SHapley Additive exPlanations (SHAP) for Landslide Susceptibility Models: Shedding Light on Explainable AI

Husam Al-Najjar[a,*], Bahareh Kalantar[b,*], Biswajeet Pradhan[c], Ghassan Beydoun[a], Naonori Ueda[b]

[a] School of Computer Science, Faculty of Engineering and IT, University of Technology Sydney, 2007 Sydney, NSW, Australia,

[b] Disaster Resilience Science Team, RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan

[c] School of Civil and Environmental Engineering, Faculty of Engineering and IT, University of Technology Sydney, 2007 Sydney, NSW, Australia

*Corresponding authors: Husam.al-najjar@uts.edu.au; bahareh.kalantar@riken.jp

Keywords: Landslide susceptibility, SHAP, XAI, Machine learning, Risk mitigation

Abstract

This research examines the effectiveness of the SHapley Additive exPlanations (SHAP) approach in enhancing the interpretability of landslide susceptibility models. With the growing popularity of machine learning, we aim to understand how geoenvironmental and physically based factors impact modelling and to explain their interactions. The study focuses on the landslide-prone region of Bhutan and compares the performance of two approaches. The first approach incorporates geoenvironmental factors, while the second integrates both geoenvironmental factors and an additional physically based model. Random Forest (RF) algorithm is used to develop and compare these landslide susceptibility models. Various evaluation metrics, including overall accuracy and precision-recall, are employed to assess the predictive capabilities of each model. The findings reveal the strengths and limitations of both models, providing valuable insights for stakeholders and decision-makers involved in land use planning and disaster preparedness. Ultimately, this research seeks to advance landslide susceptibility modelling by highlighting the role of SHAP and its interaction with geoenvironmental and physically based factors, thereby contributing to more effective risk mitigation strategies.

1. Introduction

Interpretability and explainability are increasingly recognized as critical aspects of machine learning models, especially in high-impact fields like disaster risk management, where decisions can significantly impact lives and property. In landslide susceptibility modelling, these qualities are essential, as they allow stakeholders—such as land use planners and disaster management officials—to understand and trust the predictions generated by the models, enabling more informed and effective decision-making (Arrieta et al, 2020). Landslides pose a severe natural hazard, particularly in regions with complex terrains like Bhutan, where the risk is increasing (Froude, and Petley, 2018). Accurate prediction and effective management of landslide susceptibility are vital for mitigating risks and protecting communities (Bukhari et al, 2023). Traditional approaches to landslide susceptibility modelling often rely on statistical and physically based methods (Reichenbach et al, 2018; Guzzetti et al, 1999). While physically based models, which depend on detailed geotechnical data, are known for their explainability, they are also costly and resource-intensive to implement (Chen et al, 2015; Van Westen et al, 2008).

In contrast, machine learning methods have gained significance due to their ability to process large datasets, including remote sensing data from various sensors, such as optical, radar, and LiDAR, along with historical landslide data (Hussain et al, 2022). Despite their predictive power, these models often suffer from a "black box" problem, where their internal workings are not easily interpretable, posing a challenge to their practical use (Casalicchio et al, 2019; Maxwell et al, 2021). The growing field of eXplainable AI (XAI) offers solutions to this challenge by enhancing the transparency of complex models. Among the leading XAI techniques is SHapley Additive exPlanations (SHAP), which decomposes model predictions into contributions from individual features, providing valuable insights into the model’s decision-making process (Lundberg et al, 2017). Although SHAP has been widely applied across various domains, its application to landslide susceptibility modelling remains relatively unexplored yet promising, marking a novel contribution of this study (Guidotti et al, 2022; Binu et al, 2024; Arrogante-Funes et al, 2024).

This research explores how SHAP can enhance the interpretability of landslide susceptibility models in the challenging terrain of Bhutan, where accurate geotechnical data is limited. By comparing models that incorporate physically based methods with those that do not, the study assesses the additional value SHAP brings to different modelling scenarios (Meena and Hasija, 2022). Random Forest (RF) algorithm was utilized for its effectiveness with limited training data and its performance in various scenarios, including when the data is scarce (Biau and Scornet, 2016).

