Proc. of the 9th International Symposium on Geotechnical Safety and Risk (ISGSR)
Edited by Zhongqiang Liu, Jian Dai and Kate Robinson
Published by Research Publishing, Singapore.
doi: 10.3850/GRF-25280825_isgsr-039-P055-cd

ISGSR2025

ENHANCED LANDSLIDE SUSCEPTIBILITY MAPPING ALONG HIGHWAYS USING PROGRESSIVE TREE-BASED ENSEMBLE MODELS WITH OPTIMAL NON-LANDSLIDES SELECTION

Zilin Xiang[1], Jie Dou[1], Ke Xing[1], Alessio Domenico Leto[2], Luca Simoni[2], Lorenzo Brezzi[2,*]

[1] *Badong National Observation and Research Station of Geohazards, China University of Geosciences, Lumo Road 388, 430074, Wuhan, China.*

[2] *Department of Geosciences, University of Padova, via Giovanni Gradenigo 6, 35131 Padova, Italy.*

[3] *Department ICEA, University of Padova, via Ognissanti 39, 35129 Padova, Italy.*

E-mail: ZX: zilinxiang@cug.edu.cn, JD: doujie@cug.edu.cn, KX: xingke0306@163.com AL: alessiodomenico.leto@unipd.it
LS: luca.simoni.1@studenti.unipd.it, LB: lorenzo.brezzi@unipd.it

Abstracts: Highways serve as crucial links between regions, enhancing transportation and facilitating economic activities within communities. However, strategic infrastructure such as bridges and tunnels traversing steep mountainous regions is often susceptible to landslides, resulting in damages and casualties. The geological conditions in the Calabria region of south Italy frequently led to landslide disasters, posing a significant threat to regular operations. Accurate and rapid landslide susceptibility mapping (LSM) is essential for analyzing and evaluating the degree of landslide susceptibility, enabling the prediction of severe risks that could affect strategic infrastructure such as highways. The LSM based on machine learning (ML) models exhibits significantly higher accuracy than those using traditional expert knowledge and conventional mathematical statistics models. The proper and reasonable selection of non-landslide samples in ML models significantly enhances the prediction accuracy and reliability of the regional LSM. Therefore, our work is to analyze the spatial distribution pattern between predisposing factors and landslide occurrences in the selected area. Subsequently, we utilize progressive tree-based ML models, such as decision tree (DT), random forest (RF), gradient boosting decision tree (GBDT), and their information value method (IVM)-enhanced models (i.e., IVM-DT, IVM-RF, and IVM-GBDT), to conduct the LSM. These techniques generated a classification in terms of very low, low, medium, high, and very high landslide susceptibility zones. The receiver operating characteristic (ROC) curve will be used to quantitatively evaluate and compare the prediction accuracy of these models. The outcomes of this work have the potential to facilitate rapid and precise evaluations for effective hazard management and provide theoretical guidance for the maintenance and normal operation of bridges, tunnels, and other susceptible structures along highways in the Calabria region.

Keywords: highway-landslide interaction, progressive tree-based ensemble machine learning (ML) models, landslide susceptibility mapping (LSM), spatial distribution pattern, smart risk assessment, AI-driven susceptibility map.

1. Introduction

Landslide occurrences in highway areas are becoming increasingly frequent, posing significant threats to public safety, property, and national economic development. Italy is one of the most landslide-prone regions in the world, where extreme conditions often exacerbate the impact of landslides, particularly in mountainous areas characterized by complex terrain and diverse geological environments (Brezzi et al., 2021; Xiang et al., 2023). Therefore, identifying landslide-prone areas along highways in typical regions of Italy is crucial for safeguarding lives, property, and the safety of critical infrastructure.

