Journal of Natural Resources and Environmental Management
http://dx.doi.org/10.29244/jpsl.14.2.407

RESEARCH ARTICLE

OPEN ACCESS

Check for updates

Modeling Landslide Hazard Using Machine Learning: A Case Study of Bogor, Indonesia

Boedi Tjahjono[a], Indah Firdania[a], Bambang Hendro Trisasongko[a,b]

[a] Department of Soil Science and Land Resources, Faculty of Agriculture, IPB University, IPB Darmaga Campus, Dramaga, Bogor, 16680, Indonesia
[b] Geospatial Information and Technologies for the Integrative and Intelligent Agriculture (GITIIA), Center for Regional System Analysis, Planning and Development (CRESTPENT), IPB University, IPB Baranangsiang Campus, Bogor, 16153, Indonesia

Article History
Received 07 July 2023
Revised 13 October 2023
Accepted 13 June 2024

Keywords
forest, landslide, machine learning, random modeling

ABSTRACT

Landslides occur in many parts of the world. Well-known drivers, such as geological activities, are often enhanced by violent precipitation in tropical regions, creating complex multi-hazard phenomena that complicate mitigation strategies. This research investigated the utility of spatial data, especially the digital elevation model of SRTM and Landsat 8 remotely sensed data, for the estimation of landslide distribution using a machine learning approach. Bogor Regency was chosen to demonstrate the approach considering its vast hilly/mountainous terrain and high rainfall. This study aimed to model landslide hazards in Sukajaya District using random forests and analyze the key variables contributing to the isolation of highly probable landslides. The initial model, using the default settings of random forest, demonstrated a notable accuracy of 93%, with an accuracy ranging from 91 to 94%. The three main predictors of landslides are rainfall, elevation, and slope inclination. Landslides were found to occur primarily in areas with high rainfall (2,668–3,228 mm), elevations of 500 to 1,500 m, and steep slopes (25–45%). Approximately 4,536 ha were potentially prone to landslides, while the remaining area (> 12,000 ha) appeared relatively sound.

Introduction

Natural disasters vary significantly, and each has specific characteristics that lead to complex mitigation strategies. Hydrometeorological and geological disasters are two types of naturally occurring disasters commonly found in Indonesia [1]. The first has been recurring, as Indonesia is located in a tropical region with high, often torrential, rainfall with a distinctive amplitude. This leads to disasters in the form of droughts, landslides, tornadoes, and floods. With a shift in climate patterns [2], prior expectations of drought and floods are no longer completely valid, a situation that warrants suitable adaptive mitigation planning. Geological disasters are closely associated with the geographical location of Indonesia, intertwining the Eurasian, Indo-Australian, and Pacific plates. These have been acknowledged to be very active compared to the rest of the world [3]. Their movement drives tectonic activities, which result in diverse and frequent natural disasters, such as earthquakes, tsunamis, volcanic eruptions, and landslides.

The latter has a specific occurrence scheme that has been an emerging research focus in disaster studies. Although many consider landslides to be geologically related disasters, they can also be triggered by intensive, high-volume precipitation in specific regions. Landslide disaster studies using multiple disaster sources have attracted considerable research attention [4,5]. Given the nature of hazard or disaster research, the spatial context is critical in data analysis, information extraction, and visualization. Spatial data contribute to wall-to-wall studies of landslides and other hazard studies. In general, these can be classified into vector and raster data. Data presented in vectors generally provide baseline information such as a final map for public use. This type of data does not suit the frequently updated information, which conforms to raster data.

Corresponding Author: Boedi Tjahjono ✉ boetjah@apps.ipb.ac.id 📫 Department of Soil Science and Land Resources, Faculty of Agriculture, IPB University, IPB Darmaga Campus, Dramaga, Bogor, Indonesia.

In hazard studies, both digital elevation model (DEM) and Earth observation data are useful. DEM data representing the morphology of a region have been demonstrated to be vital for estimating and mapping landslide-affected areas [6]. The resolution of the DEM is a significant benefit of the analyses [7]. Thus, the need for better DEM data in the future, within the context of spatial and temporal resolution, would make a significant contribution to landslide monitoring and other hazard studies. Combining both types of raster data has proven useful. Earlier research summarized the utility of remote sensing images [8–10]. Extending this context, multiple hazards have recently gained increasing attention. Shafique et al. [11] analyzed remote sensing data for landslide events triggered by the Kashmir earthquake. Further extension has been made in terms of multi-temporal data to understand the history of landslides [12].

