

Uneven Global Distribution and Machine Learning–Based Characterization of Landslide Events

Munish Kumar Mummadi

Saint Louis University

Harshil Sharma

Saint Louis University

Sean Kee

Saint Louis University

Khushi Chauhan

Saint Louis University

Payam Norouzzadeh

Oklahoma State University

Eli Snir

Washington University in Saint Louis

Steven Buckner

Saint Louis University

Bahareh Rahmani

Bahareh-Rahmani@utulsa.edu

University of Tulsa

Research Article

Keywords: Landslides, natural hazards, exploration data analysis, geospatial analysis, disaster risk

Posted Date: September 17th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10637323/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Uneven Global Distribution and Machine Learning–Based Characterization of Landslide Events

¹Munish Kumar Mummadi, ¹Harshil Sharma, ¹Sean Kee, ¹Khushi Chauhan, ²Payam Norouzzadeh, ³Eli Snir, ¹Steven Buckner, ⁴Bahareh Rahmani
¹Saint Louis University, Saint Louis, MO, USA
²Oklahoma State University, Still Water, OK, USA
³Washington University in Saint Louis, Saint Louis, MO, USA
⁴University of Tulsa, Tulsa, OK, USA

Corresponding Author:
Dr. Bahareh Rahmani
Bahareh.Rahmani@gmail.com

Abstract: Landslides are a danger that can cause many deaths, damage the infrastructure and disrupt the economy around the world. Landslides cause deaths. Understanding the trends in landslide occurrences and the effects of landslides is important for improving hazard awareness and for guiding risk reduction strategies. This study provides an analysis of global landslide incidents using the Global Landslide Catalog. The Global Landslide Catalog records landslide incidents from 1988 to 2017. The study uses analysis and descriptive analysis to examine the trends over time the geographic distribution, the classifications of landslide sizes and the associated death tolls. The findings indicate a notable spatial concentration of landslide events, along with a distinct correlation between the characteristics of landslides and their impact on people. This study highlights the significance of extensive observational datasets for assessing global landslide risks and lays the groundwork for future predictive modeling and risk-based evaluations.

Keywords: Landslides, natural hazards, exploration data analysis, geospatial analysis, disaster risk

1. Introduction

Landslides rank among the destructive natural hazards in the world causing loss of life damage to roads and buildings and long-term problems for people, in both rich and poor areas [18]. Landslides happen because of a mix of the shape of the land the rock layers, how people use the land and triggers like rain or earthquakes [1,2,17]. In years We have seen more climate swings and more people building in risky places and those changes have made the world worry more about landslide risk and impact [3].

Early landslide research mainly focused on the regional hazard zones using landform interpretation and multiple statistical methods [15,18]. The approaches gave insight into site specific processes. The approaches had limited applicability at spatial scales. The development of landslide inventories marked a critical advancement. The inventories enabled comparison of landslide occurrence across regions and time periods [17].

Global landslide inventories allow researchers to examine the scale spatial clustering the temporal trends and the human impacts that are not observable, in localized studies. NASA and collaborators built the Global Landslide Catalog (GLC). The Global Landslide Catalog is one of the complete global landslide inventories we have today [1,4]. The Global Landslide Catalog gathers reported landslide events from international sources. The Global Landslide Catalog uses reporting protocols for each entry. The Global Landslide Catalog provides the timing of each event the location, the size class of the landslide and the number of fatalities [2,5]. The Global Landslide Catalog suffers from reporting bias, uneven spatial coverage and limited environmental attributes. The Global Landslide Catalog remains a resource for global landslide analysis and hazard assessment [17].

When we look at the Global Landslide Catalog We see a picture of landslide activity, around the world. We rely on the Global Landslide Catalog for my research. We notice that machine learning methods have become more common as big observational data sets grow. Machine learning methods are now used in landslide susceptibility and hazard modelling. Prior studies have used tree-based ensembles support vector machines and hybrid learning frameworks to find relationships, between landslides and environmental factors [6,7,9,19,20]. We have seen that tree-based ensembles, support vector machines, hybrid learning frameworks can handle feature interactions. Statistical susceptibility models point out the strengths of ensembles support vector machines, hybrid learning frameworks in dealing with complex feature interactions. Statistical susceptibility models also point out the problems of class imbalance and feature redundancy [16].

