International Journal on Advanced Science Engineering Information Technology

Vol.11 (2021) No. 1
ISSN: 2088-5334

# High-Resolution Landslide Susceptibility Map Generation using Machine Learning (Case Study in Pacitan, Indonesia)

Mohammad Rohmaneo Darminto[a,*], Amien Widodo[b], Adillah Alfatinah[c], Hone-Jay Chu[c]

[a]*Department of Geomatics Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, 60117, Indonesia*
[b]*Department of Geophysics Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, 60117, Indonesia*
[c]*Department of Geomatics Engineering, National Cheng Kung University, No.1, Daxue Rd., East District., Tainan, 701, Taiwan*
*Corresponding author: *rohmaneo@its.ac.id*

***Abstract*** **— Landslide, one of the most disastrous natural hazards, causes damage to infrastructure worldwide and local communities. Pacitan, Indonesia is one city with high susceptibility to landslides occurrence. The conditions of landslide occurrence are assumed to be the same in the future. This study's objective is to produce a landslide susceptibility map by using machine learning methods based on topographical factors including elevation, slope, aspect, profile curvature, plan curvature, Topographic Wetness Index (TWI), distance to the river, and geological map as independent variables, whereas the landslide inventory map derived from Sentinel-2A and Landsat 7 were used as the dependent variables in the model construction. This study's datasets were constructed in three different compositions where each composition was treated as input in Random Forest, Decision Tree, and Logistic regression model. The first dataset was composed of a 70:30 ratio for training and testing sample points, the second dataset with a 60:40 ratio, and the third with a 50:50 ratio. The performance of each model using each dataset composition was analyzed using various accuracy assessments. This study also considered each topographical factor's effect on model performance by excluding several factors in model construction. From the results, random forest with the first dataset appeared to give the best performance for mapping landslide susceptibility area, shown by the highest Area Under Curve (AUC) value, Coefficient Correlation (CC), and Cohen's Kappa of 0.96, 0.92 (92%) and 0.84, respectively. Elevation and geological maps were considered as essential variables shown by significant drops in model accuracy assessment when these two factors were separately excluded, while profile curvature was the least essential variable based on the insignificant drop in the model accuracy assessment result.**

***Keywords***— **Landslide; landslide susceptibility map; random forest; decision tree; logistic regression; machine learning.**

*Manuscript received 14 Apr. 2020; revised 5 Oct. 2020; accepted 26 Nov. 2020. Date of publication 28 Feb. 2021.*


## I. INTRODUCTION

Landslides are considered as one of the most disastrous and very complex natural hazards that cause severe loss of human life and property worldwide [1]. This kind of mass movement occurs mainly in certain geological formations, steep and rugged land surfaces, and extreme climatic conditions result in a high degree of instability. Certain hydrological processes and elevation patterns due to altitude changes lead to massive differences in environmental characteristics in mountainous topography areas. Additionally, human activities like road construction or deforestation can contribute to this hazard [2], [3]. During the last several decades, many government agencies attempted to find the most appropriate solution to minimize the damage caused directly and indirectly by landslides. One of the attempts is generally performed by identifying and mapping areas susceptible to landslides. These maps are generated based on an assessment of landslide susceptibility, a spatial distribution of probabilities of landslide occurrences in a given area based on local geo-environmental factors [4], [5]. Thus, an accurate susceptibility mapping is needed as key information for many users from both private and public sectors, from governmental departments, and the scientific community on both local and international levels [6]. Various methods and techniques have been proposed to evaluate landslide susceptibility. The statistical approaches have become popular in the use of remote sensing (RS) with a geographic information system (GIS) [7]. There are many statistical approaches used in landslide susceptibility assessment, including a frequency ratio (FR) [8], [9], statistical index (SI) [10], as well as logistic regression (LR) [11], [12]. Furthermore, the approaches using machine learning techniques have become popular recently. The increasing use of machine learning method was due to

robustness and high generalization capability for landslide susceptibility analysis [7]. Machine learning methods including artificial neural network [13], fuzzy logic [14], [15], support vector machine [11], [16], random forest [7], [17], and decision tree [18]–[20] methods have been popularly applied among the others. The present study proposed the new concept to take landslide release area – as a source of landslide occurrence – into account to accurately produce a landslide release susceptibility map by considering nine topographical factors as landslide factors. This study executed using three methods: logistic regression (LR), random forest (RF), and decision tree (DT). The models' results were compared using the receiver operating characteristic (ROC) curve and statistical indices to determine the results' model accuracy. Hence, the landslide susceptibility model can show areas where landslides are more likely to be generated in the future, and it can further provide valuable information for urban planning and landslide mitigation management and prevention.

## II. Materials and Method

### A. Study Area

This research was conducted in Pacitan region, the southwest part of East Java province, Indonesia (shown in Fig. 1).

