Application of Machine Learning on Google Earth Engine to Produce Landslide Susceptibility Mapping (Case Study: Pacitan)

Hafsa Fatihul Ilmy¹, Mohammad Rohmaneo Darminto²*, Amien Widodo³

¹Geomatics Department, National Cheng Kung University, No.1 University Road, Tainan City 701, Taiwan
²Department of Geomatics Engineering, Sepuluh Nopember Institute of Technology, Indonesia, Kampus ITS, Keputih, Sukolilo, Surabaya, Indonesia 60111
³Department of Geophysics Engineering, Sepuluh Nopember Institute of Technology, Indonesia, Kampus ITS, Keputih, Sukolilo, Surabaya, Indonesia 60111

*²rohmaneo@its.ac.id

Abstract. According to the Indonesian Disaster Management Agency (BNPB), Indonesia’s losses due to landslides were estimated around hundreds of billion rupiah in 2017. Making landslide as one of the catastrophes with the greatest risk of loss and leaving a couple regions prone to landslides in Indonesia, Pacitan region is one of them. Landslide delineation therefore represents a particularly beneficial application of evolving research trend in disaster reduction, especially for the vulnerable region. In the present times of open-access satellite data, cloud computing and machine-learning algorithms is frequently used for disaster prevention monitoring. By employing Google Earth Engine, this study focuses on the susceptibility of landslide occurrence using a random forest machine-learning framework applied to digital topographic data such as elevation, slope and aspect as the independent variables and landslide inventory data obtained from Ministry of Energy and Mineral Resources Republic of Indonesia as the dependent variable. This study data sets composed from 1000 random points in Pacitan region with 70:30 ratio for training and testing sample points. The model produced good result, with overall accuracy values of 0.94, kappa values of 0.79 and 0.80 for AUC value. This model also showed that elevation is the most important variable in the landslide susceptible area. The results of this study can be used to evaluate the potential future impacts of landslide and help to optimize the management of disaster reduction in the region of Pacitan.

Keywords: Google Earth Engine, Landslide Susceptibility Map, Random Forest, Pacitan

I. Introduction

Landslides are a phenomenon of natural disasters that cause significant losses of human life and constructed environment around the world. [1], [2]. A total of 4,862 landslides were reported worldwide between 2004 and 2016, causing a total of 55,999 reported fatalities and global economic losses of tens of billions of USD. [3], [4]. These maps provide specific policies and reference for governments to make effective decisions to proactively mitigate the effects of future landslides. [5], [6]. However, it is difficult to prepare accurate map with large data sets and using the regular desktop specification would consuming a lot of time [7]. Recently, these large data sets are supported in cloud computing technologies, such as Google Earth Engine [8], Sentinel Hub EO Browser [9] or the NASA Earth Exchange [10], and with
these technologies a reliable internet connection is now all that is required to access, manipulate, and analyze large volumes of data. Along with cloud computing technology the usage of machine learning (ML) algorithm has played a role in enabling the integration of various data sets [11]. ML approaches described as optimizing the efficiency of algorithms by learning about data relations from the data itself using a training data set with as much data variability as possible [12]. ML is widely used to either predict (e.g., regression models) or describe (e.g., classification) a set of data [7], [11]. Several ML models including logistic regression [13], geographically weighted regression (GWR) [14], support vector machine (SVM) [15], [16], and Classification and Random Tree (CART) [17], [18], random forest (RF) [19], have been implemented popularly among the others.

The combination of open access data, cloud computing technology and ML algorithm gives a potential benefit to hold a large-scale reliable, and repeatable landslide mapping and monitoring around the world. Our specific objective is to improve the effectiveness of landslide susceptibility maps for decision makers based on maximizing the usage of cloud computing and open-source data. To achieve this, the present study exploiting Google Earth Engine and RF method to predict the probability of landslide susceptibility map by considering four topographical factors. We conducted a case study within Pacitan Region, East Java, Indonesia, where landslide susceptibility mapping is of particular concern due to this event occurs every year, as well as the lack of consistent, up-to-date map of Pacitan’s landslide susceptibility. Lastly, we validate the models in term of their fitness and predictive powers and generate distribution map of landslide susceptibility.