Researchers often use Dimensionality reduction methods, Principal Component Analysis (PCA) to deal with multicollinearity. Researchers also use Dimensionality reduction methods to make the model less complex. Researchers apply Dimensionality reduction methods in landslide studies [13,14]. Researchers combine PCA-based frameworks with support vector machines. Researchers also combine PCA-based frameworks with regression models. Researchers further

combine PCA-based frameworks with classifiers to make computation faster and to keep the model stable [11,12,25]. Researchers find that evidence about the benefit of PCA, for making prediction performance is still mixed especially when PCA is used on different global data sets instead of local environmental inputs.

Recently researchers have added explainable AI, ensemble learning and deep learning architectures to the toolbox for landslide analysis [10,21–24]. Deep learning can. Segment images, but deep learning still faces limits, at a global scale. Data are scarce computing costs are high, and interpretability is low. Because of those limits classical machine learning models stay relevant for comparative global analyses.

We wanted to see how the global landslide patterns and the landslide size classification appear in the Global Landslide Catalog. We use the statistical analysis and the machine learning modelling to do that. We test the classifiers. The classifiers include the Random Forest, XGBoost, CatBoost and the Support Vector Machines. We test each classifier with the PCA based dimensionality reduction and without the PCA based dimensionality reduction. We do not try to get the predictive accuracy. We try to understand the scale spatial and temporal structure of the global landslide data. We also look at the feature importance. We also look at the limitations of the reported landslide data. In my view this study clearly notes the data constraints. This study provides a foundation, for the future global landslide risk modelling and the combined hazard assessment.

2. Data Description

We used the Global Landslide Catalog. The Global Landslide Catalog contains 2,217 documented landslide events from 1996 to 2016. For each event We recorded the year the event happened the latitude and longitude the admin area the population estimate for the area the distance, to the reference point the number of deaths reported and the landslide size class code. We see that Spatial attributes support the geographic pattern analysis. Landslide size serves as a proxy for the event magnitude. Population captures exposure. Fatality variables capture the impact. Administrative division identifiers enable the aggregation and the comparative analysis, across U.S. States. Before We started modeling, we examined the dataset to make sure it was complete and consistent. We removed the records that did not have time, place or size information. We encoded the variables, and we standardized the numerical features so that machine learning could work with them and so that We could reduce the number of dimensions. The cleaned dataset then became the base for all the analyses the visualizations and predictive modeling, in this study.

Clinical Trial Number: Not applicable (N/A); this study is an observational analysis of publicly available landslide data and is not a clinical trial.

3. Methodology

We looked at the landslide data to describe the global landslide patterns across time, space and categories. We examined how the landslide frequency changed over time to find any trends or busy periods. We mapped the landslide locations with latitude and longitude. Grouped the landslides by region and country. We sorted the landslides by size. Compared each size group to the number of deaths reported. We used statistics to add up the number of landslides and the number of deaths, for each size group. We used analysis to look at the differences in human impact across the event types. We did each analysis, with computational workflows. We made visualizations to help read the findings and to show the patterns clearly. The visualizations helped me see the patterns.

3.1 Dimensionality Reduction using Principal Component Analysis

Dimensionality Reduction using Principal Component Analysis Dimensionality reduction changes model performance. We tested that change, with Principal Component Analysis. We standardized the features before we applied Principal Component Analysis so that each variable had a similar scale. We gave the machine learning models the Principal Component Analysis components that explained 95 percent of the total variation. We then compared baseline models that used the feature space with model performance that used Principal Component Analysis transformed features.

4. Results

4.1 Temporal Trends

When We look at the timing of landslide events, we see that the number of reported landslide events varies over the study period. Some years have landslide event counts. Those higher landslide events may be because the environment was more active or because reporting of landslide events got better (Fig. 1).

