a suitable models for rainfall-induced debris

556   prediction. The creation of a nationwide landslide database would solve the shortage of reliable

557   data and allow the usage of more alternative methods, which will result in more improved

558   models. The findings of this research may be used for the elaboration of rainfall-induced debris

559   mitigation policies such as post-disaster rehabilitation planning and land use management.

560   **List of abbreviations**

561   • DALEX: moDel Agnostic Language for Exploration and explanation.

562   • MASS: Modern Applied Statistics with S.

563   • SVM: Support Vector Machine.

564   • POLR: Proportional Odd Logistic Regression.

565   • EGB: Extreme Gradient Boosting.

566   • RF: Random Forest.

567   •   NIR: No information Rate.

568   •   Acc: Accuracy.

569   •   CI: confidence interval.

570   •   OOB: Out Of Bag error estimates.

571   •   CFM: Confusion Matrix.

572   •   Caret: Classification and Regression Training

573 **Acknowledgments:**

574 This research was supported by Basic Science Research Program through the National

575 Research Foundation of Korea (NRF) funded by the Ministry of Education

576 (2021R1A6A1A03044326), and the grant (NRF-2021R1C1C2003316) funded by the National

577 Research Foundation of Korea (NRF).

578 **Data availability:**

579 The datasets used and/or analyzed during the current study are available from the

580 corresponding author on reasonable request.

581 **Declaration of Competing Interest:**

582 The authors declare that there are no conflicts of interest.

583

584

585

586

587

588

References

590 Agresti, A., 2010. Analysis of ordinal categorical data (Vol. 656). John Wiley & Sons.

591 Alexander, D.,1992. On the causes of landslides: Human activities, perception, and natural

592   processes. Environmental Geology and Water Sciences, 20(3), 165-179.

593 Andersen, C. M., Bro, R., 2010. Variable selection in regression a tutorial. Journal of

594   chemometrics, 24(11-12), 728-737.

595 Au, S. W. C., 1998. Rain-induced slope instability in Hong Kong. Engineering Geology, 51(1),

596   1-36.

597 Baum, R. L., Godt, J. W., 2010. Early warning of rainfall-induced shallow landslides and debris

598   flows in the USA. Landslides, 7(3), 259-272.

599 Berti, M., Martina, M. L. V., Franceschini, S., Pignone, S., Simoni, A., Pizziolo M., 2012.

600   Probabilistic rainfall thresholds for landslide occurrence using a Bayesian approach.

601   Journal of Geophysical Research: Earth Surface, 117(F4).

602 Biau, G., Scornet, E., 2016. A random forest-guided tour. Test, 25(2), 197-227.

603 Biecek P., 2018. DALEX: Explainers for Complex Predictive Models in R. Journal of Machine

604   Learning Research, 19(84), 1-5. https://jmlr.org/papers/v19/18-416.html

605 Biecek, P., Burzykowski, T., 2021. Explanatory model analysis: Explore, explain and examine

606   predictive models. Chapman and Hall/CRC.

607 Bilder, C. R., Loughin, T. M., 2014. Analysis of categorical data with R. CRC Press.

608 Brand, E. W., Premchitt, J. E. R. A. S. A. K., Phillipson, H. B., 1984, September. Relationship

609   between rainfall and landslides in Hong Kong. In Proceedings of the 4th International

610   Symposium on Landslides (Vol. 1, No. 01, pp. 276-84). Toronto: Canadian Geotechnical

611   Society.

612 Brant, R., 1990. Assessing proportionality in the proportional odds model for ordinal logistic

613   regression. Biometrics, 46, 1171–1178.

614   Breiman, L., 2001. Random forests. Machine learning, 45(1), 5-32.

615   Brunson, J. C., 2020. ggalluvial: Layered, Grammar for Alluvial Plots. Journal of Open Source

616   Software, 5(49), doi:10.21105/joss.02017. <https://doi.org/10.21105/joss.02017>.

617   Caelen, O., 2017. A Bayesian interpretation of the confusion matrix. Annals of Mathematics

618   and Artificial Intelligence, 81(3), 429-450.