2. Materials and Methods

In order to produce the landslide susceptibility map, this study employed four main steps that are shown in Figure 1 and described as follows: (1) study area selection; (2) data sets collection and preparation, (generating 1000 points; (3) defining the data sets as training and testing data, generating random forest model using Google Earth Engine platform based on training data, validating the model using testing data to produce confusion matrix and ROC graph; (4) to generate and analyze the resultant landslide susceptibility map in study. Further details are provided in the following sections.

![Figure 1. Flowchart describing the process to predict the probabilistic landslide susceptibility mapping in Pacitan region.](image-url)
2.1 Study Area

Our study area is located in Pacitan region (110.55° - 111.25° E and 7.55° - 8.17° S) with comprises of 1,410 km² area. This region is located in southwest of East Java, Indonesia (Figure 2). Pacitan is characterized by wavy to hilly plains containing high risk of landslides due to high precipitations during rainy season and the relevant geographical and topographical conditions. [20]. Tropical climate in Pacitan region consists of two seasons: dry (April-October) and rainy (October-April). Monthly average rainfall is between 3-503 mm, with the maximum value in December (503 mm). From 2015 to 2019 landslides occurred more than 3 times a year and mostly in rainy season.

![Figure 2. Map showing the location of the study area.](image)

2.2 Data Sets

Table 1 and Figure 3 summarizes each of the topographical data sets used in this study. They contain data that were obtained from earth observation (EO) sources and the existing landslide-inventory map used to derive model-input variables, as well as training data and testing data used in the model and to assess model performance. Several satellite scenes were obtained and processed within the Google Earth Engine (GEE) platform. All data were set to have resolution of 30x30 meter to have the same resolution in the model. Furthermore, the topographical data sets and landslide inventory maps were combined. 1000 random points were generated to extract the values of 4 topographical values including elevation, slope, aspect and NDVI, location coordinates, and value of 1 (landslide) and 0 (no-landslide). The points were then split into training and testing with 70:30 ratio.
Table 1. List and description of data sets included in the study, along with variables derived from each source.

| Data Set                              | Description                                                                                                                                                                                                 | Derived Variables          |
|---------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------|
| The Shuttle Radar Topography Mission (SRTM) digital elevation data | SRTM digital elevation data is an international research effort that obtained digital elevation models on a near-global scale. This product is provided by NASA JPL at a resolution of 1 arc-second (approximately 30m). | Elevation, Slope, Aspect    |
| Optical Imagery                       | 10-m Sentinel-2 optical satellite images from 2018-2019 were acquired over the study area, provided by the European Space Agency (ESA).                                                                       | Normalized Difference Vegetation Index (NDVI) |
| Existing Landslide Inventory          | Vector polygon based Pacitan’s Landslide Inventory data, produced by Ministry of Energy and Mineral Resources Republic of Indonesia using satellite image of imagery and participatory mapping acquired between 2015 and 2019. | Landslide-Non-Landslide Classification |

Figure 3. Thematic map of the landslide topographic factors selected for Pacitan region.
2.3 Machine Learning Model and Validation

As explained in previous section, training and testing data points were extracted from elevation, slope, aspect and NDVI data to be used as variables in the model. To produce the landslide susceptibility model, the random forest (RF) machine-learning algorithm was constructed on Google Earth Engine platform. RF method have been used widely to classify satellite data for mangrove cover mapping [21], wetland mapping [22], cropland mapping [23], natural hazard susceptibility area mapping [24]–[27] and land cover mapping [28]. RF is a nonparametric machine learning method comprised of ensembles of decision trees [29]. The random forest algorithm creates multiple decision trees which classify a random subset of the training data according to the covariate predictors. The final decisions of class and model construction are determined by the majority vote among all trees [30]. This model was performed within GEE cloud platform.

The evaluation of the model’s performances required the derivation of matrices of confusion that includes true positive (TP), false positive (FP), true negative (TN), and false negative (FN) categories. The confusion matrix contains useful information to calculate model performance and accuracy [31]. Therefore, this study calculated alternative performance measures including overall accuracy (OA) [31], [32], Kappa, and the area under the receiving operation characteristic curve (AUC) [31]. The OA is the ratio of samples correctly classified by the prediction model. Kappa is a statistic that measures the agreement of two categorical items [31], [32]. In addition, the kappa statistic is used to evaluate the reliability of the landslide models. Its value ranges from −1 (non-reliable) to 1 (reliable) [33].