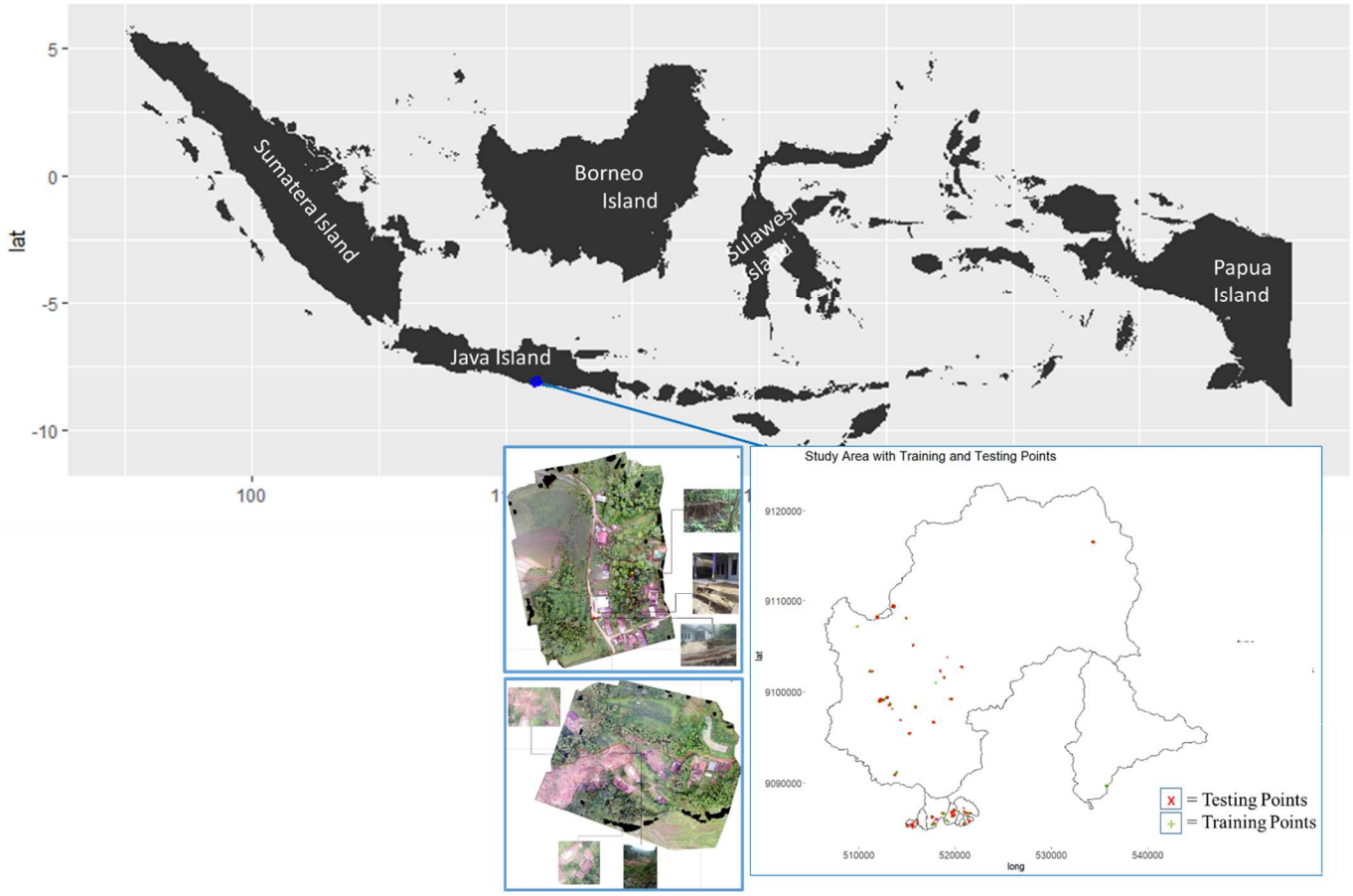


Fig. 1 Testing and training points in study area

This regency is located at coordinates 110.55° - 111.25° E and 7.55° - 8.17° S and has an area of 1.419,44 km² with an altitude 0-1,500 meters above sea level. The administration of Pacitan includes 12 subdistricts: Pacitan, Kebonagung, Arjosari, Tulakan, Ngadirojo, Punung, Pringkuku, Donorojo, Nawangan, Tegalombo, Sudimoro, and Bandar. Pacitan is a hilly region with high and steep topography; it is dominated by mountainous and rocky, while only a few places form the plains. Pacitan region has a tropical climate that experiences two seasons: dry (April-October) and rainy (October-April). The average monthly rainfall is around 3-503mm, with the highest occurring in December (503mm) and the lowest (3mm) occurring in June. This rainwater flows through three major rivers in the Pacitan region: Grindulu River - Gunungsari, Lorok River – Wonodadi, and Kedungpring River – Nawangan. These rivers become the source of irrigation for agricultural land with rice and horticultural production, considering 9.36% or around 130.15 km$^2$ of this region is used for agricultural purposes. The temperature in Pacitan is relatively stable throughout the year, with the lowest average temperature being 24°C and the highest average temperature being 26°C. The Pacitan region is known to be prone to landslides. There are varying causes for landslides, but they are mostly caused by massive rainfall during the rainy season. A major landslide event occurred in 2017 and was affected by the formation of the Cempaka Tropical Cyclone (STC) in the South Waters of Java Island (shown in Fig. 2) [21]. Tropical cyclones are hydrometeorological disasters that rarely occur in Indonesia because of their geographic position located around the equator, but some cyclones form around

Indonesian waters and impact weather conditions in Indonesia [22]. This condition caused heavy rains that triggered floods and landslides, and the Pacitan region was one of the most impacted areas during this occurrence.

Fig. 2 Landslide caused by Tropical Cyclone in the Pacitan region

### *B. Data and Research Workflow*

Data and processing workflow for this study is shown in Table 1 and Fig. 3, and a thorough explanation has been given in the following section.

TABLE I
LIST AND DESCRIPTION OF DATA SETS INCLUDED IN THE STUDY, ALONG WITH VARIABLES DERIVED FROM EACH SOURCE

| Data Set | Description | Derived Variables |
|---|---|---|
| *High Spatial Accuracy of DEM data* | Provided by Indonesia Geospatial Information Agency (BIG). It has 0.27-arcsecond (equivalent with 8.1 m) resolution and used EGM2008 as the vertical datum. | - Elevation<br>- Slope<br>- Aspect<br>- Plan Curvature<br>- Profile Curvature<br>- Ruggedness<br>- Topographic Wetness Index (TWI) |
| *Optical Satellite Data* | Sentinel-2A and Landsat 7 TM from 2016 until the beginning of 2020. | Normalized Difference Vegetation Index (NDVI) |
| *Indonesia Topographical map* | Topographical map of Indonesia was provided by Indonesia Geospatial Information Agency (BIG) | Distance to the river |
| *Geological Map* | Geological map has a 1:100,000 scale map and was provided by Indonesia Geological Research and Development Centre | Lithological information |

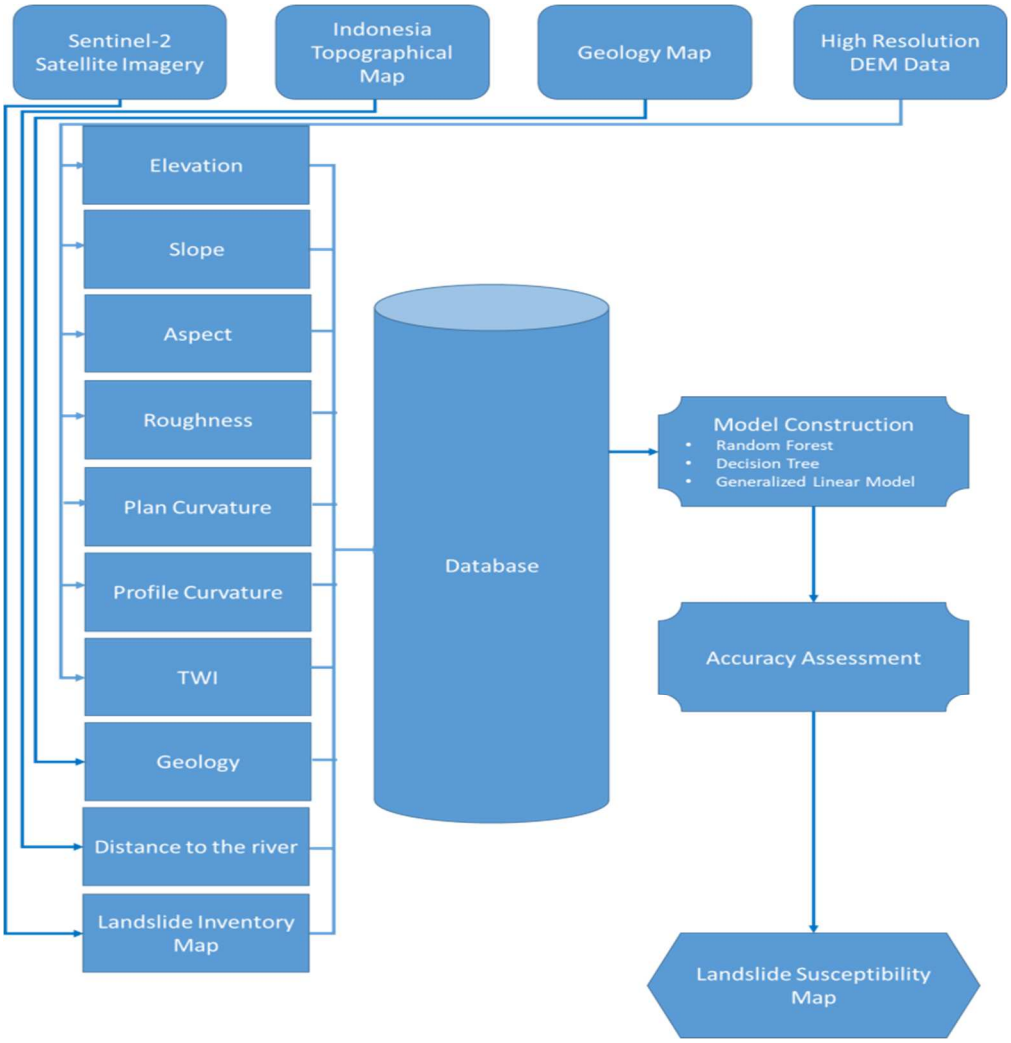


Fig. 3 Workflow of data processing

Mapping of landslide hazard was conducted by using data related to landslide triggering factors as follow:

*1) Landslide Inventory Map:* Landslides inventory in this study was collected using Sentinel-2A and Landsat 7 from 2015 to late 2019. A landslide inventory map aims to record the location and, where known, the date of occurrence and the types of mass movements that have left discernable traces in an area [23], [24]. The landslide inventory map can be generated using various methods. One of the methods is by interpreting and analyzing satellite imagery. Landslides affected the change in land cover and modified the optical properties of the land surface. Satellite sensors can measure the variations in the spectral signature of the land surface, and the images captured by satellite sensors can be used to detect and map landslides [25]. Normalized Difference Vegetation Index (NDVI) is a widely used vegetation indices to analyze land cover changes. Vegetation indices are mathematical transformations designed to assess vegetation's spectral contribution to multispectral observations [26]. NDVI calculated the intense chlorophyll pigment absorptions in the red band against the high reflectivity of plant materials in the NIR. The equation of NDVI has been described as follows:

$$NDVI = \frac{\rho_{nir} - \rho_{red}}{\rho_{nir} + \rho_{red}} \quad (1)$$

with $\rho_{nir}$ meaning Near-Infrared band and $\rho_{red}$ meaning Red band from the sensor. NDVI resulted from 0 to 1 for the vegetation indices. The area with no green vegetation could give a value close to zero, and close to positive 1 (0.8 - 0.9) indicates the highest possible density of green leaves. The low value of NDVI is usually related to the greater probability of landslide occurrence as it shows low vegetation cover [26]. Furthermore, the result from NDVI was compared with the field survey data as the ground truth to generate an accurate landslide inventory map. The inventory map was then be carried out as the response variable in the model prediction, with landslide occurred area identified as 1 and area with no landslide occurred as 0.

*2) Topographic factors:* Digital Elevation Model with resolution 8m x 8m was used to derive 8 from 10 topographic factors in this study. The topographic factors including elevation, slope, aspect slope, plan curvature, profile curvature, ruggedness, and topographic wetness index (TWI) (shown in Fig. 4). Elevation was considered as one of the factors because it is more likely that higher elevation land has a steeper slope increasing the possibility of landslides. The slope is known as the first derivative of the elevation which is calculated to quantify variation in elevation over a distance.

Slope is a crucial indicator of a DEM for specific applications, including landslide feasibility [27]. The slope aspect was calculated according to the eight-neighborhood method. The horizontal and vertical deltas were determinate using the values of the center cell and its eight neighbors. The letters from *a* to *i* were used to identify the neighbors, with *e* representing the calculated target cell. The equations used in measuring the slope angle and aspect, in this case, are as follows:

$$Aspect = \left(\frac{180}{pi}\right) \times ATAN2\left(\left[\frac{dz}{dy}\right], -\left[\frac{dz}{dx}\right]\right) \quad (2)$$

Where,

$$\left[\frac{dz}{dx}\right] = \frac{((c+2f+i)-(a+2d+g))}{8} \quad (3)$$

$$\left[\frac{dz}{dy}\right] = \frac{((g+2h+i)-(a+2b+c))}{8} \quad (4)$$

From the equations above, the aspect value has been obtained. Slope Aspect or slope direction was used to identify the downslope direction from the maximum change rate in value from each cell to its neighbors [28]. Curvature is the second derivative of elevation and measure of change of slope between two points. In order to obtain a curvature parameter, this study computed the existing slopes between two elevation surface points. This study applied two kinds of curvature as the topographic factors: profile and plan curvature. Profile curvature has a parallel direction to the maximum slope and indicates the direction of it. This variable affects the acceleration and deceleration of flow across the surface. Meanwhile, Planform curvature (commonly called plan curvature) is horizontal to the maximum slope's direction and affects the convergence and divergence of flow across the surface [29]. Terrain Ruggedness Index (TRI) is a proxy to quantify the difference between flat and mountainous landscapes [30]. It provides a rapid, objective measure of terrain heterogeneity. TRI was calculated using "DOCELL" command in ArcGIS, which calculates the sum change in elevation between a grid cell and its eight-neighbor grid cell. The last topographic factor, TWI, is commonly used to quantify topographic control on hydrological processes [31]. TWI can examine the pattern of potential soil moisture in the field and can also detect the changes in soil texture caused by an erosion process. The equation to compute the topographic wetness index is as follows:

$$TWI = \ln\left(\frac{A}{tan\beta}\right) \quad (5)$$

Where $A$ is the specific catchment's area ($m^2/m$) and $\beta$ is slope gradient (in degrees) [32]. Aside from the topography index obtained from the DEM, there were two other topography factors: distance to rivers and geology condition. Distance to rivers was created by digitizing the topographical map. This parameter was used to evaluate runoff's role and the influence of toe erosion by stream on landslide triggering [33]. Meanwhile, the geological condition was obtained from the geological map. This parameter is the main factor influences the development of landslide [34]. These 10 topographical factors were later executed as explanatory variables in the prediction models. All the topographical factor raster data have 8m x 8m resolution and are shown in Fig 4.