In this study, evaluation metrics such as overall accuracy, precision-recall, Kappa Index and the area under the receiver operating characteristic curve (AUC-ROC) are used to provide a comprehensive assessment of the models' predictive capabilities. The findings aim to lighten the strengths and limitations of integrating SHAP into landslide susceptibility models, offering critical insights for decision-makers involved in risk assessment and mitigation in complex terrains. Ultimately, this research sheds light on the role of SHAP in making landslide susceptibility models more transparent and interpretable (Samek et al, 2021), thereby supporting more informed and effective decision-making in disaster risk management.

2. Methodology

This research employed a machine learning approach, specifically RF, to develop and compare two landslide susceptibility models, each corresponding to a different setting.

The primary objective was to evaluate how SHAP can enhance the interpretability of these models (Samek et al, 2021), particularly in understanding the contribution of different factors to landslide susceptibility. The first model did not consider the physically based factors, whereas the second model incorporated a derived analysis of the physically based factor of safety (FOS). The Transient Rainfall Infiltration and Grid-Based Regional Slope-Stability (TRIGRS) model (Baum et al., 2008) was used to generate the physically based model, which includes critical parameters for assessing slope stability, such as geological/geomechanically data that account for variations in material properties and slope conditions. By integrating these factors, the second model seeks to better capture the physical processes influencing landslides, including soil cohesion, rainfall, and geological composition

2.1 Model Development

The RF was employed to predict landslide susceptibility for the two susceptibility mappings. The models were trained using a supervised learning approach, with landslide occurrence as the dependent variable. The first model primarily considers geoenvironmental factors, while the second model also uses physically based factors to assist in the training. The models were trained using a 70/30 data split, with hyperparameters tuned for optimizing predictive performance.

2.2 SHAP Analysis for Model Interpretation

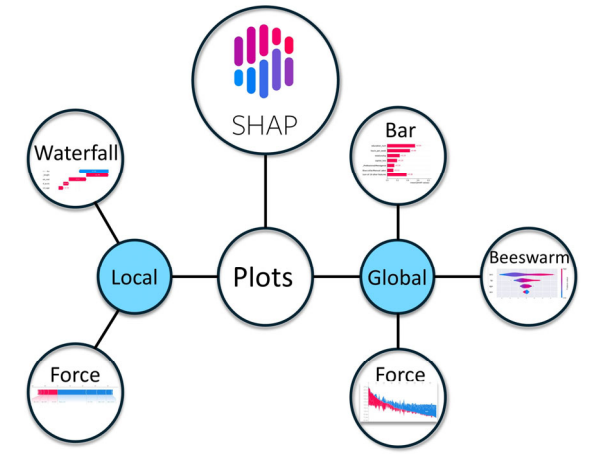
To gain insights into the contributions of individual features to the model's predictions, we employed SHAP (Lundberg and Lee, 2017), a game-theoretic approach for interpreting machine learning models. SHAP values quantify the impact of each feature on the model's output by attributing the difference between the prediction and the average prediction to the presence of a specific feature. This method provides both global and local interpretability, allowing us to understand not only the overall importance of features across the dataset but also how specific features influenced individual predictions (Lundberg and Lee, 2017). Based on Equation (1), the SHAP value represents the average marginal contribution of features (Kim and Kim, 2022).

$$\emptyset_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)] \quad (1)$$

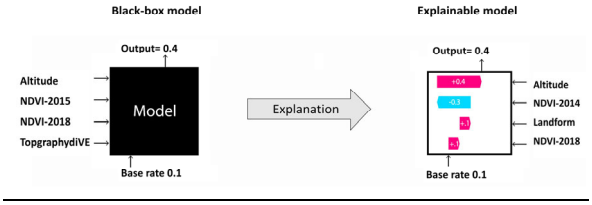
$$g(z') = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i z'_i \quad (2)$$

Here, N represents the complete set of features (38), and n refers to the total count of features in N, $N\backslash\{i\}$ denotes the set of all features in N excluding feature i. S indicates the number of subsets of N that do not include feature i, and $v(N)$ represents the baseline value, which is the predicted output for the features in N without any specific feature values. The model's output (as given in Equation 2) is obtained by summing the SHAP values for each feature (observation).
Furthermore, by defining M as the total number of input features, with $z' \epsilon \{0, 1\}M$ and $\emptyset_0$ representing the constant value when all inputs are absent, we analyzed the global impact of features using a summary plot that ranks feature importance along with their corresponding effects.