Irrespective of the type or amount of data, thematic information extraction has generally been performed using machine learning for either classification [13] or regression problems [14]. The main advantages of using this approach include optimization of the model through subsampling and options for model reimplementation with minimum data inputs. Despite this, studies on machine learning applications for hazards, especially landslides, have been conducted. Commonly used machine learning methods such as random forests have been demonstrated to provide accurate predictions of landslide vulnerability in Taiwan [15]. Tengtrairat et al. [16] indicated the benefits of a bidirectional long short-term memory (BiLSTM) machine learning model in disaster risk analysis in Thailand. Contemporary research shows significant performance of machine learning methods compared to conventional methods [17].

Although various machine learning techniques have been used in previous studies, the geographical context indicates the need for studies in specific geographies as a venue to better understand landslide characteristics or perhaps as a part of the development of generic machine learning models. This research was generally designed to develop machine-learning-based modeling as an initial step for the development of generic models in similar landscapes. Specifically, this research aimed to study the random forest approach to produce geographical estimates of landslide occurrence and examine important variables yielded by selected models. In addition, parameter optimization was performed to obtain a statistically better model.

Methodology

Study Area

This research was conducted in Sukajaya District, Bogor Regency (Figure 1). The research area has an undulating to mountainous topography, which includes the upper and middle slopes of the Halimun-Salak Mountains. With complex terrain conditions, the study area has a record of past landslides with a high likelihood of future land surface movements.

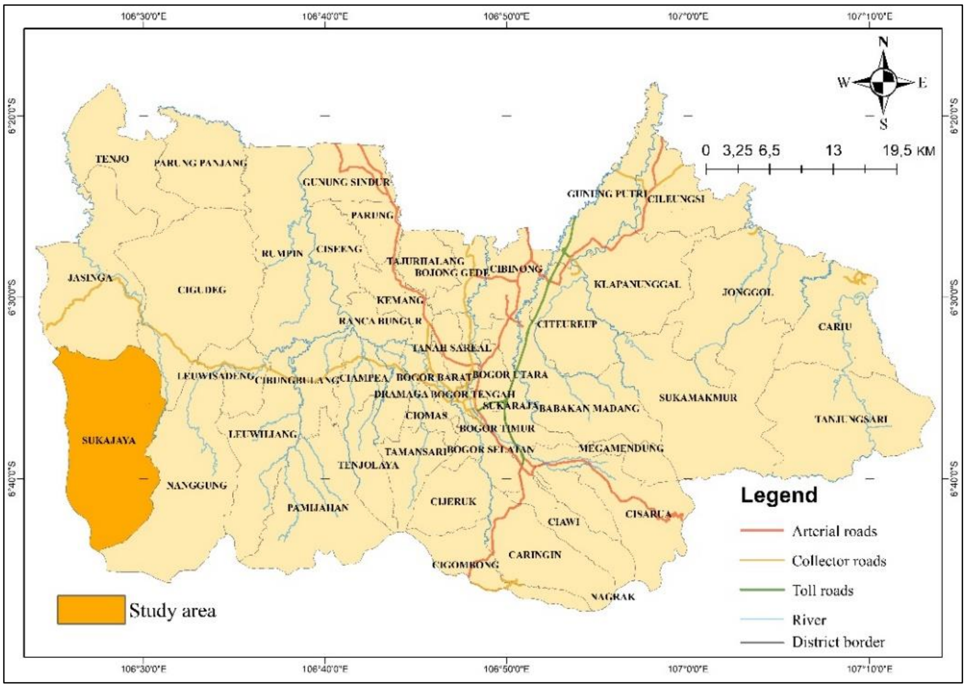

Figure 1. Study area.