619   Carmela and Mario Parise 2022. A Chronological Database about Natural and Anthropogenic

620   Sinkholes in Italy. Geosciences 12:5, pages 200

621   Causes, L., 2001. Landslide types and processes. US Geological Survey: Reston, VA, USA.

622   Chang, C. W., Lin, P. S., Tsai, C. L., (2011). Estimation of sediment volume of debris flow

623   caused by extreme rainfall in Taiwan. Engineering Geology, 123(1-2), 83-90.

624   Chang, W. J., Chou, S. H., Huang, H. P., Chao, C. Y., (2021). Development and verification of

625   coupled hydro-mechanical analysis for rainfall-induced shallow landslides. Engineering

626   Geology, 293, 106337.

627   Chang, Y. C., Chang, K. H., Wu, G. J., 2018. Application of eXtreme gradient boosting trees in

628   the construction of credit risk assessment models for financial institutions. Applied Soft

629   Computing, 73, 914-920.

630   Charles W. W., Shi, Q., 1998. A numerical investigation of the stability of unsaturated soil

631   slopes subjected to transient seepage. Computers and Geotechnics, 22(1), 1-28.

632   Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano,

633   I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J., 2022. xgboost: Extreme

634   Gradient   Boosting.   R   package   version   1.6.0.1   https://CRAN.R

635   project.org/package=xgboost

636   Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., 2015. Xgboost: an

637   extreme gradient boosting. R package version 0.4-2, 1(4), 1-4.

638   Chen, Y., Withanage, K. R., Uchimura, T., Mao, W., Nie, W., 2020. Shear deformation and

639 failure of unsaturated sandy soils in surface layers of slopes during rainwater infiltration.

640 Measurement, 149, 107001.

641 Chinkulkijniwat, A., Salee, R., Horpibulsuk, S., Arulrajah, A., & Hoy, M. (2022). Landslide

642 rainfall threshold for landslide warning in Northern Thailand. Geomatics, Natural

643 Hazards and Risk, 13(1), 2425-2441.

644 Chough, S. K., Kwon, S. T., Ree, J. H., Choi, D. K., 2000. Tectonic and sedimentary evolution

645 of the Korean peninsula: a review and new view. Earth-Science Reviews, 52(1-3), 175-

646 235.

647 Coppola, L., Reder, A., Tarantino, A., Mannara, G., Pagano, L., 2022. Pre-failure suction-

648 induced deformation to inform early warning of shallow landslides: Proof of concept at

649 slope model scale. Engineering Geology, 106834.

650 Crawford, M. M., Bryson, L. S., Woolery, E. W., Wang, Z., 2019. Long-term landslide

651 monitoring using soil-water relationships and electrical data to estimate suction stress.

652 Engineering Geology, 251, 146-157.

653 Dahal, R. K., Hasegawa, S., Nonomura, A., Yamanaka, M., Masuda, T., Nishino, K., 2008.

654 GIS-based weights-of-evidence modeling of rainfall-induced landslides in small

655 catchments for landslide susceptibility mapping. Environmental Geology, 54(2), 311-

656 324.

657 de Jesús Arce-Mojica, T., Nehren, U., Sudmeier-Rieux, K., Miranda, P. J., Anhuf, D., 2019.

658 Nature-based solutions (NbS) for reducing the risk of shallow landslides: where do we

659 stand? International journal of disaster risk reduction, 41, 101293.

660 DeGraff, J. V.,and Romesburg, H. C. 2020. Regional landslide—susceptibility assessment for

661 wildland management: a matrix approach. In Thresholds in geomorphology (pp. 401-

662 414). Routledge.

663 Di Napoli, M., Carotenuto, F., Cevasco, A., Confuorto, P., Di Martire, D., Firpo, M., ...and

664   Calcaterra, D. 2020. Machine learning ensemble modelling as a tool to improve

665   landslide susceptibility mapping reliability. Landslides, 17(8), 1897-1914.

666 Diaz-Uriarte, R., de Andrés, S. A., 2005. Variable selection from random forests: application

667   to gene expression data. arXiv preprint q-bio/0503025.

668 Ding, C. H., Dubchak, I., 2001. Multi-class protein fold recognition using support vector

669   machines and neural networks. Bioinformatics, 17(4), 349-358.

670 Fox, J., Weisberg, S., 2019. An R companion to applied regression 3rd ed Sage Thousand Oaks.

671 Franzluebbers, A. J., 2002. Water infiltration and soil structure related to organic matter and its

672   stratification with depth. Soil and Tillage research, 66(2), 197-205.