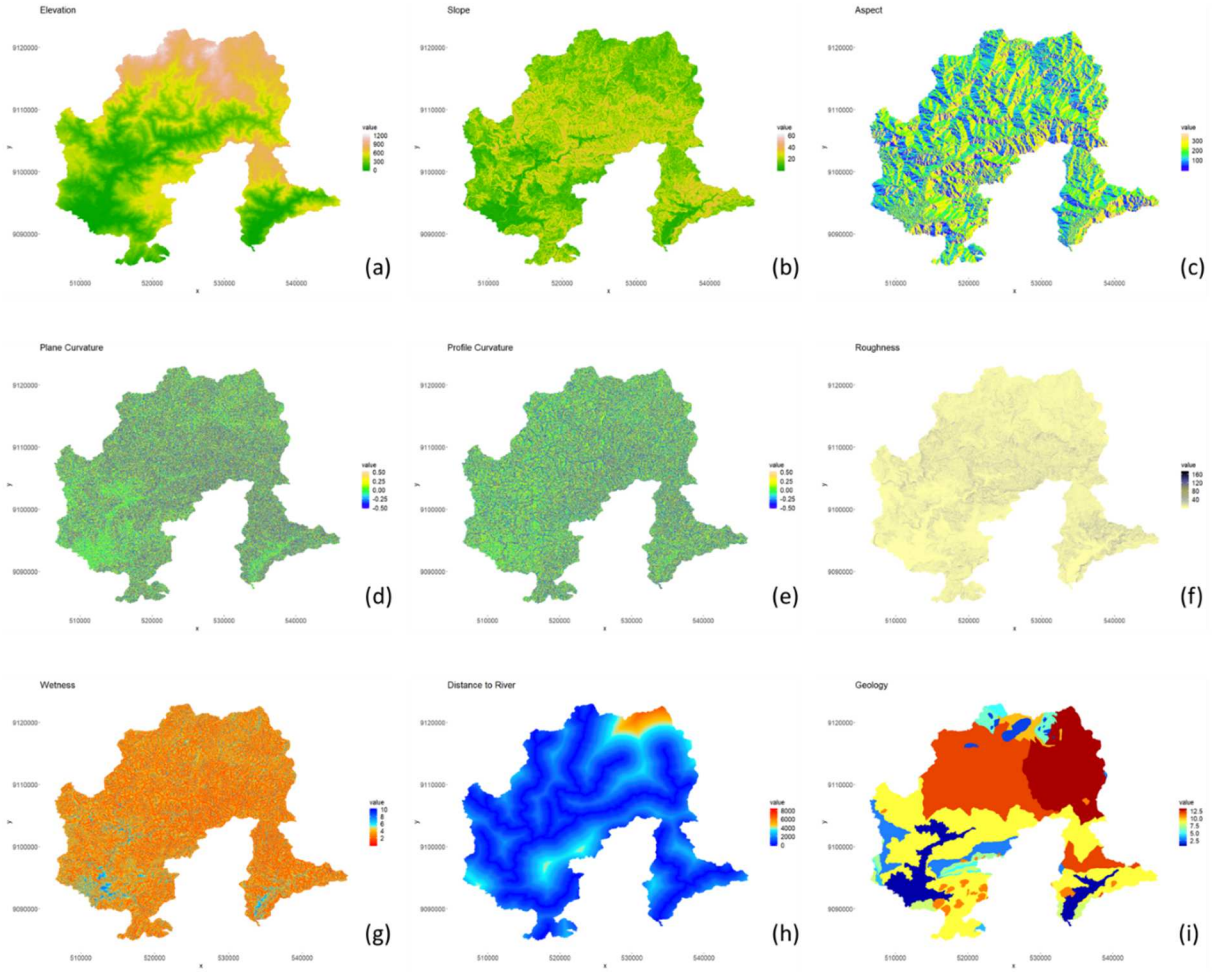


Fig. 4 Topography factors for landslide susceptibility prediction; (a) elevation; (b) slope; (c) aspect; (d) plane curvature; (e) profile curvature; (f) roughness; (g) TWI; (h) distance to the river; and (i) geology

*3) Training and Testing Dataset Preparation:* Furthermore, all the prepared topographical factors raster and landslide inventory maps were bundled together before generating several random points to extract the values. Considering the possibility of locating random points in an area with null value or overlapped with other points in the same cell size, the generation of random points needed to be filtered out to evaluate all those possibilities. After the extraction, each point contained ten topographical values and the information from the landslide inventory map. The point's dataset was, then split into training and testing following the scenario of each composition.

### *C. Model Construction*

This study applied the generated dataset to construct landslide susceptibility models with landslide information as the dependent variable and the topographical factors as the explanatory variables. Models were constructed with three machine learning techniques including, RF, DT, and Generalized Linear Model (GLM) from logistic regression. The effect of data composition is also being considered in this study. Assuming the different ratios of data partition affect the model performance, this study determined three datasets based on different training and testing data composition. The first dataset comprises a 70:30 ratio for the Training and Testing sample points, whereas the second dataset is composed of a 60:40 ratio, and the third with a 50:50 ratio.

*1) Random Forest:* RF is the popular ensemble learning method developed by Breiman [35], widely used for classification, regression, clustering, and interaction detection [7]. This method generates thousands of random binary trees to form a forest so that each tree depends on the values of randomly chosen vectors distributed evenly among all trees in the forest. Each tree is grown based on a bootstrap sample using a classification and regression trees (CART) procedure with a random subset of variables selected at each node [36]. The "out-of-bag" (OOB) error rate is calculated using observations left out of the bootstrap sample for each tree grown on a bootstrap sample. The majority vote determines the final decisions of class and model construction among all trees.

*2) Decision Tree:* DT is a data mining technique for solving classification and prediction problems. Data mining consists of different methods and algorithms used for discovering the knowledge of large data sets [37]. This technique can find and describe structural patterns in data as a structural tree. It does not require advanced knowledge of the relationship between all the input variables and an objective variable. DT is used for solving classification as well as regression problems. When a DT is used for

classification tasks, it is most referred to as a classification tree, and when it is used for regression tasks, it is called a regression tree. DT has a simple hierarchical structure that is easy to understand, and it consists of nodes and leaves. The node includes testing a particular attribute and the leaf denotes a class. DT classifies instances by sorting them down the tree from the root to some leaf node, which gives a classification that applies to all instances that reach the leaf. The tree complexity is measured by one of the following metrics: the total number of nodes, total number of leaves, tree depth, and number of attributes used [38]. DT was executed using "rpart" package in R software.

*3) Logistic Regression:* The logistic regression model is applied to establish the relationship between dependent and independent variables [39]. Logistic regression looks for the most suitable equation as in linear regression while using a different method. It used the maximum likelihood method instead the least-square method as in the linear regression [40]. For this study, the Generalized Linear Model (GLM) is chosen as the logistic regression. GLM is a means of modeling the relationship between a variable whose outcome we wish to predict and one or more explanatory variables. The predicted variable is called the target variable and is denoted by $y$ [41]. The relationship between $\mu i$ (the model prediction) and the predictors is as follows:

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \quad (6)$$

Equation 7 states that some specified transformation of $\mu i$ (denoted $g(\mu i)$) is equal to the intercept (denoted $\beta 0$) plus a linear combination of the predictors and the coefficients which are denoted by $\beta 1 \ldots \beta p$. By applying this concept, environmental data acted as predictor variables that calculate the value of $g(\mu i)$ as the model prediction. In this study, GLM proceeds Program-R using 'glm' package.

### *D. Accuracy Assessment*

The evaluation of the model's performances required the derivation of matrices of confusion that identified true positive (a), false positive (b), false negative (c), and true negative (d) cases predicted by each model. The basic concept of a confusion matrix is displayed in Table 2.

TABLE II
CONFUSION MATRIX CONCEPT

| | | Actual | |
|---|---|---|---|
| Predicted | | Positive (+) | Negative (-) |
| | Positive (+) | True Positive | False Negative |
| | Negative (-) | False Positive | True Negative |

The confusion matrix is needed to calculate most of the measures of classification accuracy from the prediction model. Therefore, this study calculated alternative performance measures including overall prediction success (CC), Cohen's Kappa, and the area under the receiving operation characteristic curve (AUC). The CC is the ratio of samples correctly classified by the prediction model [42]. Cohen's Kappa is a statistic that measures the agreement of two categorical items [43]. Calculation of Cohen's Kappa may be performed according to the following equation:

$$Pr_a = \frac{agree_{yes} + agree_{no}}{event_{total}} \quad (7)$$

$$Pr_e = \left[\frac{Pred_{yes_{total}}}{event_{total}} \times \frac{Actual_{yes_{total}}}{event_{total}}\right] + \left[\frac{Pred_{no_{total}}}{event_{total}} \times \frac{Actual_{no_{total}}}{event_{total}}\right] \quad (8)$$

$$\kappa = \frac{Pr_a - Pr_e}{1 - Pr_e} \quad (9)$$

where $Pr_a$ represents the observed agreement and $Pr_e$ represents chance agreement. A previous study conducted by McKenna and Castiglione (2014) used the categorization presented in Table 3 to assess the significance of Cohen's Kappa values [44].

TABLE III
COHEN'S KAPPA VALUE CATEGORIZATION

| Cohen's Kappa Value | Agreement Categorization |
|---|---|
| < 0.01 | No agreement |
| 0.01 – 0.20 | Slight agreement |
| 0.21 – 0.40 | Fair agreement |
| 0.41 – 0.60 | Moderate agreement |
| 0.61 – 0.80 | Substantial agreement |
| > 0.80 | Almost perfect Agreement |

The Receiver Operating Characteristic (ROC) curve is a graphical method representing the relation between the False Positive fraction and the sensitivity for a range of thresholds. This method has been widely used to measure the performance of prediction models. This curve is obtained by plotting all combinations of True Positive Rate (TPR) and proportions of False Positive Rate (FPR), which may be obtained by varying the decision threshold. Correlation of predictive capability and AUC could be quantified as follows: excellent (0.9–1), very good (0.8–0.9), good (0.7–0.8), average (0.6–0.7), and poor (0.5–0.6) [45]

## III. RESULT AND DISCUSSION

### *A. Model Validation and Comparison*

Each model's performance was analyzed using the Coefficient Correlation, Cohen's Kappa, and Area Under Curve assessments. The following Table 4 shows the comparison of each model validation result.