Data Collection and Analysis

The landslide point inventory was carried out through observations and interpretations of high-resolution Google Earth images within the last three years (Figure 2). Some of these observation points were validated using field surveys conducted between February and March 2021. Road signs related to landslide occurrences were recorded, indicating evidence of past disasters. In addition, landslide scars (barren) or prior scars (usually non-woody vegetation) were also collected; some were informed by locals. All data (3,040 sample points), either landslide (1,021 points) or non-landslide (2,019 points) categories, were combined into a spatial database, which served as a sample set for data analysis using a machine learning approach. Figure 3 shows the spatial coverage of the sample dataset.

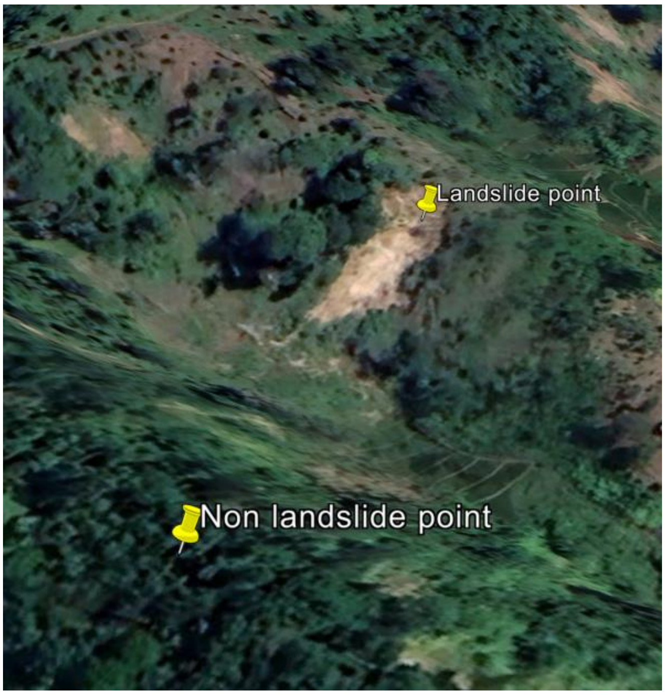

Figure 2. Identifying landslide events from Google Earth.

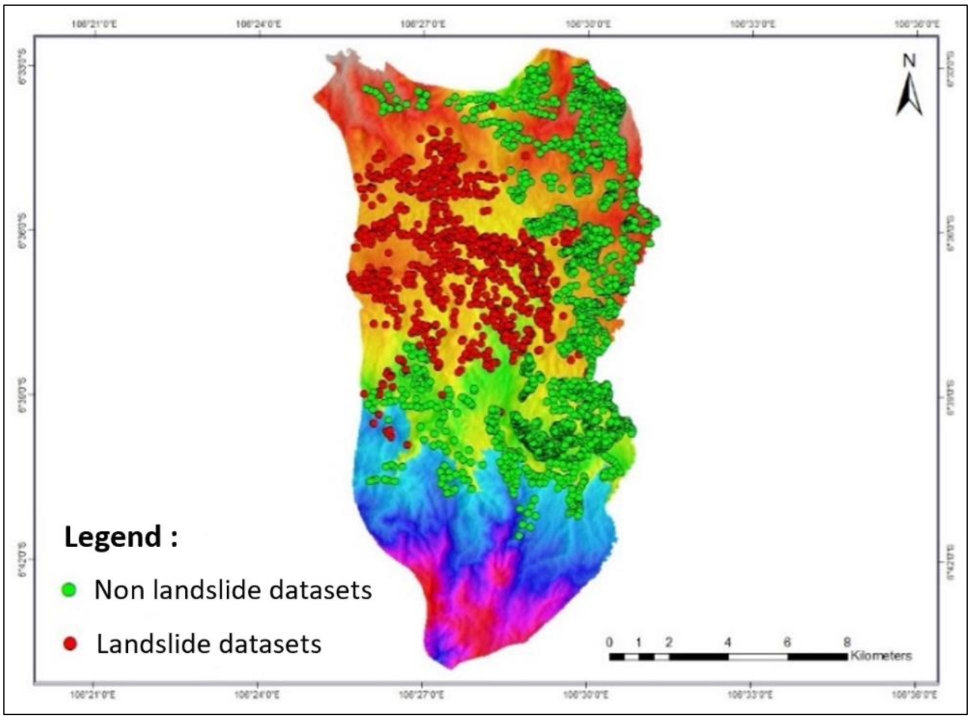

Figure 3. Spatial distribution of the landslide point dataset.