673 Frattini, P., Crosta, G.,and Carrara, A. 2010. Techniques for evaluating the performance of

674   landslide susceptibility models. Engineering geology, 111(1-4), 62-72.

675 Friedman, J. H., 2001. Greedy function approximation: a gradient boosting machine. Annals

676   of Statistics, 1189-1232.

677 Gariano, S. L., Guzzetti, F. 2016. Landslides in a changing climate. Earth-Science Reviews,

678   162, 227-252.

679 Garson, G., 2021. Data Analytics for the Social Sciences: Applications in R.

680   10.4324/9781003109396.

681 Georganos, S., Grippa, T., Vanhuysse, S., Lennert, M., Shimoni, M., Wolff, E., 2018. Very high-

682   resolution object-based land use–land cover urban classification using extreme gradient

683   boosting. IEEE geoscience and remote sensing letters, 15(4), 607-611.

684 Hakim, W. L., Rezaie, F., Nur, A. S., Panahi, M., Khosravi, K., Lee, C. W., Lee, S., 2022.

685   Convolutional neural network (CNN) with metaheuristic optimization algorithms for

686   landslide susceptibility mapping in Icheon, South Korea. Journal of environmental

687   management, 305, 114367.

688 Harrell, F. E., 2015. Ordinal logistic regression. In Regression modeling strategies (pp. 311-

689   325). Springer, Cham.

690   He, Q., Wang, M. and Liu, K.,2021. Rapidly assessing earthquake-induced landslide

691   susceptibility on a global scale using random forest. Geomorphology, 391, 107889.

692   Hidalgo, C. A., Vega, J. A., Parra Obando, M., 2017. Effect of the Rainfall Infiltration Processes

693   on the Landslide Hazard Assessment of Unsaturated Soils in Tropical Mountainous

694   Regions. In T. H. II, & P. Rao (Eds.), Engineering and Mathematical Topics in Rainfall.

695   IntechOpen. https://doi.org/ 10.5772/intechopen.70821.

696   Ho, K. K. S., and Ko, F. W. Y., 2009. Application of quantified risk analysis in landslide risk

697   management practice: Hong Kong experience. Georisk, 3(3), 134-146.

698   Hong, H., Pradhan, B., Xu, C., Bui, D. T., 2015. Spatial prediction of landslide hazard at the

699   Yihuang area (China) using two-class kernel logistic regression, alternating decision tree

700   and support vector machines. Catena, 133, 266-281.

701   Huang, F., Chen, J., Du, Z., Yao, C., Huang, J., Jiang, Q., ..., Li, S., 2020. Landslide

702   susceptibility prediction considering regional soil erosion based on machine-learning

703   models. ISPRS International Journal of Geo-Information, 9(6), 377.

704   Huang, F., Chen, J., Liu, W., Huang, J., Hong, H., and Chen, W., 2022. Regional rainfall-

705   induced landslide hazard warning based on landslide susceptibility mapping and a

706   critical rainfall threshold. Geomorphology, 408, 108236.

707   Huang, G., Zheng, M., Peng, J., 2021. Effect of Vegetation Roots on the Threshold of Slope

708   Instability Induced by Rainfall and Runoff. Geofluids.

709   Huang, Y., Zhao, L., 2018. Review on landslide susceptibility mapping using support vector

710   machines. Catena, 165, 520-529.

711   Ju, N., Huang, J., He, C., Van Asch, T. W. J., Huang, R., Fan, X., ... , Wang, J. (2020). Landslide

712   early warning, case studies from Southwest China. Engineering Geology, 279, 105917.

713   Kabani, A., El-Sakka, M. R., 2016, July. Object detection and localization using deep

714   convolutional networks with softmax activation and multi-class log loss. In International

715   Conference on Image Analysis and Recognition (pp. 358-366). Springer, Cham.

716   Kadavi, P. R., Lee, C. W., Lee, S., 2019. Landslide-susceptibility mapping in Gangwon-do,

717   South Korea, using logistic regression and decision tree models. Environmental Earth

718   Sciences, 78(4), 1-17.

719   Kainthura P. and Sharma N., 2022. Machine learning driven landslide susceptibility prediction

720   for the Uttarkashi region of Uttarakhand in India. Georisk: Assessment and Managem

721   Kim, J. C., Lee, S., Jung, H. S., Lee, S., 2018. Landslide susceptibility mapping using random

722   forest and boosted tree models in Pyeong-Chang, Korea. Geocarto international, 33(9),

723   1000-1015.