Figure 1(a) provides a broad overview of SHAP visualization techniques used for model interpretability, and Figure 1(b) demonstrates the process of converting a black-box model into an explainable one.



(a)



(b)

Figure 1. SHAP concept; (a) SHAP visualization techniques, (b) TreeSHAP concept for a landslide susceptibility model, comparing a black-box model with an explainable model. (adapted with modifications from: https://github.com/shap/shap)

SHAP analysis was applied to assess the importance of geoenvironmental and physically based factors in predicting landslide susceptibility. By visualizing SHAP values through Beeswarm plot, we were able to identify the most influential factors driving the model's predictions and gain a deeper understanding of the underlying mechanisms that contribute to landslide occurrences. This interpretability is crucial for validating the model's behaviour and ensuring that the predictions align with domain knowledge.

2.3 Model Evaluation and Analysis

The performance of the two models was evaluated using several metrics, including overall accuracy, precision, recall, Kappa Index and (AUC-ROC). Additionally, the interpretability of the models, as provided by the SHAP analysis, was assessed in terms of how well it could explain the model's predictions and highlight key factors contributing to landslide susceptibility. A comparative analysis was conducted to assess the impact of including the physically based model (FOS) in selecting landslide and non-landslide points for training. The study focused on evaluating both the predictive performance and interpretability of the models. More specifically, SHAP values were analyzed to determine which features had the greatest influence in each model and how the presence of the FOS affected the feature importance rankings. This comparison provided insights into the trade-offs between model complexity, predictive accuracy, and interpretability.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-81-2025 | © Author(s) 2025. CC BY 4.0 License.

3. Data and Study Area

3.1 Data Collection and Study Area

The study area is in a landslide-prone region of Bhutan, as shown in Figure 2, where 269 historical landslide occurrences were documented from 1998 to 2015. The dataset used in this study included a combination of remote sensing data, geotechnical information, and other relevant environmental factors known to influence landslide susceptibility. Remote sensing data was obtained from various sources, including optical (Landsat 8), radar sensors (ALOS PALSAR), while geotechnical data was gathered from field surveys and existing databases from Border Roads Organization of the Government of India (https://www.Bro.gov.in) and the National Center for Hydrology and Meteorology of the Royal Government of Bhutan (http://www.hydromet.gov.bt).

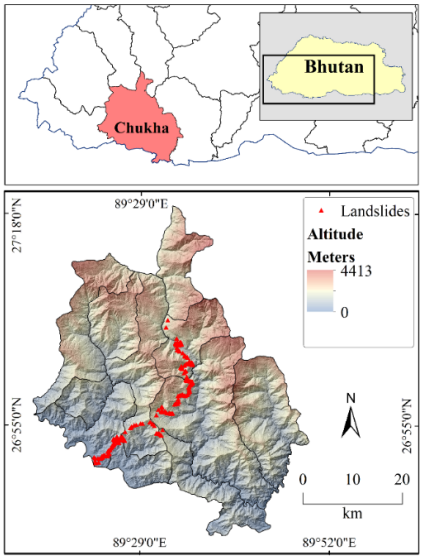


Figure 2. Study area, showing the altitude and landslide data points.

The data underwent preprocessing to ensure consistency and quality. This involved data cleaning, normalization, handling of missing values, and data balancing.
The factors considered in the models included 38 features, including landform, topographic indices (e.g., Topographic Diversity Index, mTPI), Synthetic Aperture Radar (SAR) data (from 2007 to 2015) with HH and VH polarizations, Normalized Difference Vegetation Index (NDVI) data (from 2013 to 2019), terrain attributes (e.g., Topographic Wetness Index (TWI), Sediment Transport Index (STI), total curvature, slope, slope length, aspect, and altitude), proximity factors (e.g., distance to streams), and lithological and geological characteristics.