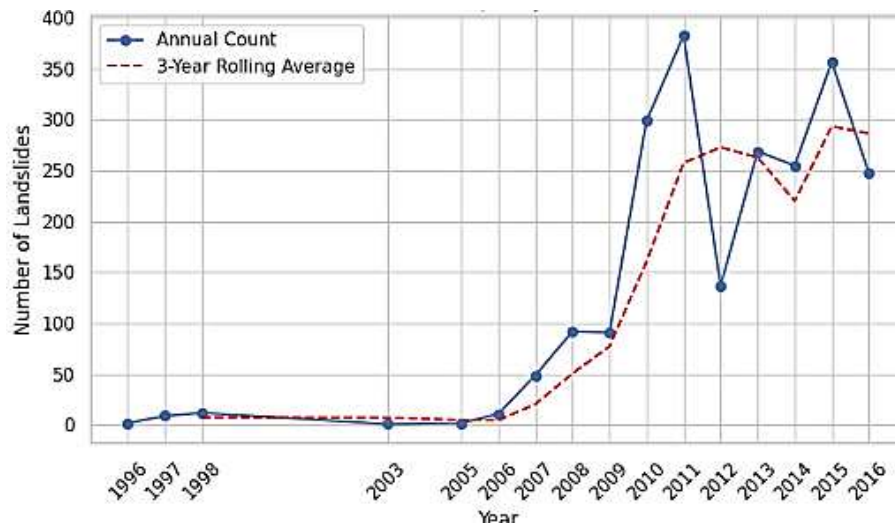


Figure 1:

Over Time

Landslide Frequency

4.2 Spatial Distribution

We notice that landslide events are not spread evenly across the map. We see numbers of landslide events in places with mountains, a lot of rain and active tectonic movement. We find that the pattern of landslide events lines up, with the areas that are already known to be landslide prone (Fig. 2).

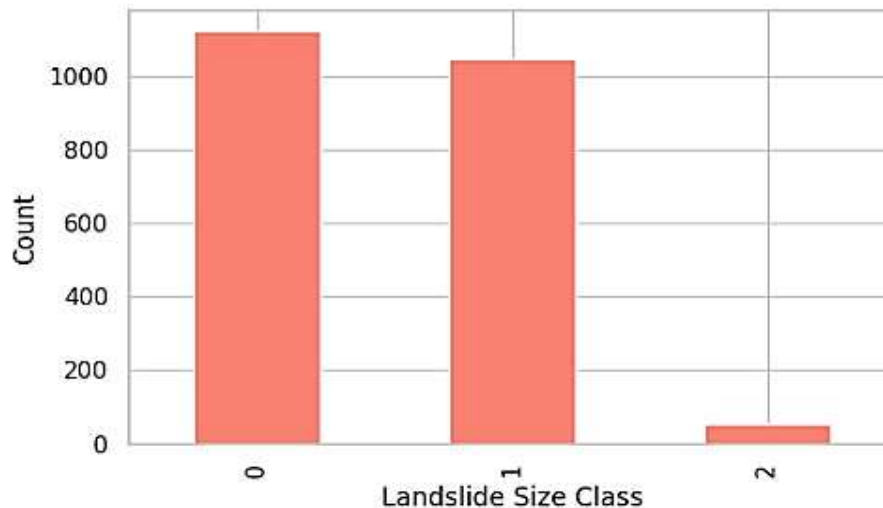


Figure 2: US Landslide Locations Size Color Fatality Size

4.3 Landslide Size Characteristics

Landslide Size Characteristics Figure 3 shows the distribution of landslide size categories. Medium-sized landslides make up most of the reported events while large events are relatively rare. The imbalance matters for hazard assessment and predictive modeling. Large landslides occur often but large landslides still have a high potential, for severe human impact. The need to consider both frequency and consequence in landslide risk analysis is clear (Fig. 3).

4.4 Feature Importance and Model Interpretation

The researchers performed feature importance analysis on Random Forest, XGBoost, CatBoost and Support Vector Machine models. Feature importance analysis looked at how each input variable contributed to the predictions and Feature importance analysis found that spatial and temporal features such as event year geographic coordinates and population-related attributes consistently ranked among the predictors, across all models as shown in Figure 4.

We see that the fatality count shows limited importance because the fatality count is an outcome not a cause of landslide size. We see that the importance rankings stay the same across tree-based models and permutation based SVM analysis. That shows the models agree on relationships even though the models are built differently. We see that the results show that location and time play the role in shaping landslide size patterns around the world.