724   Kim, J., Lee, K., Jeong, S., & Kim, G., 2014. GIS-based prediction method of landslide

725   susceptibility using a rainfall infiltration-groundwater flow model. Engineering geology,

726   182, 63-78.

727   Klose, M., Maurischat, P., Damm, B., 2016. Landslide impacts in Germany: A historical and

728   socioeconomic perspective. Landslides, 13(1), 183-199.

729   Kockelman, W. J., 1986. Some techniques for reducing landslide hazards. Bulletin of the

730   Association of Engineering Geologists, 23(1), 29-52.

731   Lacroix, P., Handwerger, A. L., Bièvre, G., 2020. Life and death of slow-moving landslides.

732   Nature Reviews Earth & Environment, 1(8), 404-419.

733   Lechner, M., Okasa, G., 2020. orf: Ordered Random Forests. R package version 0.1, 3.

734   Lee, S. G., Winter, M. G., 2019. The effects of debris flow in the Republic of Korea and some

735   issues for successful risk reduction. Engineering Geology, 251, 172-189.

736   Lee, S. W., Kim, G. H., Yune, C. Y., Ryu, H. J., Hong, S. J., 2012. Development of landslide-

737   risk prediction model thorough database construction. Journal of the Korean

738   geotechnical society, 28(4), 23-33.

739   Lee, S. W., Kim, G., Yune, C. Y., Ryu, H. J., 2013. Development of landslide-risk assessment

740   model for mountainous regions in eastern Korea. Disaster advances, 6(6), 70-79.

741   Lee, S., Hong, S. M., Jung, H. S., 2017. A support vector machine for landslide susceptibility

742   mapping in Gangwon Province, Korea. Sustainability, 9(1), 48.

743   Lee, S., Lee, M. J., Jung, H. S., Lee, S., 2020. Landslide susceptibility mapping using Naïve

744   Bayes and Bayesian network models in Umyeonsan, Korea. Geocarto international,

745   35(15), 1665-1679.

746   Li, Z., Liu, P., Wang, W., Xu, C., 2012. Using support vector machine models for crash injury

747   severity analysis. Accident Analysis & Prevention, 45, 478-486.

748   Lin, L., Lin, Q., & Wang, Y. (2017). Landslide susceptibility mapping on a global scale using

749   the method of logistic regression. Natural Hazards and Earth System Sciences, 17(8),

750   1411-1424.

751   Liu, G., Dai E., Xu, X., Wu, W., Xiang A., 2018. Quantitative Assessment of Regional Debris-

752   Flow Risk: A Case Study in Southwest China. Sustainability, 10, 2223.

753   https://doi.org/10.3390/su10072223

754   Liu, Y., Xu, C., Huang, B., Ren, X., Liu, C., Hu, B., Chen, Z., 2020. Landslide displacement

755   prediction based on multi-source data fusion and sensitivity states. Engineering Geology,

756   271, 105608.

757   Liu, Z., Gilbert, G., Cepeda, J. M., Lysdahl, A. O. K., Piciullo, L., Hefre, H., Lacasse, S., 2021.

758   Modeling of shallow landslides with machine learning algorithms. Geoscience Frontiers,

759   12(1), 385-393.

760   Lombardo, L., & Mai, P. M. (2018). Presenting logistic regression-based landslide

761   susceptibility results. Engineering geology, 244, 14-24.

762   Louviere, J. J., Hensher, D. A., Swait, J. D., 2000. Stated choice methods: analysis and

763   applications. Cambridge university press.

764   Marjanović, M., Kovačević, M., Bajat, B., Voženílek, V., 2011. Landslide susceptibility

765   assessment using SVM machine learning algorithm. Engineering Geology, 123(3), 225-

766   234.

767   Martinez-Taboada, F., Redondo, J. I., 2020. Variable importance plot (mean decrease accuracy

768   and mean decrease Gini). Plos One, 15(4), e0230799.

769   Max K., 2022. caret: Classification and Regression Training. R package version 6.0-92.

770   https://CRAN.R-project.org/package=caret.