Landslide susceptibility mapping (LSM) involves investigating various factors influencing landslide occurrence, such as geology, topography, climate, and land use, in conjunction with historical landslide distribution data. By employing statistical analysis, geographic information system (GIS) techniques, and model development methods, this process evaluates the likelihood and spatial distribution of landslides within a given region. Advanced machine learning (ML) approaches like random forest (RF), decision tree (DT), and support vector machine (SVM), have been increasingly employed in LSM due to their lower statistical limitations and automatic inductive features (Dou et al., 2019). For example, Dou et al. (2019) utilized DT and RF models to conduct LSM for rainfall-induced landslides on Izu Oshima Island, Japan. The results demonstrated that the predictive performance of the RF model was superior to that of the DT model. Huang et al. (2020) evaluated the impact of different landslide boundary representations (points, circles, and polygons) on LSM using SVM and RF models. Results show that polygon-based models significantly improve LSM accuracy and reduce uncertainties, with RF outperforming SVM overall. A pivotal step in ML-based LSM is the precise collection of positive (landslide) and negative (non-landslide) samples, as the quality and balance of these datasets play a crucial role in determining the predictive accuracy of the models. Dou et al. (2020) highlighted that the careful selection of non-landslide samples is crucial in LSM to prevent overestimating landslide risks. Building on this, Hu et al. (2020) proposed a hybrid approach that integrated fractal theory with ML models for selecting these samples, achieving highly accurate susceptibility maps. Therefore, optimizing strategies for non-landslide sample selection is vital to enhance the accuracy and reliability of LSM models. This study focuses on Calabria, southern Italy, using comprehensive data, including a detailed landslide inventory, lithological information, DEMs, road networks, and hydrographic data. Totally, eight predisposing factors were identified, including the specific influence of distance to roads, as road construction can destabilize slopes by undermining their base. Using the Natural Breaks method, roads were classified into 9 distance categories, and hazard levels were predicted considering, among other factors, the slope proximity to roads. Advanced tree-based ML algorithms (DT, RF, GBDT) and their IVM-enhanced variants (IVM-DT, IVM-RF, IVM-GBDT) evaluated landslide susceptibility, comparing model performance and optimizing negative sample conditions. The results offer theoretical support for mitigating landslide risks along highways, with reliable susceptibility maps being key to risk assessment.

2. Case study

The Calabria region, located in southern Italy at the junction between the southern Apennines and the Mediterranean geological domain, is primarily shaped by the Apennine-Maghrebian orogenic belt and the complex geodynamics of the Calabrian Arc. The area is characterized by well-developed thrust and normal faults. Approximately 42% of the region consists of mountainous terrain, 49% is hilly, and only 9% comprises plains, with pronounced elevation differences. The geology of the region is very complex and heterogeneous, dominated by metamorphic, sedimentary, and volcanic rocks. The climate of Calabria is influenced by both the sea and mountainous topography, leading to higher rainfall on the western coast compared to the eastern coast, especially during winter and autumn, with relatively low precipitation in summer. The complex geological structures and rainfall patterns make this region one of the most landslide-prone areas in Italy. The selected study area, spanning from 16°5′0″E to 16°23′50″E and 38°56′20″N to 39°11′10″N, encompasses a total of 842 landslides distributed across the region.