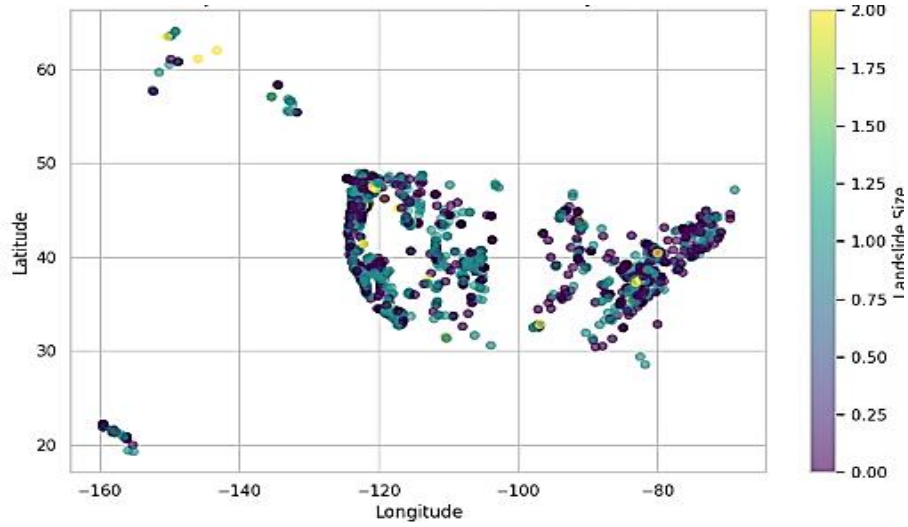


Figure 3: Landslide Size Distribution

4.5 Classification Performance

We evaluated classification performance with confusion matrices, accuracy scores and class-wise precision and recall metrics in a classification that separates small landslides from non-small landslides. When we trained models on the feature set the Support Vector Machine got the highest accuracy, at 58.3 percent. The Random Forest got an accuracy of 57.7 percent. The CatBoost got an accuracy of 55.6 percent. The XGBoost got an accuracy of 55.4 percent. We saw performance improvements when We trained on PCA-transformed features. SVM with PCA gave overall accuracy at 59.7 percent. The XGBoost with PCA followed at 59.2 percent.

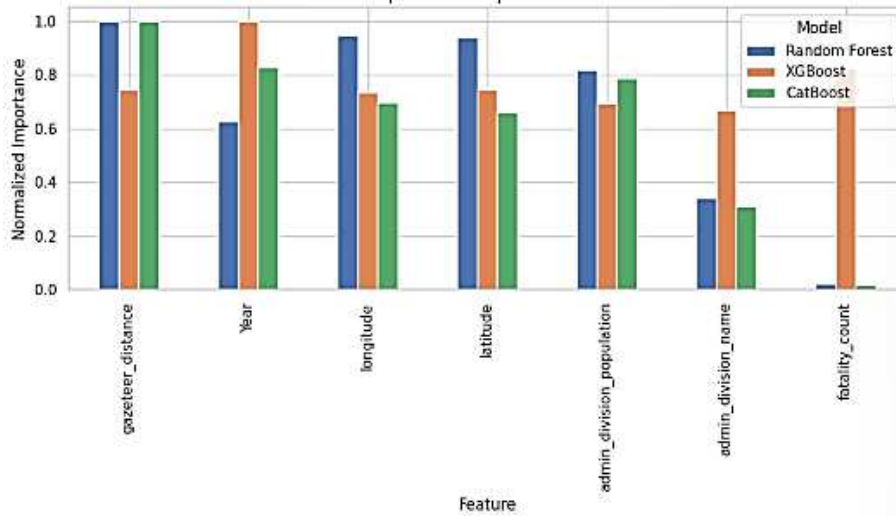


Figure 4: Feature Importance Comparison

The Random Forest with PCA was next at 58.6 percent. The CatBoost, with PCA slipped to 57.0 percent. We see that the confusion matrices show that the models have moderate ability to tell small landslides from non-small landslides. The classifications happen near the class boundaries. We see that the results show that the features, for the size categories, overlap a lot. We see that the results also point out the challenges of classifying landslide size when using the globally reported data (Fig. 5).

4.6 Effect of PCA on Model Performance

We looked at the numerical features with Principal Component Analysis (PCA). We wanted to see how the variant pattern of the data affects the classification performance. The scree plot and the cumulative variance plot show that 95 percent of the total variance is captured by the first six principal components. When We project the landslide size categories onto the two principal components We see a lot of overlap between the classes. The overlap is especially strong between the class and the medium class. The overlap shows that the classes are not well separated in space.