This study assessed proxies of landslide events derived from three types of primary data. Terrain conditions were represented by the Shuttle Radar Topography Mission (SRTM) DEM data with a spatial resolution of 30 m downloaded from the US Geological Survey (USGS) Website. This resolution was chosen to match the Landsat 8 data. These images with level 2 processing were downloaded; hence, standard radiometric correction was performed internally using USGS. Landsat data represent the dynamics of land cover, which is an essential element in landslide modeling. The third dataset was rainfall obtained from the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) downloaded from their website (https://www.chc.ucsb.edu/data/chirps, accessed on 2 July 2021). All datasets were spatially co-registered according to the Landsat data.

Figure 4 illustrates the research framework used in this study. DEM data were processed using QGIS software to generate five terrain variables: elevation, slope, aspect, Topographic Wetness Index (TWI), and slope curvature. Rózycka et al. [18] showed that although common terrain parameters, such as slope, were quite successful, the application of TWI in modeling offers opportunities for more detailed information. Changes in slope appearance due to landslides were indicated by high TWI values. Previous research indicated that DEM showing a specific plan curvature indicated high potential hazards compared to concave or convex slopes [19].

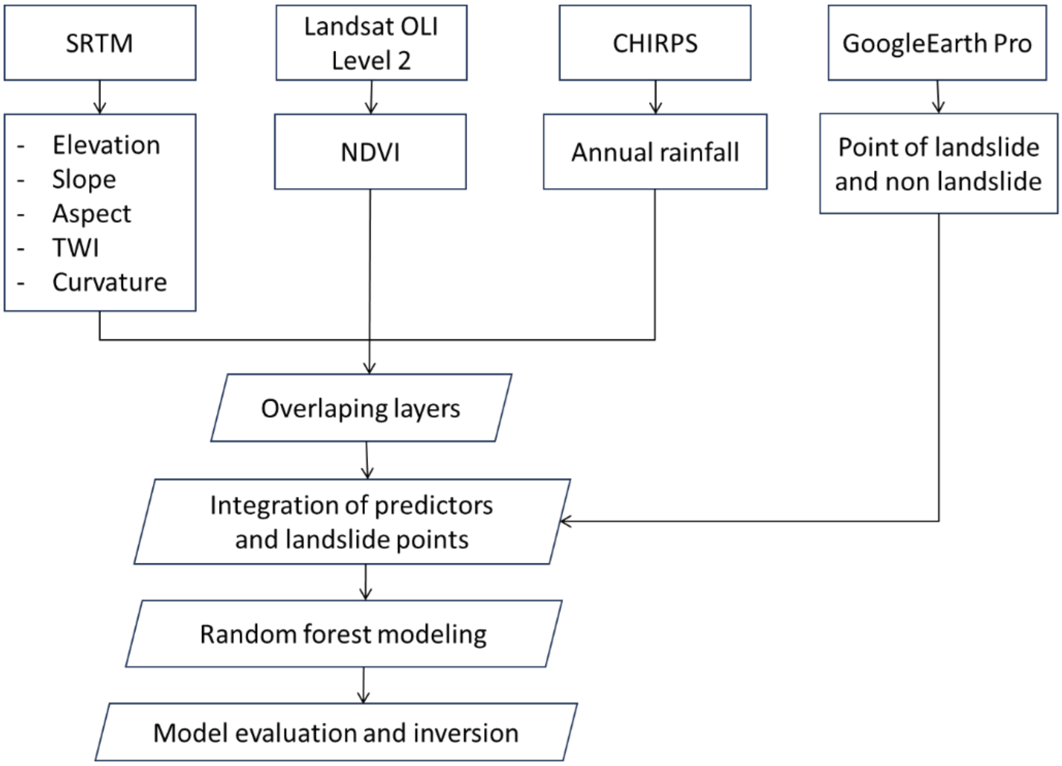

Figure 4. Research framework.