TABLE IV
PREDICTION MODEL VALIDATION RESULTS

| Model | First Dataset | | | Second Dataset | | | Third Dataset | | |
|---|---|---|---|---|---|---|---|---|---|
| | CC | CK | AUC | CC | CK | AUC | CC | CK | AUC |
| RF | 0.92 | 0.84 | 0.96 | 0.89 | 0.8 | 0.95 | 0.87 | 0.75 | 0.95 |
| DT | 0.81 | 0.61 | 0.86 | 0.79 | 0.58 | 0.83 | 0.8 | 0.61 | 0.83 |
| GLM | 0.91 | 0.82 | 0.85 | 0.88 | 0.79 | 0.83 | 0.85 | 0.77 | 0.82 |

RF using the first dataset showed the best performance compare to other datasets. It achieved 0.96 for its AUC, 0.92 (92%) for CC, and 0.84 for its Kappa value. DT achieved the lowest performance among other models. The least one was shown in the second dataset with 0.79 (79%), 0.58, and 0.83 for the CC, kappa, and AUC values. The ROC of each model has been displayed in Fig. 5.

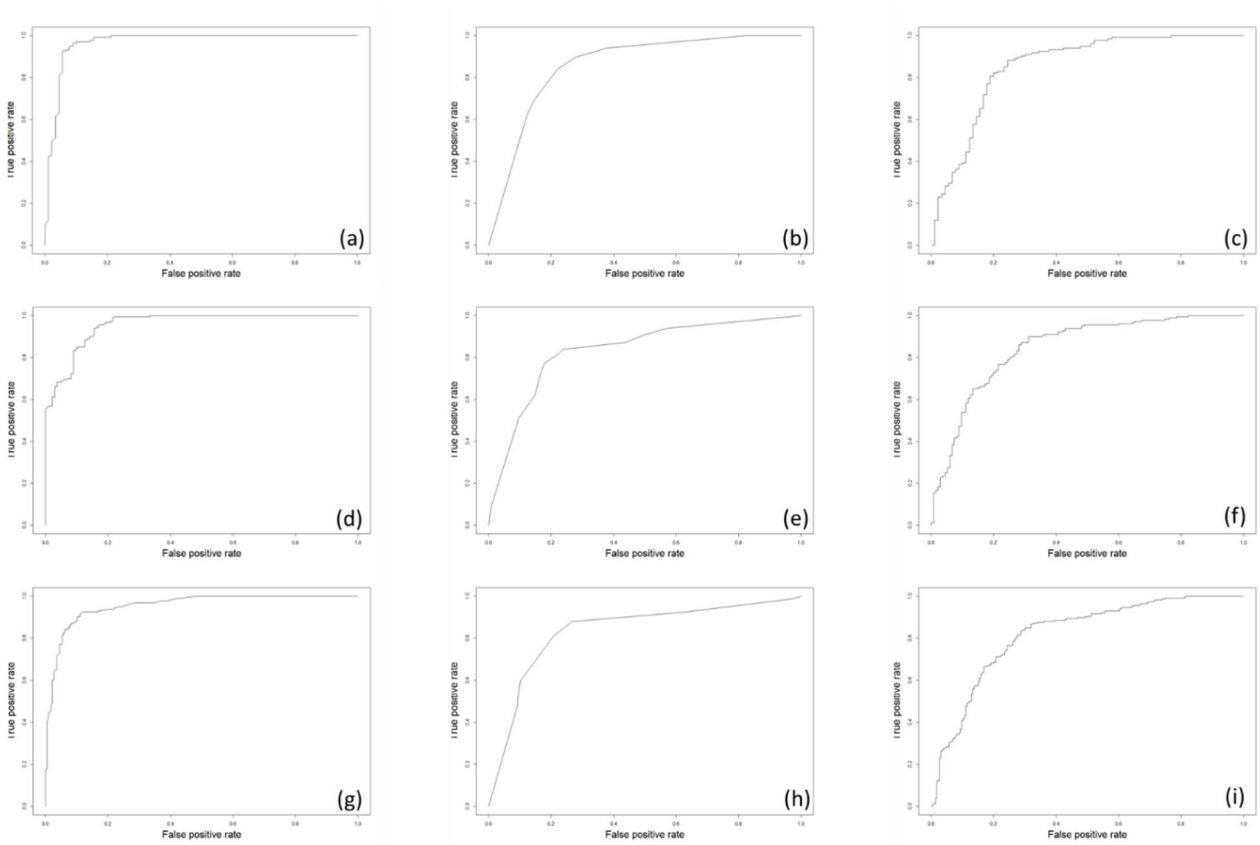


Fig. 5 Region of Curve (ROC) of the first dataset: (a) random forest, (b) decision tree, (c) glm; the second dataset: (d) random forest, (e) decision tree, **(f) glm;** and the third dataset: (g) random forest, (h) decision tree, and (i) glm.

These figures showed that RF has almost perfect results in all datasets, shown by the AUC values present in Table 4, where it achieved over 0.9. DT and GLM appeared to have similar results, where DT has a smoother curve compare to GLM (shown in Fig. 5). The overall results of AUC and CC of all models were over 0.8, which demonstrates good performance in producing landslide susceptibility map. Kappa value indicated the agreement between classes in the prediction model and the reliability of data collection.

Overall, RF and GLM achieved over 0.75 for its kappa value, which indicates less than 25% of the analyzed data are erroneous and show almost perfect agreement in all its classes. Meanwhile, DT achieved lower kappa value in all datasets compared to the other two methods with the lowest value of 0.58 using the second dataset. Based on the results, the RF model using the first dataset shows the best performance.

Furthermore, the effect of topographical factors to the models was also analyzed. These factors greatly impact the landslide susceptibility mapping [7]. Each factor may not make an equal contribution, and it can affect the model prediction result differently. The importance of topographical factors was analyzed using the best performance model of RF using the first dataset. Variable's importance for the model is shown in Fig. 6. The elevation is considered the most important variable for landslide occurrence, while profile curvature was the least critical. To provide deeper analysis, elevation, geological map, distance to the river, topographical wetness index, slope, and profile curvature were separated into six model constructions.

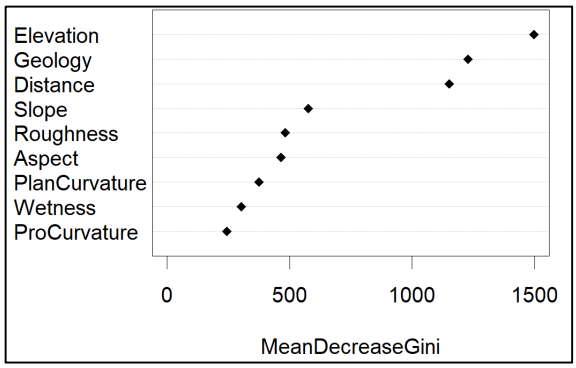


Fig. 6 Variables importance in RF model

The RF model's accuracy value shown in Table 5 significantly dropped as elevation was excluded from the topographical factors.

TABLE V
ACCURACY FROM MODEL WITH VARIABLE DEDUCTION

| Model | Accuracy Assessment | | |
|---|---|---|---|
| | CC | CK | AUC |
| RF first dataset | 0.92 | 0.84 | 0.96 |
| RF w/o Elevation | 0.88 | 0.74 | 0.94 |
| RF w/o Geology | 0.89 | 0.79 | 0.96 |
| RF w/o Distance | 0.89 | 0.79 | 0.95 |
| RF w/o Wetness | 0.91 | 0.82 | 0.96 |
| RF w/o Slope | 0.91 | 0.82 | 0.97 |
| RF w/o Profile Curvature | 0.92 | 0.84 | 0.97 |

It showed 0.88 of Coefficient Correlation, 0.74 Cohen's Kappa, and 0.94 AUC curve. The accuracy drops also occurred when distance and wetness were excluded from the model as shown in Table 5. The insignificant drops of accuracy occurred as slope and wetness were excluded, where the accuracy merely decrease by 0.01 in Coefficient Correlation, and no changes of accuracy occurred when profile curvature was excluded from the model. The resulting accuracy and kappa value were similar with the original model, and the AUC value was even higher when profile curvature was eliminated. These results showed the effects of each topographical factors employed in this study to landslide susceptibility modeling.