FOS was also calculated using geotechnical parameters and incorporated as an additional input feature. The preprocessing of ALOS PALSAR was done based on ALOS data user’s handbook. The models were implemented using Python, with key libraries including scikit-learn for machine learning, SHAP for interpretability analysis, and ArcGIS for data processing and visualization.

4. Results and Discussion

Table 1 summarizes the performance metrics for the landslide susceptibility models with and without the incorporation of a physically based model. The model incorporating the physically based factor (FOS) shows slightly higher overall accuracy and precision, suggesting that including this factor improves the model's predictive capabilities. The ROC value also indicates better discrimination ability in the model with the physically based factor. Also, the increase in Kappa Index implies a better agreement and reliability in classification. Although the second model outperformed the first model, it has a slightly lower recall. This implies that the model may be slightly less sensitive to true positives, but this is minimal.

Metric	Without physically based model	Incorporating physically based model
F1-score	0.77	0.80
Precision	0.85	0.95
Recall	0.70	0.69
Kappa Index	0.72	0.76
ROC	0.94	0.95
Overall Accuracy	0.91	0.92

Table 1. Result of two models; (a) without physically based model; (b) Incorporating physically based model.

The SHAP analysis presented in Figure 3 provides insights into the importance and impact of various features on the model outputs for landslide susceptibility. It displays the top 20 contributing features among all features. It shows the SHAP values for two models: one without the physically based model (Figure 3a) and one incorporating the physically based model (Figure 3b). In both models, altitude emerges as the most influential feature, with consistently high SHAP values. This suggests that altitude plays a critical role in determining landslide susceptibility in the study area. Other significant features include NDVI from various years (e.g., NDVI-2014, NDVI-2018, NDVI-2015) and ALOS-Topography-derived Indices (ALOS-TpogrphydiVE and ALOS mTPI), indicating that vegetation cover and topographical attributes are also crucial in landslide prediction. When the physically based model is incorporated (Figure 3b), the feature rankings slightly change. Particularly, FOS and geology become more prominent, highlighting the additional value that physically based features bring to the model. The incorporation of physically based parameters seems to redistribute the importance among certain features. For instance, FOS appears as an important feature in the physically based model, which was not highlighted in the first model.

The SHAP plots also indicate the directionality (positive or negative) of the feature impacts. For example, in the case of altitude, the high SHAP values (in red) suggest that higher altitudes contribute positively to landslide susceptibility, whereas lower SHAP values (in blue) indicate a negative contribution. The variance in SHAP values for each feature suggests the interaction effects and non-linear relationships between the features and the model output. The spread of these values also indicates the robustness and consistency of each feature's contribution. The comparison between Figure 3a and 3b reveals that incorporating physically based models leads to a more diverse set of features being considered as important. This inclusion helps capture the complex interactions in the data, potentially leading to improved prediction accuracy. The model without the physically based inputs relies heavily on altitude and vegetation indices, while the model with the physically based

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-81-2025 | © Author(s) 2025. CC BY 4.0 License.

inputs considers a more balanced combination of topographical, geological, and vegetation features.