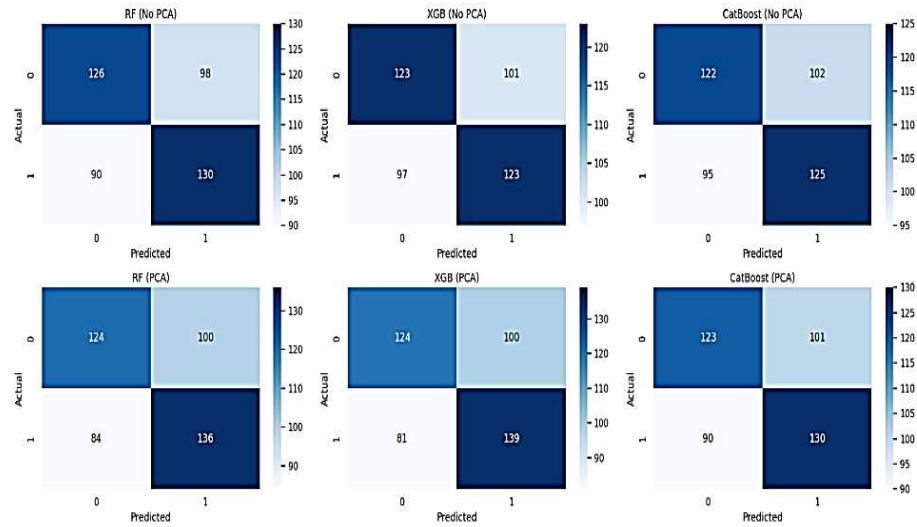


Figure 5: Confusion matrices for binary landslide size classification across machine learning model

This matches the observation. The models that used the PCA-transformed features did not do better than the models that used the feature set. The dimensionality reduction reduces the information, for the landslide size classification (Fig. 6). We think the study shows a pattern over time and space in the landslide activity that people report. We think the study shows that the landslide frequency rises over time. We think the study shows that the rise matches reporting, climate changes and more exposure in hazard-prone regions. We think the study shows that the landslide events cluster together in space and that the clustering points to the need to consider the context especially in areas, with complex terrain and high population density. We see that the time and location variables predict landslide size best. We notice that human impact measures such as fatality counts add little to prediction ability. With many classification models' overall accuracy stays moderate, at about 55 to 60 percent.

This shows class imbalance overlapping feature patterns and limited environmental data in the Global Landslide Catalog. We tried PCA to reduce dimensions. PCA does not really improve classification performance. PCA result shows the value of keeping the feature representations for the landslide size analysis. The models capture patterns in the landslide behaviour. The models also point out the limits of the landslide inventories, for detailed prediction.