### *B. Land Susceptibility Map*

The prediction models were used to generate landslide susceptibility maps in the study area as shown in Fig. 7.

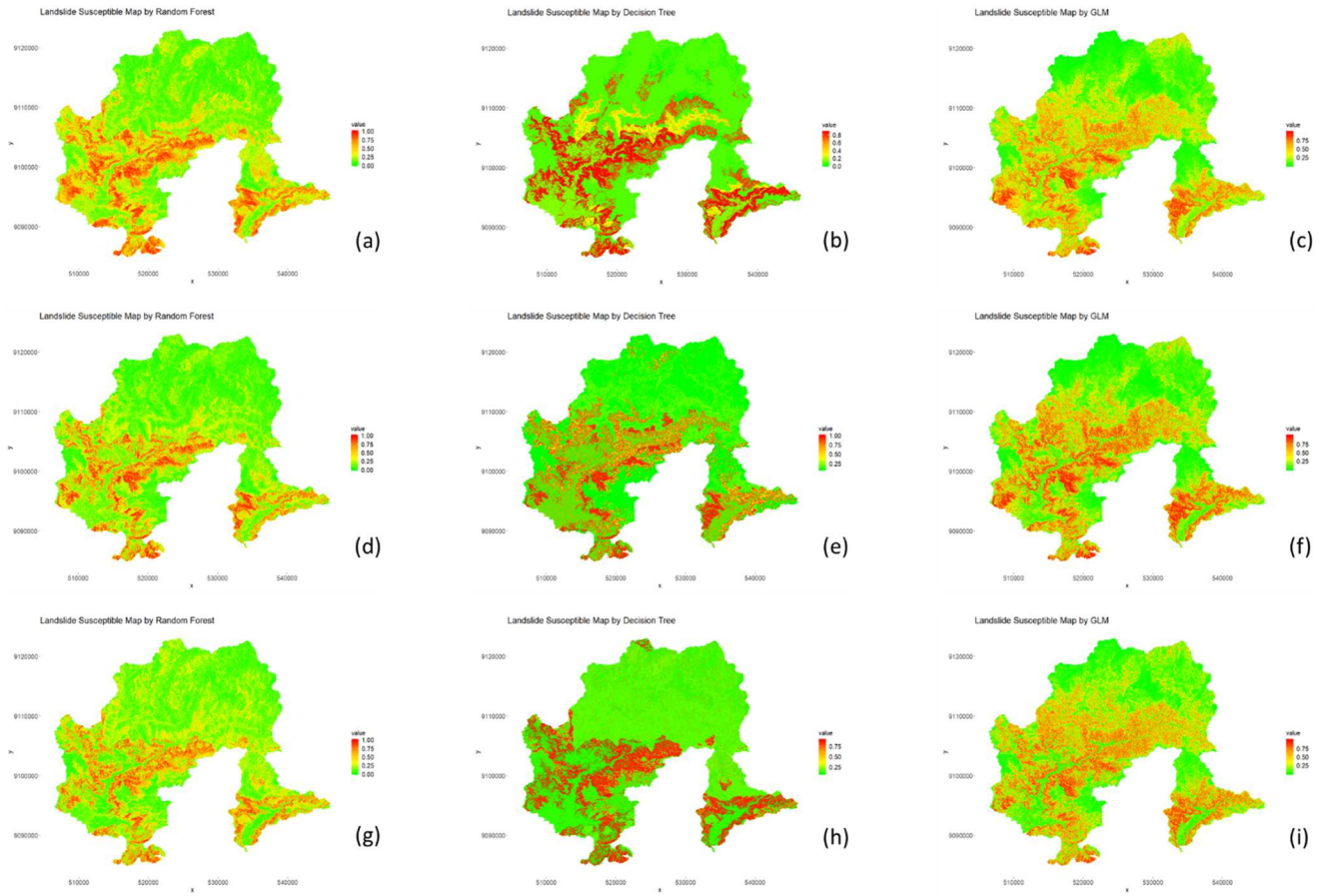


Fig. 7 Landslide susceptibility maps using the first dataset: (a) random forest, (b) decision tree, (c) glm; the second dataset: (d) random forest, (e) decision tree, and (f) glm; and the third dataset: (g) random forest, (h) decision tree, and (i) glm.

The maps showed the susceptibility of landslide occurrence based on the constructed models and have 8m x 8m for the resolution. The maps' value ranges from 0 to 1, showing the area with low to high susceptibility of landslide occurrences. The red color in the maps indicates high susceptibility to landslide event, while the green color indicates areas with low susceptibility. Based on the results, RF and GLM showed similar landslide susceptible areas. The maps showed insignificant differences in all models showing good accuracy to give a constant prediction.

Meanwhile, the maps generated by DT showed significant results with larger area predicted as high susceptibility using the third dataset compared to the other two dataset results. DT is commonly known for suffering from the major instability [46], explaining the significant differences in the map results. The overall distribution of landslide results from all models showed similar patterns for the susceptible landslide prediction, discovering that the area with high susceptibility to landslide is located in the western region of Pacitan regency.

### *C. Discussion*

In the validation of model performance, this study discovered that RF has the best performance among all the applied models. However, RF's AUC value showed a particular result with very high value, almost reaching a value of 1. This result indicates that the model has an overfitting problem. It also showed that the training data excessively trained the RF. A study conducted by Park [7] derived similar results in his study associated with a poor generalization of training data which increased the errors. There can be many reasons for over fitting. In addition, this study also has a much smaller landslide area compared to the non-landslide area; thus, the lack of information could make the model incapable of learning and predicting precisely despite its high accuracy rate.

Based on Table 4, the model using the first dataset has proven to have the highest accuracy among all datasets. And the accuracy for RF and GLM model was decreasing when

the training and testing ratio was equally set (third dataset). This circumstance indicated that if the suitable data partition to generate the RF and GLM model was using datasets with training data ratio bigger than the testing data. Arabameri *et al* also used several methods, which are RF, Alternative DT (ADTree), and Fisher's Linear Discriminant Function (FLDA) to do landslide susceptibility mapping in their research [47]. The result also indicated that RF resulted in the best result for landslide susceptibility mapping compared to another. However, this research only employed 70:30 partition for its dataset, so the effect of training and testing dataset ratio was not taken into account. We conducted three different ratio datasets in each model to investigate the effect of dataset partition on the model's performance. RF can perform the most stable compared to the other models when employed with a different data partition.

On the other hand, DT showed better performance using the third dataset compared to other datasets. The minor differences in each DT model accuracy assessment led to inconsistencies in DT map results. These inconsistencies were indicated in the maps generated by DT, showing different maps prediction compared to RF and GLM maps even though with only minor accuracy differences. Overall, RF is proven to produce the best model among other models in this study.

Besides, topographical factors are considered to hold great importance as well in model performances [7]. As mentioned in the previous section, elevation is considered the most important variable for land susceptibility mapping. Elevation was known to be frequently used in landslide susceptibility studies. The study was conducted in an area with hilly topography. Thus, it portrays areas with different relative relief. In this study, the landslide occurred mostly in intermediate elevations around 300-600 meters above mean sea level. In his study, Dai encountered similar circumstances where landslide events mostly happened in intermediate elevation. He stated that this circumstance was possible because in high elevation, the topography is mostly characterized by weathered rocks, and the shear strength of these is much higher, causing this area to be less invulnerable. In low elevation, the frequency of landslides is low due to the gentle terrain, and it requires a higher perched water table to initiate landslides because this area is usually covered with thick colluvium and residual soils. Meanwhile, intermediate elevation slopes tend to be covered by a thin layer of colluvium, which causes the area to be more prone to landslides.