The availability of Landsat 8 Surface Reflectance (SR, Level 2) data accelerates data processing, considering that this type of data has undergone radiometric calibration. Thus, SR data can be directly inserted into the Normalized Difference Vegetation Index (NDVI) equation, as proposed by Rouse et al. [20], using bands 4 and 5. As a good representation of land cover conditions, NDVI has become one of the main variables in spatial analyses related to natural disasters, including landslides. NDVI analysis of landslide disasters has been proven in previous studies [21–23]. Rainfall is highly correlated with landslide events. High, frequent downpours would destabilize soils in sloping land surfaces; hence, this type of precipitation was ingested in the analysis. Hong et al. [24] presented a global analysis of the influence of rainfall on landslide events and concluded that rainfall is a potential proxy for landslide modeling.

Data analysis for classifying landslide or non-landslide classes was carried out in R software using a source code written in the RStudio software. All raster and vector data were converted employing 'raster' package. All sample data and raster attributes at corresponding locations were collected in the same set of data frames, an R terminology for data preparation prior to modeling. Shortly before modeling, samples were partitioned into training and testing data at a ratio of 70:30 using a ten-fold cross-validation resampling technique. Random forest modeling was carried out using commonly used R package, the 'randomForest' package. This approach was developed as an extension of conventional decision trees. While decision-making in decision tree methods is fairly robust for simple classification problems, a single decision-making process would not suit complex problems. Thus, an ensemble approach would theoretically reduce the bias. The ensemble approach has been adopted in many newly developed decision tree algorithms, including Extreme Gradient Boosting.

The random forest base model was examined through default parameterization, which was then revised by analyzing the variance of accuracy with several arbitrary parameter values. This procedure is known as tuning. This research focused on tuning the n-tree parameter, considering that the number of decision trees has been shown to be very important for enhancing the overall accuracy [25]. The overall accuracy was computed using validation data independent of the training data, which was 30% of the total sampling pixels. First, we assessed the confusion matrix to better understand the deviation. Next, we computed the overall accuracy by summarizing the diagonal of the confusion matrix. To analyze the variables with the highest contribution, this study used the variable importance approach. The model with the highest accuracy was subsequently inverted to generate estimates of the landslide susceptibility distribution. The inversion method was implemented using the 'raster' package in R.

Results and Discussion

Model Accuracy

The use of the random forest model with default settings resulted in an overall accuracy of 93%. The confidence interval ranged from 91 to 94%, suggesting that a random forest model is an excellent choice for the initial modeling process. The obtained results were better suited than similar models reported in India [26], Japan [27], and Greece [28]. However, the number of samples plays a significant role. This maiden outcome should then be broadened to include a diverse landscape, especially in areas with different precipitation levels. Machine learning algorithms generally have several parameters that need to be tuned. This is typically used to optimize the performance of the initial model. Compared to equivalent algorithms such as Support Vector Machine (SVM), the random forest method has fewer parameters to adjust. Therefore, this method is generally advisable, particularly when standard processing yields weak models. In this study, only the n-tree parameter was investigated. The results of the experiments with varying accuracies owing to changing n-tree parameters are presented in Table 1.

Table 1. Configuration of n-tree number and accuracy.

N-tree setting	Overall accuracy
100	0.925
200	0.926
300	0.929
400	0.929
500	**0.930**
600	0.929
700	0.928
800	0.927
900	0.927
1000	0.928

A tuning experiment indicated that the optimal n-tree for the data was 500. Although the results of overall accuracy at various values were not significantly different (roughly under 1%), the pattern of accuracy indicated diminishing returns. This indicates that setting excessive values can reduce accuracy, a condition similar to that in a previous report [25]. The results of model inversion are presented in Figure 5. This landslide estimation map indicated clustered locations of landslide disasters in the study area, particularly in the northern flank of the mountain range. However, the likelihood of landslides is low in most research areas. Further spatial analysis by calculating impacted areas showed that there was a high chance of landslides covering approximately 4,536 ha, whereas safe areas covered more than 12,000 ha.