771   Mayer, R., Plank, C., Bohner, A., Kollarits, S., Corsini, A., Ronchetti, F., ... and Jindra, P., 2008.

772   Monitor: Hazard monitoring for risk assessment and risk communication. Georisk, 2(4),

773   195-222.

774   McColl, S. T., 2022. Landslide causes and triggers. In Landslide Hazards, Risks, and Disasters

775   (pp. 13-41). Elsevier.

776   McCullagh, P., 1980. Regression models for ordinal data. Journal of the Royal Statistical

777   Society: Series B (Methodological), 42(2), 109-127

778   McNulty, K., 2021. Handbook of Regression Modeling in People Analytics: With Examples in

779   R and Python. Chapman and Hall/CRC.

780   Meena, S. R., Ghorbanzadeh, O., van Westen, C. J., Nachappa, T. G., Blaschke, T., Singh, R.

781   P., Sarkar, R., 2021. Rapid mapping of landslides in the Western Ghats (India) triggered

782   by 2018 extreme monsoon rainfall using a deep learning approach. Landslides, 18(5),

783   1937-1950.

784   Meena, S. R., Puliero, S., Bhuyan, K., Floris, M., Catani, F., 2022. Assessing the importance

785   of conditioning factor selection in landslide susceptibility for the province of Belluno

786   (region of Veneto, northeastern Italy). Natural hazards and earth system sciences, 22(4),

787   1395-1417.

788   Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2021. e1071: Misc

789   Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071),

790   TU Wien. R package version 1.7-9, https://CRAN.R-project.org/package=e1071

791   Nandi A.and Shakoor, A., 2008. Application of logistic regression model for slope instability

792   prediction in Cuyahoga River Watershed, Ohio, USA. Georisk, 2(1), 16-27.

793   Negi, H. S., Kumar, A., Rao, N. N., Thakur, N. K., Shekhar, M. S., 2020. Susceptibility

794   assessment of rainfall-induced debris flow zones in Ladakh–Nubra region, Indian

795   Himalaya. Journal of Earth System Science, 129(1), 1-20.

796   NGII, 2018. Digital elevation model, NGII (National Geographical Information Institute), the

797   Ministry of Land, Infrastructure and Transport, Korea.

798   Ngo, P. T. T., Panahi, M., Khosravi, K., Ghorbanzadeh, O., Kariminejad, N., Cerda, A., Lee,

799   S., 2021. Evaluation of deep learning algorithms for national scale landslide

800   susceptibility mapping of Iran. Geoscience Frontiers, 12(2), 505-519.

801   Noble, W. S., 2006. What is a support vector machine? Nature Biotechnology, 24(12), 1565-

802   1567.

803   Ozioko, O. H., Igwe, O., 2020. GIS-based landslide susceptibility mapping using heuristic and

804   bivariate statistical methods for Iva Valley and environs of Southeast Nigeria.

805   Environmental monitoring and assessment, 192(2), 1-19.

806   Panahi, M., Gayen, A., Pourghasemi, H. R., Rezaie, F., Lee, S., 2020. Spatial prediction of

807   landslide susceptibility using hybrid support vector regression (SVR) and the adaptive

808   neuro-fuzzy inference system (ANFIS) with various metaheuristic algorithms. Science

809   of the Total Environment, 741, 139937.

810   Park, S., Choi, C., Kim, B., Kim, J., 2013. Landslide susceptibility mapping using frequency

811   ratio, analytic hierarchy process, logistic regression, and artificial neural network

812   methods at the Inje area, Korea. Environmental earth sciences, 68(5), 1443-1464.

813   Park, S., Kim, J., 2019. Landslide susceptibility mapping based on random forest and boosted

814   regression tree models, and a comparison of their performance. Applied Sciences, 9(5),

815   942.

816   Peruccacci, S., Brunetti, M. T., Gariano, S. L., Melillo, M., Rossi, M., & Guzzetti, F., 2017.

817   Rainfall thresholds for possible landslide occurrence in Italy. Geomorphology, 290, 39-

818   57.

819   Pham, B. T., Tien Bui, D., Prakash, I., 2018. Bagging-based support vector machines for spatial

820   prediction of landslides. Environmental Earth Sciences, 77(4), 1-17.