The landslide susceptibility map generated using RF and the first datasets, as shown in Fig. 8, predicted around 7,434 kilometers area in Pacitan region was classified as being moderate to highly susceptible to a landslide. These areas are mostly located in the southern part of the Pacitan region, while the northern part mostly has low susceptibility. The southern part of Pacitan region was widely known to be in high risk of landslide. Aside from the intermediate/moderate elevation, this area's geological condition also played an important role. The mass movements often occurred in this study area due to high weathering control where large amounts of clay minerals were produced; these included smectite, illite, and kaolin. The presence of smectite, itlite, and kaolin in the weathered zone triggers landslide in this area [48]. This result also proves that the geological condition was one of the most important factors in our model construction, as shown in Fig. 7.

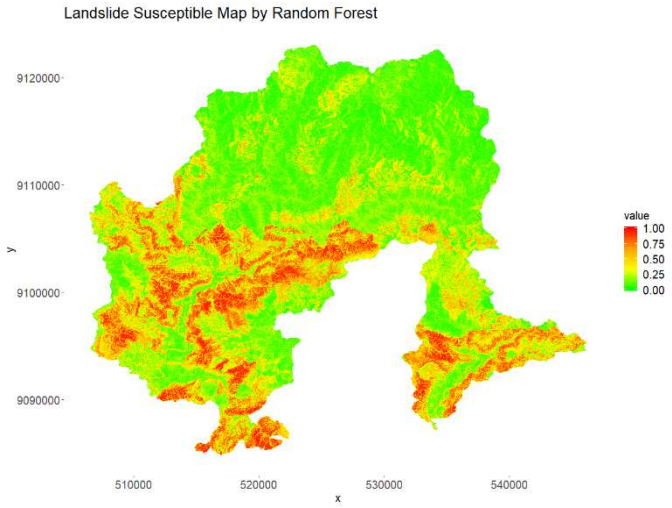


Fig. 8 Landslide susceptibility map generated by RF using the first dataset

## IV. Conclusion

Elevation was considered the most important factor in generating landslide susceptibility maps as shown in significant accuracy drops by model that excluded the given factor. The exclusion of profile curvature showed an insignificant drop in the model's performance; this has proved the least effect of this model construction factor.

Comparing the three datasets composition in three different machine learning methods showed various effects of each method in handling different training and testing dataset ratios. The topographical factors employed in this study, including elevation, slope, aspect, angle, curvature, distance to the river, and geological factor, were proven to be eligible for landslide susceptibility mapping. The accuracy assessment results showed the RF model to have the best performance while DT showed the lowest one.

The overall AUC and CC values of all models were more significant than 0.75, which concludes that the landslide susceptibility maps constructed in this study have good accuracy in the spatial prediction, and they were able to serve this purpose. The determination of variables' importance was strongly related to the techniques and variable combination used. Thus, the importance of each variable can be different if applied to other techniques.

Overall, the model with the best performance in this study was the RF model with 70:30 data ratio. However, because the limitation of landslide events used in our research landslide inventory map, we suggest further research with more landslide samples for a better result. Another machine learning method, such as Boosted Regression Tree (BRT) also can employ to see if the model could be yield a better result that RF.

## Acknowledgment

This research was funded by Penelitian Pemula ITS 2019 under grant number 1209/PKS/ITS/2019.

## REFERENCES


[1] A. Arabameri, B. Pradhan, K. Rezaei, M. Sohrabi, and Z. Kalantari, "GIS-based landslide susceptibility mapping using numerical risk factor bivariate model and its ensemble with linear multivariate regression and boosted regression tree algorithms," *J. Mt. Sci.*, vol. 16, no. 3, pp. 595–618, 2019.

[2] Y. Cui, D. Cheng, C. E. Choi, W. Jin, Y. Lei, and J. S. Kargel, "The cost of rapid and haphazard urbanization: lessons learned from the Freetown landslide disaster," *Landslides*, no. March, pp. 1167–1176, 2019.

[3] M. S. Rahman, B. Ahmed, and L. Di, "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.*, vol. 14, no. 10, pp. 1919–1937, 2017.

[4] P. Aleotti and R. Chowdhury, "Landslide hazard assessment: summary review and new perspectives," *Bull. Eng. Geol. Environ.*, vol. 58, no. 1, pp. 21–44, 1999.

[5] L.-J. Wang, M. Guo, K. Sawada, J. Lin, and J. Zhang, "Landslide susceptibility mapping in Mizunami City, Japan: A comparison between logistic regression, bivariate statistical analysis and multivariate adaptive regression spline models," *CATENA*, vol. 135, pp. 271–282, 2015.

[6] R. Fell, J. Corominas, C. Bonnard, L. Cascini, E. Leroi, and W. Z. Savage, "Guidelines for landslide susceptibility, hazard and risk zoning for land use planning," *Eng. Geol.*, vol. 102, no. 3, pp. 85–98, 2008.

[7] S. Park and J. Kim, "Landslide susceptibility mapping based on random forest and boosted regression tree models, and a comparison of their performance," *Appl. Sci.*, vol. 9, no. 5, 2019.

[8] H. Khan, M. Shafique, M. A. Khan, M. A. Bacha, S. U. Shah, and C. Calligaris, "Landslide susceptibility assessment using Frequency Ratio, a case study of northern Pakistan," *Egypt. J. Remote Sens. Sp. Sci.*, vol. 22, no. 1, pp. 11–24, 2019.

[9] S. Razavizadeh, K. Solaimani, M. Massironi, and A. Kavian, "Mapping landslide susceptibility with frequency ratio, statistical index, and weights of evidence models: a case study in northern Iran," *Environ. Earth Sci.*, vol. 76, no. 14, p. 499, 2017.

[10] S. Mandal and K. Mandal, "Bivariate statistical index for landslide susceptibility mapping in the Rorachu river basin of eastern Sikkim Himalaya, India," *Spat. Inf. Res.*, vol. 26, no. 1, pp. 59–75, 2018.

[11] Z. Xie, G. Chen, X. Meng, Y. Zhang, L. Qiao, and L. Tan, "A comparative study of landslide susceptibility mapping using weight of evidence, logistic regression and support vector machine and evaluated by SBAS-InSAR monitoring: Zhouqu to Wudu segment in Bailong River Basin, China," *Environ. Earth Sci.*, vol. 76, no. 8, p. 313, 2017.

[12] L. Lin, Q. Lin, and Y. Wang, "Landslide susceptibility mapping on a global scale using the method of logistic regression," *Nat. Hazards Earth Syst. Sci.*, vol. 17, no. 8, pp. 1411–1424, 2017.

[13] B. Zeng, W. Xiang, J. Rohn, D. Ehret, and X. Chen, "Assessment of shallow landslide susceptibility using an artificial neural network in Enshi region, China," *Nat. Hazards Earth Syst. Sci. Discuss.*, no. 388, pp. 1–46, 2017.

[14] D. Salcedo, O. P. Almeida, B. Morales, and T. Toulkeridis, "Landslide susceptibility mapping using fuzzy logic and multi-criteria evaluation techniques in the city of Quito, Ecuador," *Chinese J. Sensors Actuators*, vol. 11, no. 11, pp. 45–55, 2018.

[15] S. M. Fatemi Aghda, V. Bagheri, and M. Razifard, "Landslide Susceptibility Mapping Using Fuzzy Logic System and Its Influences on Mainlines in Lashgarak Region, Tehran, Iran," *Geotech. Geol. Eng.*, vol. 36, no. 2, pp. 915–937, 2018.