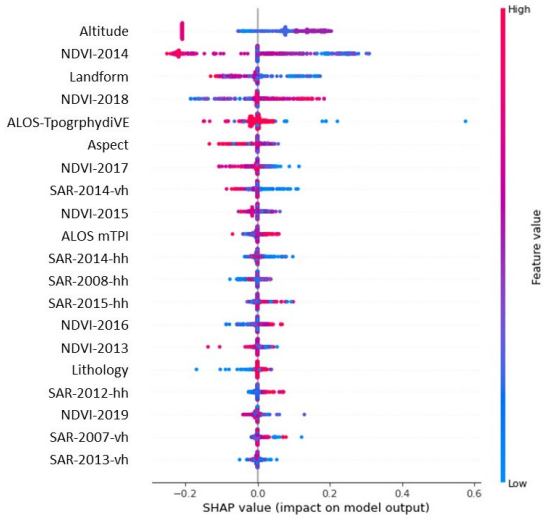
Despite slope being a known key factor in landslide susceptibility, the SHAP analysis did not highlight it as the most important feature. This could be due to multicollinearity with other topographic features (e.g., altitude, TWI, STI), which might capture similar information, reducing slope’s apparent importance. Additionally, interaction effects with other variables like soil type or vegetation could explain why slope’s contribution wasn’t as prominent. The influence of slope may also vary across different spatial scales; at larger regional scales, other factors, such as geology and vegetation, might dominate. Furthermore, the preprocessing steps and the model’s structure could have led to a stronger focus on variables like altitude and vegetation indices, which might have more direct relationships with landslide occurrence. Finally, the SHAP methodology, while powerful, prioritizes predictive accuracy and may not always align with theoretical expectations, particularly in complex, multi-factorial processes like landslides. Overall, this result emphasizes the complex, non-linear relationships in landslide susceptibility and suggests the need for further exploration of feature interactions and spatial variability in future models.

While the analysis provides a detailed understanding of feature contributions, the applicability of the model varies depending on the scale. Given SHAP's ability to explain the contribution of various features and their interactions, the model is particularly suited for regional and local beneficiaries, as it offers detailed, understandable insights into landslide susceptibility at these scales. At the global level, the model could serve as an initial screening tool or a guide for more localized studies, but it would require refinement for practical implementation in specific areas.

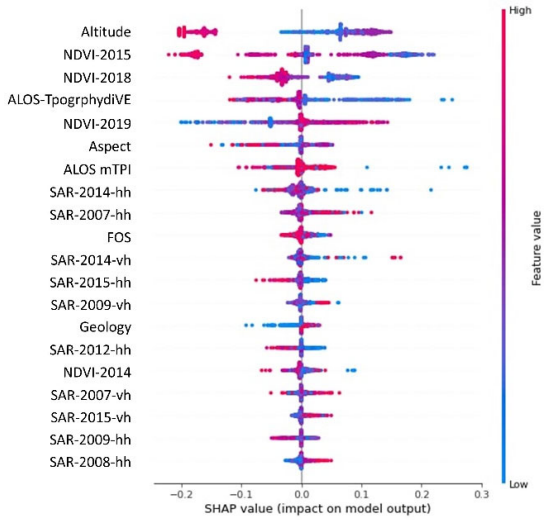
This study highlights SHAP’s role in translating machine learning insights into actionable strategies for landslide-prone regions. By identifying key risk factors, SHAP enhances land-use policies, enabling data-driven zoning and adaptive hazard mapping. It also refines early warning systems, improving alerts for at-risk communities and guiding public awareness campaigns. In urban planning, SHAP could inform resilient infrastructure design and supports GIS-based decision tools, ensuring new developments align with geospatial risk constraints. Furthermore, its explainability can aid in cost-effective risk mitigation, helping allocate resources efficiently. Ultimately, SHAP can bridge the gap between black-box models and real-world disaster resilience strategies, making landslide susceptibility assessments more transparent and actionable.

5. Conclusion

While incorporating physically based model enhances the model's overall ability to correctly identify and differentiate cases, it might come at the cost of being too conservative in predicting positives, which slightly lowers the recall. This trade-off is common in models that aim to reduce false positives but may inadvertently increase false negatives. Physically based factors alone may not capture all the essential or accurate information needed to fully predict landslide susceptibility, particularly at large scales. The integration approach can improve model performance by providing essential physical insights, but it must be combined with high-quality data and possibly other empirical or data-driven methods to ensure comprehensive and accurate predictions.
The slight decrease in recall in the second model might reflect these limitations, where the physically based factors could be missing or misrepresenting certain key aspects, leading to a more conservative prediction approach. When data are scarce or historical records are limited, the inclusion of a physically based model might be considered an additional step to enhance model robustness, despite the potential for increased processing demands. Ultimately, this research contributes to the advancement of landslide susceptibility modelling by clarifying the role of SHAP in conjunction with physically based models, offering valuable insights for stakeholders and decision-makers in land use planning and disaster preparedness.