821   Polemio, M., and Petrucci, O. 2000. Rainfall as a landslide triggering factor an overview of

822   recent international research. Landslides in research, theory and practice.

823   Popescu, M. E. 2002. Landslide causal factors and landslide remediatial options. In 3rd

824   international conference on landslides, slope stability and safety of infra-structures (pp.

825   61-81). CI-Premier PTE LTD Singapore.

826   Rahman, H. A. A., Wah, Y. B., Huat, O. S., 2021. Predictive Performance of Logistic

827   Regression for Imbalanced Data with Categorical Covariate. Pertanika Journal of

828   Science & Technology, 29(1).

829   Rahardjo, H., Nistor, M. M., Gofar, N., Satyanaga, A., Xiaosheng, Q., and Chui Yee, S. I., 2020.

830   Spatial distribution, variation and trend of five-day antecedent rainfall in Singapore.

831   Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards,

832   14(3), 177-191.

833   Raja, N. B., Çiçek, I., Türkoğlu, N., Aydin, O., Kawasaki, A., 2017. Landslide susceptibility

834   mapping of the Sera River Basin using the logistic regression model. Natural Hazards,

835   85(3), 1323-1346.

836   Rengers, F. K., McGuire, L. A., Oakley, N. S., Kean, J. W., Staley, D. M., and Tang, H. ,2020.

837   Landslides after wildfire: Initiation, magnitude, and mobility. Landslides, 17(11), 2631-

838   2641.

839   Rivas, V., Remondo, J., Bonachea, J., & Sánchez-Espeso, J. (2022). Rainfall and weather
840   conditions inducing intense landslide activity in northern Spain (Deba, Guipúzcoa).
841   Physical Geography, 43(4), 419-439.

842   Sameen, M. I., Pradhan, B., Lee, S., 2020. Application of convolutional neural networks
843   featuring Bayesian optimization for landslide susceptibility assessment. Catena, 186,
844   104249.

845   Sarkar, R., Dikshit, A., Hazarika, H., Yamada, K., Subba, K., 2019. Probabilistic rainfall
846   thresholds for landslide occurrences in Bhutan. International Journal of Recent
847   Technology and Engineering.

848   Sarkar, R., Dorji, K., 2019. Determination of the probabilities of landslide events—a case study
849   of Bhutan. Hydrology, 6(2), 52.

850   Segoni, S., Piciullo, L., Gariano, S. L., 2018. A review of the recent literature on rainfall
851   thresholds for landslide occurrence. Landslides, 15(8), 1483-1501

852   Shahabi, H., Hashim, M., 2015. Landslide susceptibility mapping using GIS-based statistical
853   models and Remote sensing data in the tropical environment. Scientific reports, 5(1), 1-
854   15.

855   Šilhán, K.; and Stoffel, M., 2015. Impacts of age-dependent tree sensitivity and dating
856   approaches on dendrogeomorphic time series of landslides. Geomorphology, 236, 34-
857   43.

858   Su, C., Wang, B., Lv, Y., Zhang, M., Peng, D., Bate, B.,and Zhang, S. 2022. Improved landslide
859   susceptibility mapping using unsupervised and supervised collaborative machine
860   learning models. Georisk: Assessment and Management of Risk for Engineered Systems
861   and Geohazards, 1-19.

862   Sun, D., Xu, J., Wen, H., and Wang, D., 2021. Assessment of landslide susceptibility mapping
863   based on Bayesian hyperparameter optimization: A comparison between logistic

864   regression and random forest. Engineering Geology, 281, 105972.

865   Sun, D., Wen, H., Wang, D., and Xu, J. 2020. A random forest model of landslide susceptibility

866   mapping based on hyperparameter optimization using Bayes algorithm. Geomorphology,

867   362, 107201.

868   Susmaga, R., 2004. Confusion matrix visualization. In Intelligent information processing and

869   web mining (pp. 107-116). Springer, Berlin, Heidelberg.

870   Takara, K., Yamashiki, Y., Sassa, K., Ibrahim, A. B., & Fukuoka, H., 2010. A distributed

871   hydrological–geotechnical model using satellite-derived rainfall estimates for shallow

872   landslide prediction system at a catchment scale. Landslides, 7(3), 237-258.