The Receiver Operating Characteristic (ROC) curve is a graphical method that represents the relation between the False Positive fraction and the sensitivity for a range of thresholds [34]. This method has been widely used to measure the performance of probabilistic 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 [35]. The AUC value is the area under the ROC curve that validates quantitatively how well the general performance of landslide models [36]. Higher AUC value indicate better performance of landslide models and 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) [36],[37].

3. Results and Discussions

The performance of random forest model that was analyzed using the training dataset to produce landslide susceptibility showed high (0.94) overall accuracy (OA) with kappa index of 0.79. In addition, the model rates (model success and prediction performance) were analyzed using the training and testing datasets (Figure 4). Both success and prediction rate have good AUC value showing 0.80 and 0.81, respectively. Overall, the AUC values of all models were greater than 0.7. These results showed that the random forest model generated in this study has good accuracy and indicating reasonable accuracy in the spatial prediction of landslide susceptibility. Furthermore, the landslide susceptibility map produced different results depending on the variables used in each model. In this study, the topographical data used in landslides model made different contributions to the models. Figure 5 illustrates the importance of each topographical data, calculated in the RF model. Based on the figure, aspect and slope have similar influence which also indicate that these two variables have
higher predictive capability than elevation and NDVI. This conditions also reflecting that aspect and slope are important topographic conditions in landslide susceptibility. This relationship also demonstrated by other topographical-based landslide area mapping efforts such as described by [36] where the second influenced variable were elevation and NDVI.

Figure 4. Analysis of the receiver operating characteristic (ROC) curve for the landslide susceptibility map: (a) success rate curve using the training data set, and (b) prediction rate curve using the validation data set.

Figure 5. Relative importance of each landslide independent variable calculated in random forest model.

Figure 6 illustrate the landslide susceptibility map in Pacitan region produced using random forest model in GEE and multi-source remote sensing imagery. The map value ranges from 0 to 1 represented by green to red colors where area with low to high susceptible condition based on the elevation, slope, aspect, and NDVI information. The distribution of landslide susceptibility area with high value is around the center, south east, south west and coastal area (south) of Pacitan region.
4. Conclusion

This study focused on landslide susceptibility map generation at Pacitan region, East Java, Indonesia using cloud computing and open-sources data sets based on machine learning method. For this purpose, landslide-related spatial data consisting of a landslide inventory, and landslide topographical factors were collected and prepared on Google Earth Engine (GEE) cloud computing platform. Four landslide topographical factors were constructed from earth observation (EO) sources. These factors included elevation, slope, aspect and NDVI. The random forest (RF) model was implemented in GEE using ML method. The landslide susceptibility map produced in this study has proved the capability of cloud computing and open-source data for providing rapid landslide susceptibility map that can be updated regularly for decision makers and urban planners. The results could also be used to enhance disaster management as well as to prevent economic losses and causalities in the future. In future study, the accuracy of landslide susceptibility map could be enhanced by selecting more topographical factors as well annual precipitation and compare the model with other model using different algorithms.

Acknowledgement

This research was funded by research grant Penelitian Pemula ITS 2019 under grant number 1209/PKS/ITS/2019 given to Mohammad Rohmaneo Darminto.

References

[1] D. Petley, “Global patterns of loss of life from landslides,” *Geology*, vol. 40, no. 10, pp. 927–930, Oct. 2012, doi: 10.1130/G33217.1.

[2] S. Lee, J.-H. Ryu, J.-S. Won, and H.-J. Park, “Determination and application of the weights for landslide susceptibility mapping using an artificial neural network,” *Eng. Geol.*, vol. 71, no. 3, pp. 289–302, 2004, doi:https://doi.org/10.1016/S0013-
7952(03)00142-X.

[3] M. J. Froude and D. N. Petley, “Global fatal landslide occurrence from 2004 to 2016,” *Nat. Hazards Earth Syst. Sci.*, vol. 18, no. 8, pp. 2161–2181, 2018, doi: 10.5194/nhess-18-2161-2018.

[4] E. Intrieri, T. Carlà, and G. Gigli, “Forecasting the time of failure of landslides at slope-scale: A literature review,” *Earth-Science Rev.*, vol. 193, pp. 333–349, 2019, doi: https://doi.org/10.1016/j.earscirev.2019.03.019.