(a)



(b)

Figure 3. Results of SHAP analysis for the two models: (a) without the physically based model; (b) incorporating the physically based model.

Acknowledgements

The authors gratefully acknowledge the Faculty of Engineering and Information Technology at the University of Technology Sydney for providing the essential facilities that supported this research. Special thanks are extended to the RIKEN Centre for Advanced Intelligence Project (AIP) in Tokyo, Japan, for their

valuable fund support. The authors also wish to express their appreciation to the Border Roads Organisation, Government of India (http://www.bro.gov.in), for facilitating the collection of landslide inventories through the Project DANTAK framework. Furthermore, we sincerely thank the National Center for Hydrology and Meteorology of the Royal Government of Bhutan (http://www.hydromet.gov.bt) and Prof. Raju Sarkar for their invaluable contribution in providing the data used in this study.

References

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115.

Arrogante-Funes, P., Bruzón, A. G., Álvarez-Ripado, A., Arrogante-Funes, F., Martín-González, F., & Novillo, C. J. (2024). Assessment of the regeneration of landslides areas using unsupervised and supervised methods and explainable machine learning models. Landslides, 21(2), 275-290.

Baum, R. L., Savage, W. Z., & Godt, J. W. (2008). TRIGRS: a Fortran program for transient rainfall infiltration and grid-based regional slope-stability analysis, version 2.0 (p. 75). Reston, VA, USA: US Geological Survey.

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

Binu, K. E., Anoopkumar, L. T., Sunil, M., Jose, M., & Preetha, K. G. (2024, June). Dynamic Landslide Prediction, Monitoring, and Early Warning with Explainable AI: A Comprehensive Approach. In 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 960-965). IEEE.

Bukhari, M. H., da Silva, P. F., Pilz, J., Istanbulluoglu, E., Görüm, T., Lee, J., ... & Haque, U. (2023). Community perceptions of landslide risk and susceptibility: a multi-country study. Landslides, 20(6), 1321-1334.

Casalicchio, G., Molnar, C., & Bischl, B. (2019). Visualizing the feature importance for black box models. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18 (pp. 655-670). Springer International Publishing.

Chen, H. X., Zhang, L. M., Gao, L., Zhu, H., & Zhang, S. (2015). Presenting regional shallow landslide movement on three-dimensional digital terrain. Engineering geology, 195, 122-134.

Froude, M. J., & Petley, D. N. (2018). Global fatal landslide occurrence from 2004 to 2016. Natural Hazards and Earth System Sciences, 18(8), 2161-2181.

Guidotti, R., Monreale, A., Ruggieri, S., Naretto, F., Turini, F., Pedreschi, D., & Giannotti, F. (2022). Stable and actionable explanations of black-box models through factual and counterfactual rules. Data Mining and Knowledge Discovery, 1-38.

Guzzetti, F., Carrara, A., Cardinali, M., & Reichenbach, P. (1999). Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy. Geomorphology, 31(1-4), 181-216.

Hussain, Y., Schlögel, R., Innocenti, A., Hamza, O., Iannucci, R., Martino, S., & Havenith, H. B. (2022). Review on the geophysical and UAV-based methods applied to landslides. Remote Sensing, 14(18), 4564.

Kim, Yesuel, Kim, Youngchul, 2022. Explainable heat-related mortality with random forest and SHapley Additive exPlanations (SHAP) models. Sustain. Cities Soc. 79, 103677.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Maxwell, A. E., Sharma, M., & Donaldson, K. A. (2021). Explainable boosting machines for slope failure spatial predictive modeling. Remote Sensing, 13(24), 4991.

Meena, J., & Hasija, Y. (2022). Application of explainable artificial intelligence in the identification of Squamous Cell Carcinoma biomarkers. Computers in Biology and Medicine, 146, 105505.

Reichenbach, P., Rossi, M., Malamud, B. D., Mihir, M., & Guzzetti, F. (2018). A review of statistically-based landslide susceptibility models. Earth-science reviews, 180, 60-91.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE, 109(3), 247-278.

Van Westen, C. J., Castellanos, E., & Kuriakose, S. L. (2008). Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview. Engineering geology, 102(3-4), 112-131.