873   Tang, D., Li, D. Q., and Cao, Z. J., 2017. Slope stability analysis in the Three Gorges Reservoir

874   Area considering effect of antecedent rainfall. Georisk: Assessment and Management of

875   Risk for Engineered Systems and Geohazards, 11(2), 161-172.

876   Taylor, F. E., Tarolli, P., Malamud, B. D., 2020. Preface: Landslide–transport network

877   interactions. Natural Hazards and Earth System Sciences, 20(10), 2585-2590.

878   Team, R. C., 2021. R: A language and environment for statistical computing. R Foundation for

879   Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

880   Tiwari, B., Ajmera, B., Gonzalez, A., and Sonbol, H. 2020. Impact of wildfire on triggering

881   mudslides—a case study of 2018 Montecito debris flows. In Geo-Congress 2020:

882   Engineering, Monitoring, and Management of Geotechnical Infrastructure (pp. 40-49).

883   Reston, VA: American Society of Civil Engineers.

884   Turner, A. K., 2018. Social and environmental impacts of landslides. Innovative Infrastructure

885   Solutions, 3(1), 1-25.

886   Vahedifard, F., Sehat, S., and Aanstoos, J. V. 2017. Effects of rainfall, geomorphological and

887   geometrical variables on vulnerability of the lower Mississippi River levee system to

888   slump slides. Georisk: Assessment and Management of Risk for Engineered Systems

889   and Geohazards, 11(3), 257-271.

890   Van der Beek, P., 2021. Stressed rocks cause big landslides. Nature Geoscience, 14(5), 261-

891   262.

892   Van Tien, P., Luong, L. H., Duc, D. M., Trinh, P. T., Quynh, D. T., Lan, N. C.,..., Loi, D. H.,

893   2021. Rainfall-induced catastrophic landslide in Quang Tri Province: the deadliest single

894   landslide event in Vietnam in 2020.

895   Wang, X., Xiao, Y., Shi, W., Ren, J., Liang, F., Lu, J., ..., Yu, X., 2022b. Forensic analysis and

896   numerical simulation of a catastrophic landslide of dissolved and fractured rock slope

897   subject to underground mining. Landslides, 19(5), 1045-1067.

898   Wang, Y., Tang, H., Huang, J., Wen, T., Ma, J., Zhang, J., 2022a. A comparative study of

899   different machine learning methods for reservoir landslide displacement prediction.

900   Engineering Geology, 298, 106544.

901   Wickham, H., François, R., Henry, L., Müller, K., 2022. dplyr: A Grammar of Data

902   Manipulation. R package version 1.0.9, https://CRAN.R-project.org/package=dplyr

903   Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

904   Winter, M. G., 2020. Debris flows. Geological Society, London, Engineering Geology Special

905   Publications, 29(1), 163-185.

906   Woo, C., Kwon, H., Lee, C., Kim, K., 2014. Landslide hazard prediction map based on logistic

907   regression model for applying in the whole country of South Korea. Journal of the

908   Korean Society of Hazard Mitigation, 14(6), 117-123.

909   Yesilnacar, E., and Topal, T. A. M. E. R., 2005. Landslide susceptibility mapping: a comparison

910   of logistic regression and neural networks methods in a medium scale study, Hendek

911   region (Turkey). Engineering Geology, 79(3-4), 251-266.

912   Yilmaz, I., 2009. Landslide susceptibility mapping using frequency ratio, logistic regression,

913   artificial neural networks and their comparison: a case study from Kat landslides

914   (Tokat—Turkey). Computers & Geosciences, 35(6), 1125-1138.

915   Zeng, H., Tang, C. S., Zhu, C., Vahedifard, F., Cheng, Q., Shi, B., 2022. Desiccation cracking

916   of soil subjected to different environmental relative humidity conditions. Engineering

917   Geology, 297, 106536.

918   Zhao, Y., Wang, R., Jiang, Y., Liu, H., Wei, Z., 2019. GIS-based logistic regression for rainfall-

919   induced landslide susceptibility mapping under different grid sizes in Yueqing,

920   Southeastern China. Engineering Geology, 259, 105147.

921   Zhu, H., and Zhang, L. 2019. Root-soil-water hydrological interaction and its impact on slope

922   stability. Georisk: Assessment and Management of Risk for Engineered Systems and

923   Geohazards, 13(4), 349-359.

924