[5] L. Piciullo, M. Calvello, and J. M. Cepeda, “Territorial early warning systems for rainfall-induced landslides,” *Earth-Science Rev.*, vol. 179, pp. 228–247, 2018, doi: https://doi.org/10.1016/j.earscirev.2018.02.013.

[6] Q. C. Tran et al., “Novel Ensemble Landslide Predictive Models Based on the Hyperpipes Algorithm: A Case Study in the Nam Dam Commune, Vietnam,” *Applied Sciences*, vol. 10, no. 11. 2020, doi: 10.3390/app10113710.

[7] J. N. Hird, E. R. DeLancey, G. J. McDermid, and J. Kariyeva, “Google earth engine, open-access satellite data, and machine learning in support of large-area probabilistic wetland mapping,” *Remote Sens.*, vol. 9, no. 12, 2017, doi: 10.3390/rs9121315.

[8] Google, “A Planetary-Scale Platform for Earth Science Data & Analysis.”

[9] L. Sinergise Laboratory for geographical information systems, “Sentinel Hub EO Browser,” [Online]. Available: https://www.sentinel-hub.com/explore/eobrowser/.

[10] NASA, “National Aeronautics and Space Administration Welcome to the NASA Earth Exchange (NEX).” https://nex.nasa.gov/nex/ (accessed Jul. 22, 2020).

[11] B. Waske, M. Fauvel, J. A. Benediktsson, and J. Chanussot, “Machine Learning Techniques in Remote Sensing Data Analysis,” *Kernel Methods for Remote Sensing Data Analysis*. pp. 1–24, Oct. 23, 2009, doi: 10.1002/9780470748992.ch1.

[12] D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker, “Machine learning in geosciences and remote sensing,” *Geosci. Front.*, vol. 7, no. 1, pp. 3–10, 2016, doi: https://doi.org/10.1016/j.gsf.2015.07.003.

[13] S. Z. Mousavi, A. Kavian, K. Soleimani, S. R. Mousavi, and A. Shirzadi, “GIS-based spatial prediction of landslide susceptibility using logistic regression model,” *Geomatics, Nat. Hazards Risk*, vol. 2, no. 1, pp. 33–50, Mar. 2011, doi: 10.1080/19475705.2010.532975.

[14] Y. Li, X. Liu, Z. Han, and J. Dou, “Spatial Proximity-Based Geographically Weighted Regression Model for Landslide Susceptibility Assessment: A Case Study of Qingchuan Area, China,” *Applied Sciences*, vol. 10, no. 3. 2020, doi: 10.3390/app10031107.

[15] H. Hong et al., “A novel hybrid integration model using support vector machines and random subspace for weather-triggered landslide susceptibility assessment in the Wuning area (China),” *Environ. Earth Sci.*, vol. 76, no. 19, p. 652, 2017, doi: 10.1007/s12665-017-6981-2.

[16] M. Abedini, B. Ghasemian, A. Shirzadi, and D. T. Bui, “A comparative study of support vector machine and logistic model tree classifiers for shallow landslide susceptibility modeling,” *Environ. Earth Sci.*, vol. 78, no. 18, p. 560, 2019, doi: 10.1007/s12665-019-8562-z.

[17] A. Akgun, S. Dag, and F. Bulut, “Landslide susceptibility mapping for a landslide-prone area (Fındıklı, NE of Turkey) by likelihood-frequency ratio and weighted linear combination models,” *Environ. Geol.*, vol. 54, no. 6, pp. 1127–1143, 2008, doi: 10.1007/s00254-007-0882-8.

[18] H. A. Nefeslioglu, E. Sezer, C. Gokceoglu, A. S. Bozkir, and T. Y. Duman, “Assessment of Landslide Susceptibility by Decision Trees in the Metropolitan Area of Istanbul, Turkey,” *Math. Probl. Eng.*, vol. 2010, p. 901095, 2010, doi: 10.1155/2010/901095.
[19] H. R. Pourghasemi and N. Kerle, “Random forests and evidential belief function-based landslide susceptibility assessment in Western Mazandaran Province, Iran,” 
*Environ. Earth Sci.*, vol. 75, no. 3, p. 185, 2016, doi: 10.1007/s12665-015-4950-1.