[16] S. Lee, S. M. Hong, and H. S. Jung, "A support vector machine for landslide susceptibility mapping in Gangwon Province, Korea," *Sustain.*, vol. 9, no. 1, pp. 15–19, 2017.

[17] J.-C. Kim, S. Lee, H.-S. Jung, and S. Lee, "Landslide susceptibility mapping using random forest and boosted tree models in Pyeong-Chang, Korea," *Geocarto Int.*, vol. 33, no. 9, pp. 1000–1015, Sep. 2018.

[18] P. R. Kadavi, C.-W. Lee, and S. Lee, "Landslide-susceptibility mapping in Gangwon-do, South Korea, using logistic regression and decision tree models," *Environ. Earth Sci.*, vol. 78, no. 4, p. 116, 2019.

[19] H. Hong *et al.*, "Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China)," *CATENA*, vol. 163, pp. 399–413, 2018.

[20] S. J. Park, C. W. Lee, S. Lee, and M. J. Lee, "Landslide susceptibility mapping and comparison using decision tree models: A case study of Jumunjin Area, Korea," *Remote Sens.*, vol. 10, no. 10, 2018.

[21] BNPB, "Buku Saku Tanggap Tangkas Tangguh Menghadapi Bencana," *Badan Nas. Penanggulangan Bencana*, p. 62, 2012.

[22] R. Azgha and M. Mukminan, "Analysis of the influence of tropical cyclones on rainfall in Indonesia," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 271, no. 1, 2019.

[23] G. F. Wieczorek, "Preparing A Detailed Landslide-Inventory Map for Hazard Evaluation and Reduction.," *Bull. Assoc. Eng. Geol.*, vol. 21, no. 3, pp. 337–342, 1983.

[24] F. Guzzetti, M. Cardinali, P. Reichenbach, and A. Carrara, "Comparing Landslide Maps: A Case Study in the Upper Tiber River Basin, Central Italy," *Environ. Manage.*, vol. 25, no. 3, pp. 247–263, 2000.

[25] M. Alvioli, A. C. Mondini, F. Fiorucci, M. Cardinali, and I. Marchesini, "Topography-driven satellite imagery analysis for landslide mapping," *Geomatics, Nat. Hazards Risk*, vol. 9, no. 1, pp. 544–567, 2018.

[26] M. R. M. Salleh *et al.*, "Geospatial approach for landslide activity assessment and mapping based on vegetation anomalies," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch.*, vol. 42, no. 4/W9, pp. 201–215, 2018.

[27] N. Saleem, M. Enamul Huq, N. Y. D. Twumasi, A. Javed, and A. Sajjad, "Parameters derived from and/or used with digital elevation models (DEMs) for landslide susceptibility mapping and landslide risk assessment: A review," *ISPRS Int. J. Geo-Information*, vol. 8, no. 12, 2019.

[28] M. S. Alkhasawneh, U. K. Ngah, L. T. Tay, N. A. Mat Isa, and M. S. Al-batah, "Determination of Important Topographic Factors for Landslide Mapping Analysis Using MLP Network," *Sci. World J.*, vol. 2013, p. 415023, 2013.

[29] M. Di Napoli, P. Marsiglia, D. Di Martire, M. Ramondini, S. L. Ullo, and D. Calcaterra, "Landslide susceptibility assessment of wildfire burnt areas through earth-observation techniques and a machine learning-based approach," *Remote Sens.*, vol. 12, no. 15, 2020.

[30] Y. Goulamoussène, C. Bedeau, L. Descroix, L. Linguet, and B. Hérault, "Environmental control of natural gap size distribution in tropical forests," *Biogeosciences*, vol. 14, no. 2, pp. 353–364, 2017.

[31] L. Jiang, D. Ling, M. Zhao, C. Wang, Q. Liang, and K. Liu, "Effective identification of terrain positions from gridded DEM data using multimodal classification integration," *ISPRS Int. J. Geo-Information*, vol. 7, no. 11, 2018.

[32] M. Różycka, P. Migoń, and A. Michniewicz, "Topographic Wetness Index and Terrain Ruggedness Index in geomorphic characterisation of landslide terrains, on examples from the Sudetes, SW Poland," *Zeitschrift für Geomorphol. Suppl. Issues*, vol. 61, no. 2, pp. 61–80, 2017.

[33] W. Chen *et al.*, "GIS-based landslide susceptibility evaluation using a novel hybrid integration approach of bivariate statistical based random forest method," *CATENA*, vol. 164, pp. 135–149, 2018.

[34] H. D. Skilodimou, G. D. Bathrellos, E. Koskeridou, K. Soukis, and D. Rozos, "Physical and anthropogenic factors related to landslide activity in the northern Peloponnese, Greece," *Land*, vol. 7, no. 3, 2018.

[35] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[36] O. Rahmati *et al.*, "Groundwater spring potential modelling: Comprising the capability and robustness of three different modeling approaches," *J. Hydrol.*, vol. 565, pp. 248–261, 2018.

[37] D. Šebalj, J. Franjković, and K. Hodak, "Shopping intention prediction using decision trees," *Millenium - J. Educ. Technol. Heal.*, no. No. 4 (2017): Series 2, n.º 4, 2017.

[38] L. Rokach and O. Maimon, *Data Mining With Decision Trees: Theory and Applications*, 2nd ed. USA: World Scientific Publishing Co., Inc., 2014.

[39] P. M. Atkinson and R. Massari, "Generalised Linear Modelling Of Susceptibility to Landsliding in the Central Apennines, Italy," *Comput. Geosci.*, vol. 24, no. 4, pp. 373–385, 1998.

[40] F. Djeddaoui, M. Chadli, and R. Gloaguen, "Desertification susceptibility mapping using logistic regression analysis in the Djelfa Area, Algeria," *Remote Sens.*, vol. 9, no. 10, pp. 1–26, 2017.

[41] M. Goldburd, D. Guller, A. Khare, and D. Tevet, *Generalized Linear Models for Insurance Rating*, Second Edi. Arlington, Virginia: Casualty Actuarial Society, 2020.

[42] S. S. Rwanga and J. M. Ndambuki, "Accuracy Assessment of Land Use/Land Cover Classification Using Remote Sensing and GIS," *Int. J. Geosci.*, vol. 08, no. 04, pp. 611–622, 2017.

[43] R. E. Alexander, "A Comparison of GLM , GAM , and GWR Modeling of Fish Distribution and Abundance in Lake Ontario," University of Southern California, 2016.
[44] J. E. McKenna Jr. and C. Castiglione, "Model distribution of Silver Chub (Macrhybopsis storeriana) in western Lake Erie," *Am. Midl. Nat.*, vol. 171, no. 2, pp. 301–310, 2014.
[45] D. Tien Bui *et al.*, "Spatial prediction of rainfall-induced landslides for the Lao Cai area (Vietnam) using a hybrid intelligent approach of least squares support vector machines inference model and artificial bee colony optimization," *Landslides*, vol. 14, no. 2, pp. 447–458, 2017.
[46] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees," *Classification and Regression Trees*. pp. 1–358, 1984.
[47] A. Arabameri, S. Saha, J. Roy, W. Chen, T. Blaschke, and D. T. Bui, "Landslide susceptibility evaluation and management using different machine learning methods in the Gallicash River Watershed, Iran," *Remote Sens.*, vol. 12, no. 3, 2020.
[48] N. Dzakiya, R. A. Hidayah, and Larikiansyah, "Analisis Potensi Longsor Menggunakan Metode Geolistrik Konfigurasi Dipole-dipole di Desa Kasihan Kecamatan Tegalombo Kabupaten Pacitan Jawa Timur," *J. Mater. dan Pembelajaran Fis.*, vol. 2, no. 8, pp. 17–22, 2018.