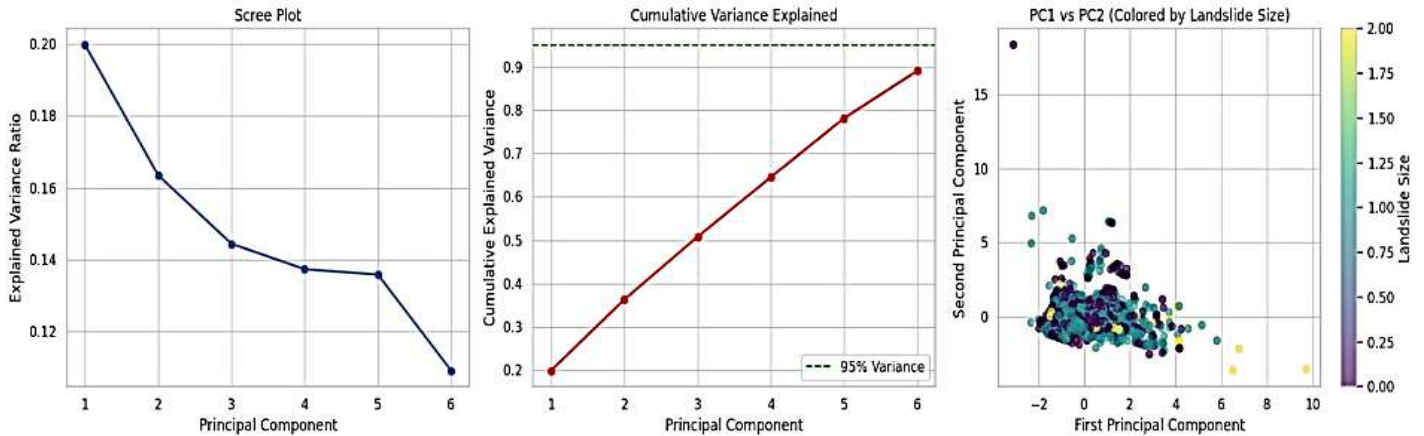


Figure 6: Principal Component Analysis (PCA) results showing (a) the scree plot of explained variance by principal component, (b) cumulative explained variance used to select retained components, and (c) the projection of landslide size categories onto the first two principal components (PC1 and PC2). Six components retain approximately 95% of the total variance, with substantial overlap observed between size classes.

5 Discussion

6 Conclusion

We looked at landslide occurrences using the Global Landslide Catalog. We focused on patterns, temporal trends and landslide size classification. The results show geographic clustering. The results also show a rise in reported landslide

frequency over time. The rise comes from the combined effect of exposure, terrain and reporting practices. We have seen that Machine learning models show that the spatial and temporal features give clues for the landslide size. We also notice that the class imbalance and a big overlap between the size categories limit the classification performance. We think moderate accuracy is average.

We have observed that most of the misclassifications happen between the size classes instead of between distinct events. We have tried Dimensionality reduction with PCA and Dimensionality reduction, with PCA does not improve the model performance by a lot. We conclude that the original feature representations keep the information. These findings highlight both the potential and limitations of data-driven approaches for global landslide analysis. While the Global Landslide Catalog enables large-scale comparative studies, incomplete reporting and the absence of detailed environmental variables limit predictive precision and causal interpretation. Future work should prioritize improved data completeness and the integration of environmental and socioeconomic drivers to strengthen global landslide risk assessment and modeling efforts.

Consent to Publish declaration: not applicable.

Consent to Participate declaration: not applicable.

Data Availability declaration: The data used in this study are publicly available from the NASA Global Landslide Catalog and can be accessed through the NASA Open Data Portal. The datasets analyzed during the current study are available from the corresponding public data source. <https://data.nasa.gov/dataset/global-landslide-catalog-export>

Author Contribution

Munish Kumar Mummadi, Write the Original Paper, Programming, Machine Learning Modeling

Harshil Sharma, Update the Original Paper, Programming, Predictive Modeling

Sean Kee, Literature Review, Data Curation, Visualization, Data Analysis

Khushi Chauhan, Literature Review, Data Curation, Data Analysis

Payam Norouzzadeh, Supervision, Methodology, Review the Paper

Eli Snir, Supervision, Validation, Review the Paper

Steven Buckner, Administration, Define the Project, Validation, Funding, Review the Paper

Bahareh Rahmani, Administration, Define the Project, Methodology, Validation, Review the Paper

Funding

Saint Louis University internal grant

Foundational Interdisciplinary Research Experience (FIRE)

References

- [1] Kirschbaum, D. B., Adler, R., Hong, Y., Lerner-Lam, A., & Huffman, J. J. (2012). Global distribution of extreme precipitation and high-impact landslides in 2010 relative to previous years. *Journal of Hydrometeorology*, 13(5), 1536–1551. <https://doi.org/10.1175/JHM-D-11-0102.1>
- [2] Kirschbaum, D. B., Stanley, T. A., & Zhou, Y. (2015). Spatial and temporal analysis of a global landslide catalog. *Geomorphology*, 249, 4–15. <https://doi.org/10.1016/j.geomorph.2015.03.016>
- [3] Kirschbaum, D. B., & Stanley, T. A. (2020). Developing a global landslide nowcasting system. *Natural Hazards and Earth System Sciences*, 20(7), 1985–2002. <https://doi.org/10.5194/nhess-20-1985-2020>
- [4] NASA Global Landslide Catalogue. (n.d.). NASA Open Data. <https://data.nasa.gov/dataset/global-landslide-catalog-export>
- [5] ArcGIS Online. Global Landslide Catalogue. <https://www.arcgis.com/home/item.html?id=b80a04dce63d4cc1bf081a6174ee2586>
- [6] Ado, M., et al. (2022). Landslide susceptibility mapping using machine learning: Trends, limitations, and future directions. *Remote Sensing*, 14(13), 3029. <https://www.mdpi.com/2072-4292/14/13/3029>
- [7] Youssef, A. M., et al. (2021). Evaluation of machine learning models for landslide susceptibility mapping. *Geoscience Frontiers*. <https://doi.org/10.1016/j.gsf.2020.05.010>
- [8] Tjahjono, B. T. (2024). Modelling landslide hazard using spatial data and machine learning tools. *International Journal of Remote Sensing*. <https://doi.org/10.1080/01431161.2023.0000000>