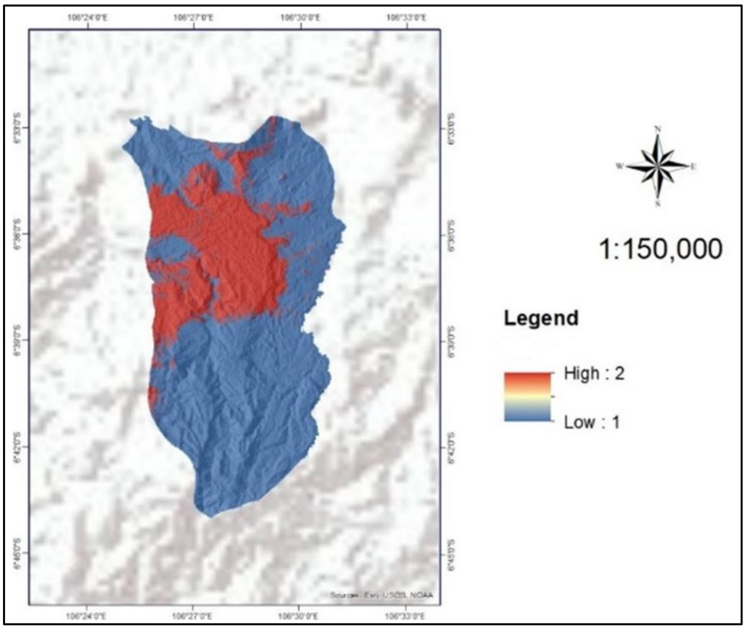

Figure 5. Estimation of landslide occurrence using the best model.

Contributing Variables

Figure 6 ranks the variables responsible for the targeted classes. The three most influential landslide predictors in the study area are rainfall, elevation, and slope. Landslide events were suspected to occur at locations with very high rainfall (2,668–3,228 mm), elevations ranging from 500 to 1,500 m above sea level, and steep slopes (25–45%). These proxies are widely understood as important variables in landslide event modeling. A very high level of precipitation could rapidly increase the soil moisture and put more weight on the soil column. While this would have less impact over flat terrain, in hilly or mountainous regions, gravity would trigger a substantial pull, leading to landslide events. Steep slopes also initiate landslides when gravitational force permits, even with substantial vegetation cover (see also the case presented in Figure 2).

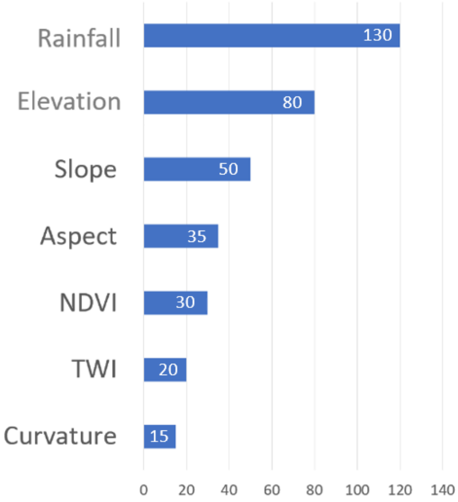

Figure 6. Response variables.

Conclusion

The random forest method, as a machine learning approach, was shown to be excellent for landslide hazard modeling because it yielded > 90% accuracy. Rainfall, elevation, and slope were the dominant factors that caused landslides in the study area. Tuning the n-tree parameters slightly indicates diminishing returns. The best outcome was provided by n-tree = 500, despite a difference of less than 1% compared with the other settings. Model inversion indicated landslide hazard areas covering approximately 4,536 ha, which is believed to be important for inclusion in future land use planning or developing mitigation strategies. Although high classification confidence was outlined in this study, the implementation of the model was fairly challenging. The primary considerations include the extent to which the model is applicable to different environments. This requires extensive near-future research in similar landscapes to investigate the variations and model overfitting. In terms of possible implementation in mitigation planning, an extension of this research, covering the entire area of Halimun-Salak National Park, would significantly contribute to better spatial planning at regency levels.

Acknowledgments

We would like to thank the Division of Remote Sensing and Spatial Information, Department of Soil Science and Land Resources, IPB University, for their support through the provision of the computational facilities. We also acknowledge the contributions of the reviewers in constructing a much better manuscript.