[20] Suprapto, “Tanah Longsor dan Banjir Bencana yang Mematikan di Indonesia (Data Tahun 2008-2017),” 2019.

[21] R. Jhonnerie, V. P. Siregar, B. Nababan, L. B. Prasetyo, and S. Wouthuyzen, “Random Forest Classification for Mangrove Land Cover Mapping Using Landsat 5 TM and Alos Palsar Imageries,” *Procedia Environ. Sci.*, vol. 24, pp. 215–221, 2015, doi: https://doi.org/10.1016/j.proenv.2015.03.028.

[22] O. Mutanga, E. Adam, and M. A. Cho, “High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm,” *Int. J. Appl. Earth Obs. Geoinf.*, vol. 18, pp. 399–406, 2012, doi: https://doi.org/10.1016/j.jag.2012.03.012.

[23] P. Teluguntla et al., “A 30-m landsat-derived cropland extent product of Australia and China using random forest machine learning algorithm on Google Earth Engine cloud computing platform,” *ISPRS J. Photogramm. Remote Sens.*, vol. 144, pp. 325–340, 2018, doi: https://doi.org/10.1016/j.isprsjprs.2018.07.017.

[24] F. Catani, D. Lagomarsino, S. Segoni, and V. Tofani, “Landslide susceptibility estimation by random forests technique: Sensitivity and scaling issues,” *Nat. Hazards Earth Syst. Sci.*, vol. 13, pp. 2815–2831, Nov. 2013, doi: 10.5194/nhess-13-2815-2013.

[25] H. R. Pourghasemi et al., “Assessing and mapping multi-hazard risk susceptibility using a machine learning technique,” *Sci. Rep.*, vol. 10, no. 1, p. 3203, 2020, doi: 10.1038/s41598-020-60191-3.

[26] S. Lee, J.-C. Kim, H.-S. Jung, M. J. Lee, and S. Lee, “Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea,” *Geomatics, Nat. Hazards Risk*, vol. 8, no. 2, pp. 1185–1203, Dec. 2017, doi: 10.1080/19475705.2017.1308971.

[27] K. Taalab, T. Cheng, and Y. Zhang, “Mapping landslide susceptibility and types using Random Forest,” *Big Earth Data*, vol. 2, no. 2, pp. 159–178, Apr. 2018, doi: 10.1080/20964471.2018.1472392.

[28] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, “Random Forests for land cover classification,” *Pattern Recognit. Lett.*, vol. 27, no. 4, pp. 294–300, 2006, doi: https://doi.org/10.1016/j.patrec.2005.08.011.

[29] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[30] N. Micheletti et al., “Machine Learning Feature Selection Methods for Landslide Susceptibility Mapping,” *Math. Geosci.*, vol. 46, no. 1, pp. 33–57, 2014, doi: 10.1007/s11004-013-9511-0.

[31] S. Begueria, “Validation and Evaluation of Predictive Models in Hazard Assessment and Risk Management,” *Nat. Hazards*, vol. 37, no. 3, pp. 315–329, 2006, doi: 10.1007/s11069-005-5182-6.

[32] R. E. Alexander, “A Comparison of GLM, GAM, and GWR Modeling of Fish Distribution and Abundance in Lake Ontario,” University of Southern California, 2016.

[33] 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, doi: 10.1674/0003-0031-171.2.301.

[34] W. Thuiller, M. B. Araújo, and S. Lavorel, “Generalized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales,” *J. Veg. Sci.*, vol. 14, no. 5, pp. 669–680, 2003, doi: 10.1111/j.1654-1103.2003.tb02199.x.

[35] C. Baeza, N. Lantada, and J. Moya, “Validation and evaluation of two multivariate statistical models for predictive shallow landslide susceptibility mapping of the Eastern
Pyrenees (Spain),” *Environ. Earth Sci.*, vol. 61, no. 3, pp. 507–523, 2010, doi: 10.1007/s12665-009-0361-5.

[36] B. T. Pham, B. Pradhan, D. Tien Bui, I. Prakash, and M. B. Dholakia, “A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India),” *Environ. Model. Softw.*, vol. 84, pp. 240–250, 2016, doi: https://doi.org/10.1016/j.envsoft.2016.07.005.

[37] 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, doi: 10.1007/s10346-016-0711-9.