- [9] Hossain, M. I., et al. (2025). Landslide susceptibility mapping with Random Forest and MaxEnt. *Environmental Monitoring and Assessment*. <https://doi.org/10.1007/s43621-025-02084-X>
- [10] Deng, N., et al. (2022). Hybrid machine learning models for landslide susceptibility analysis. *Frontiers in Environmental Science*. <https://www.frontiersin.org/articles/10.3389/fenvs.2022.1009433/full>
- [11] Xu, Z., et al. (2024). Seismic landslide susceptibility assessment using PCA and SVM. *Scientific Reports*. <https://doi.org/10.1038/s41598-023-48196-0>
- [12] Li, L., et al. (2021). Landslide prediction using improved PCA and support vector regression. *Journal of Mountain Science*. <https://doi.org/10.1007/s11629-020-6396-5>
- [13] Sabokbar, H. F. (2014). Application of principal component analysis in landslide susceptibility modelling. *Geographical Information Sciences*. <https://doi.org/10.1080/13658816.2014.000000>
- [14] Nandi, A., & Shakoor, A. (2009). Multivariate principal component analysis for landslide susceptibility assessment. *Environmental Geology*, 58, 827–836. <https://doi.org/10.1007/s00254-009-1508-z>
- [15] Carrara, A., et al. (1991). Multivariate statistical analysis for landslide hazard zonation. *Geological Society of America Special Papers*, 267, 1–53.
- [16] Reichenbach, J., Rossi, M., Malamud, B. D., Mihir, M., & Guzzetti, F. (2018). Review of statistically-based landslide susceptibility models. *Earth-Science Reviews*, 180, 60–91. <https://doi.org/10.1016/j.earscirev.2018.03.001>
- [17] Malamud, B. D., Turcotte, D. L., Guzzetti, F., & Reichenbach, P. (2004). Landslide inventories and their statistical properties. *Earth Surface Processes and Landforms*, 29(6), 687–711. <https://doi.org/10.1002/esp.1064>
- [18] Guzzetti, F., Carrara, A., Cardinali, M., & Reichenbach, P. (1999). Landslide hazard evaluation: Review of current techniques and applications. *Geomorphology*, 31(1-4), 181–216. [https://doi.org/10.1016/S0169-555X\(99\)00078-1](https://doi.org/10.1016/S0169-555X(99)00078-1)
- [19] Shirzadi, A., et al. (2018). Hybrid machine learning techniques for landslide susceptibility mapping. *Sensors*, 18(3), 872. <https://doi.org/10.3390/s18030872>
- [20] Pham, B. T., et al. (2020). Ensemble machine learning models for landslide susceptibility analysis. *Advances in Space Research*, 66(3), 747–766. <https://doi.org/10.1016/j.asr.2020.05.025>
- [21] Inan, M. S. K., & Rahman, I. (2022). Explainable AI and feature selection in landslide susceptibility mapping. *arXiv:2201.03225*. <https://arxiv.org/abs/2201.03225>
- [22] Ganerød, A. J., et al. (2023). Ensemble deep learning for global landslide detection. *arXiv:2310.05959*. <https://arxiv.org/abs/2310.05959>
- [23] Lu, Z., et al. (2023). Iterative CNN-based landslide segmentation framework. *arXiv:2302.12420*. <https://arxiv.org/abs/2302.12420>
- [24] Guo, Z., et al. (2023). A Python system for regional landslide susceptibility assessment using machine learning. *PMC Articles*. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10660045/>
- [25] Ullah, I. (2025). Landslide susceptibility modelling using PCA and data mining techniques. *International Journal of Geo-Information*, Article ahead of print. <https://doi.org/10.1080/19475705.2025.2537221>