References

1. Suprayoga, H. Disaster Management in the implementation of the 2030 Sustainable Development Goals in Indonesia. *J. Indones. Sustain. Dev. Plan.* **2020**, *1*, 105–111, doi:10.46456/jisdep.v1i1.49.
2. Abdillah, A.; Buchari, R.A.; Widianingsih, I.; Nurasa, H. Climate change governance for urban resilience for Indonesia: A systematic literature review. *Cogent Soc. Sci.* **2023**, *9*, 2235170, doi:10.1080/23311886.2023.2235170.
3. Rusydy, I.; Fathani, T.F.; Al-Huda, N.; Sugiarto; Iqbal, K.; Jamaluddin, K.; Meilianda, E. integrated approach in studying rock and soil slope stability in a tropical and active tectonic country. *Environ. Earth Sci.* **2021**, *80*, 1–20, doi:10.1007/s12665-020-09357-w.
4. Maqsoom, A.; Aslam, B.; Khalil, U.; Kazmi, Z.A.; Azam, S.; Mehmood, T.; Nawaz, A. Landslide susceptibility mapping along the China Pakistan Economic Corridor (CPEC) route using multi-criteria decision-making method. *Model. Earth Syst. Environ.* **2022**, *8*, 1519–1533, doi:10.1007/s40808-021-01226-0.
5. Fan, X.; Yang, F.; Siva Subramanian, S.; Xu, Q.; Feng, Z.; Mavrouli, O.; Peng, M.; Ouyang, C.; Jansen, J.D.; Huang, R. Prediction of a multi-hazard chain by an integrated numerical simulation approach: The Baige Landslide, Jinsha River, China. *Landslides* **2020**, *17*, 147–164, doi:10.1007/s10346-019-01313-5.
6. Dahal, R.K.; Hasegawa, S.; Nonomura, A.; Yamanaka, M.; Dhakal, S. DEM-Based deterministic landslide hazard analysis in the Lesser Himalaya of Nepal. *Georisk* **2008**, *2*, 161–178, doi:10.1080/17499510802285379.
7. Claessens, L.; Heuvelink, G.B.M.; Schoorl, J.M.; Veldkamp, A. DEM resolution effects on shallow landslide hazard and soil redistribution modelling. *Earth Surf. Process. Landforms* **2005**, *30*, 461–477, doi:10.1002/esp.1155.
8. Mantovani, F.; Soeters, R.; Van Westen, C.J. Remote sensing techniques for landslide studies and hazard zonation in Europe. *Geomorphology* **1996**, *15*, 213–225, doi:10.1016/0169-555x(95)00071-c.
9. Mukherjee, S. Microzonation of seismic and landslide prone areas for alternate highway alignment in a part of western coast of India using remote sensing techniques. *J. Indian Soc. Remote Sens.* **1999**, *27*, 81–90, doi:10.1007/BF02990804.
10. Malkawi, A.I.H.; Saleh, B.; Al-Sheriadeh, M.S.; Hamza, M.S. Mapping of landslide hazard zones in Jordan using remote sensing and GIS. *J. Urban Plan. Dev.* **2000**, *126*, 1–17.
11. Shafique, M.; van der Meijde, M.; Khan, M.A. A review of the 2005 Kashmir earthquake-induced landslides; from a remote sensing prospective. *J. Asian Earth Sci.* **2016**, *118*, 68–80, doi:10.1016/j.jseaes.2016.01.002.
12. Lee, C.F.; Huang, W.K.; Chang, Y.L.; Chi, S.Y.; Liao, W.C. Regional landslide susceptibility assessment using multi-stage remote sensing data along the coastal range highway in northeastern Taiwan. *Geomorphology* **2018**, *300*, 113–127, doi:10.1016/j.geomorph.2017.10.019.

13. Panuju, D.R.; Paull, D.J.; Trisasongko, B.H. Combining binary and post-classification change analysis of augmented ALOS Backscatter for identifying subtle land cover changes. *Remote Sensing* **2019**, *11*, 1–24, doi:10.3390/rs11010100.

14. Trisasongko, B.H.; Paull, D.J.; Griffin, A.L.; Jia, X.; Panuju, D.R. On the relationship between the circumference of rubber trees and L-Band waves. *Int. J. Remote Sens.* **2019**, *40*, 6395–6417, doi:10.1080/01431161.2019.1591650.

15. Chang, K.T.; Merghadi, A.; Yunus, A.P.; Pham, B.T.; Dou, J. Evaluating scale effects of topographic variables in landslide susceptibility models using GIS-Based machine learning techniques. *Sci. Rep.* **2019**, *9*, 1–22, doi:10.1038/s41598-019-48773-2.

16. Tengtrairat, N.; Woo, W.L.; Parathai, P.; Aryupong, C.; Jitsangiam, P.; Rinchumphu, D. Automated landslide-risk prediction using web GIS and machine learning models. *Sensors (Basel).* **2021**, *21*, 1–32, doi:10.3390/s21134620.

17. Akinci, H.; Zeybek, M. Comparing classical statistic and machine learning models in landslide susceptibility mapping in Ardanuc (Artvin), Turkey. *Nat. Hazards* **2021**, *108*, 1515–1543, doi:10.1007/s11069-021-04743-4.

18. Rózycka, M.; Migoń, P.; Michniewicz, A. Topographic wetness index and terrain ruggedness index in geomorphic characterisation of landslide terrains, on examples from the Sudetes, SW Poland. *Zeitschrift fur Geomorphol.* **2017**, *61*, 61–80, doi:10.1127/zfg_suppl/2016/0328.

19. Ohlmacher, G.C. Plan curvature and landslide probability in regions dominated by earth flows and earth slides. *Eng. Geol.* **2007**, *91*, 117–134, doi:10.1016/j.enggeo.2007.01.005.

20. Rouse, J.W.; Haas, R.H.; Schell, J.A.; Deering, D.W. Monitoring vegetation systems in the great plains with ERTS. In Proceedings of the 3rd ERTS Symposium, Washington DC, USA, 10–14 December 1974.

21. Yang, W.; Wang, M.; Shi, P. Using MODIS NDVI time series to identify geographic patterns of landslides in vegetated regions. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 707–710, doi:10.1109/LGRS.2012.2219576.

22. Hsieh, H.C.; Chung, C.H.; Huang, C.Y. Using the NDVI and mean shift segmentation to extract landslide areas in the lioukuei experimental forest region with multi-temporal FORMOSAT-2 images. *Taiwan J. For. Sci.* **2017**, *32*, 203–222.

23. Sajadi, P.; Sang, Y.F.; Gholamnia, M.; Bonafoni, S.; Brocca, L.; Pradhan, B.; Singh, A. Performance evaluation of long NDVI timeseries from AVHRR, MODIS and Landsat sensors over landslide-prone locations in Qinghai-Tibetan Plateau. *Remote Sensing* **2021**, *13*, 1–27, doi:10.3390/rs13163172.

24. Hong, Y.; Adler, R.F.; Huffman, G. An Experimental global prediction system for rainfall-triggered landslides using satellite remote sensing and geospatial datasets. *IEEE Trans. Geosci. Remote Sensing* **2007**, *45*, 1671–1680, doi:10.1109/TGRS.2006.888436.

25. Trisasongko, B.H.; Panuju, D.R.; Paull, D.J.; Jia, X.; Griffin, A.L. Comparing six pixel-wise classifiers for tropical rural land cover mapping using four forms of fully polarimetric sar data. *Int. J. Remote Sens.* **2017**, *38*, 3274–3293, doi:10.1080/01431161.2017.1292072.

26. Pandey, A.; Dabral, P.P.; Chowdary, V.M.; Yadav, N.K. Landslide hazard zonation using remote sensing and GIS: A case study of Dikrong River Basin, Arunachal Pradesh, India. *Environ. Geol.* **2008**, *54*, 1517–1529, doi:10.1007/s00254-007-0933-1.

27. Hasegawa, S.; Dahal, R.K.; Nishimura, T.; Nonomura, A.; Yamanaka, M. DEM-based analysis of earthquake-induced shallow landslide susceptibility. *Geotech. Geol. Eng.* **2009**, *27*, 419–430, doi:10.1007/s10706-008-9242-z.

28. Tsangaratos, P.; Loupasakis, C.; Nikolakopoulos, K.; Angelitsa, V.; Ilia, I. Developing a landslide susceptibility map based on remote sensing, fuzzy logic and expert knowledge of the Island of Lefkada, Greece. *Environ. Earth Sci.* **2018**, *77*, 363, doi:10.1007/s12665-018-7548-6.