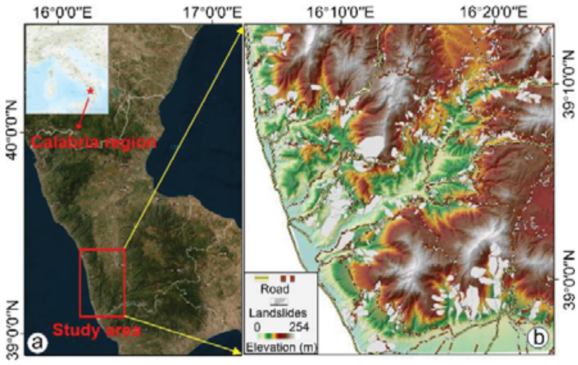

Fig. 1. Overview of the study area

3. Data collection and methods

This study extensively collected data for the Calabria region, including a landslide inventory, a 1:250,000 geological map, road network data, hydrographic data, and high-resolution DEM (with a 10 m resolution). In addition, based on the data obtained, we selected geological and environmental factors related to landslide occurrence, including elevation, slope, roughness, aspect, lithology, profile curvature, distance to roads, and distance to water. All data was downloaded from freely accessible Italian online portals (https://tinitaly.pi.ingv.it/ http://www.pcn.minambiente.it/ https://www.isprambiente.gov.it/). In the current study, both progressive ML models (DT, RF, and GBDT) and ensemble ML models (IVM-DT, IVM-RF, and IVM-GBDT) were used to perform the LSM. DT: A DT model features a hierarchical structure resembling a tree and functions as a non-parametric supervised learning approach. It facilitates the identification of complex, non-linear, and non-additive relationships between independent and target variables (Dou et al., 2019). Its core principle lies in constructing a predictive framework that estimates the dependent variable through decision rules generated from data-driven analysis. RF: RF is an ensemble learning method that constructs numerous individual DTs using randomly selected subsets of data to achieve classification and prediction tasks. The underlying principle of the RF algorithm is to create multiple DTs on randomly sampled datasets, enhancing robustness. This approach mitigates the risk of overfitting by relying on the aggregated outcomes of the ensemble (Zhou et al., 2021). GBDT: GBDT is an iterative ensemble approach that combines multiple DTs to improve the performance of a single predictive model by leveraging the strengths of multiple learners. By incorporating a regularization function, the GBDT algorithm effectively mitigates overfitting associated with traditional DT models, refining the training outcomes. Through successive iterations, it builds a series of weak classifiers, primarily in the form of classification and regression trees, to achieve enhanced predictive accuracy (Chen et al., 2020). IVM: The IVM is a probabilistic statistical approach for landslide risk assessment. It quantifies the contribution of various influencing factors to landslide occurrence by analyzing the spatial correlation between factor categories and the distribution of landslides.

The core of the method lies in calculating the ratio of the landslide occurrence probability for each category to the overall probability, which is then transformed logarithmically to obtain the information value.

Finally, the IV values of all influencing factors are aggregated to produce a landslide susceptibility map (Chen et al., 2020). Model Evaluation and Assessment: The receiver operating characteristic (ROC) curve serves as a robust tool for evaluating model performance by measuring predictive accuracy across various probability thresholds. The area under the ROC curve (AUC) is widely regarded as a key metric for model validation, with values ranging between 0.5 and 1. A higher AUC reflects superior model efficiency (Zhou et al., 2021). Therefore, this study utilized AUC as the primary criterion to assess the effectiveness of the LSM models.

4. Results

4.1. *IVM-based Non-landslide Samples Selection*

The area was subdivided into 901,967 grid cells, each with a spatial resolution of 10 m. Within the overlapping region, 842 landslide polygons were transformed into 114,857 grid cells for the IVM model. The model was subsequently applied to all grid cells to conduct LSM and generate a preliminary susceptibility map (Fig. 2a). Using the Jenks natural breaks classification method in ArcGIS 10.8, 5 susceptibility categories were delineated: very low (5.22%), low (17.13%), moderate (21.03%), high (14.22%), and very high (42.40%). To address the issue of unreasonable non-landslide samples, 842 non-landslide grid cells were randomly selected from the very low and low susceptibility zones, ensuring a balance with the number of landslide samples (Fig. 2b).

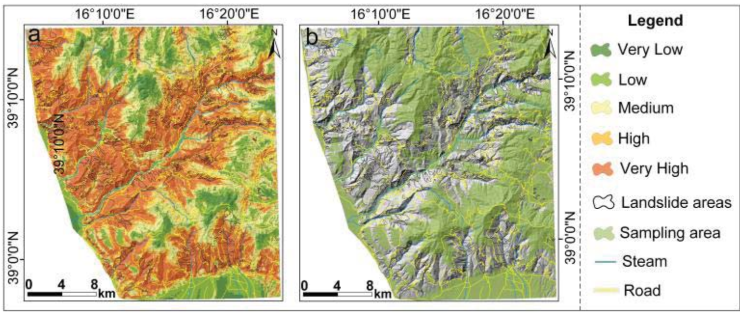

Fig. 2.LSM based on IVM: preliminary susceptibility map (a) and non-landslide sampling area (b)

4.2. *LSM based on ML and ensemble models*

In this section, progressive ML models (DT, RF, and GBDT) and ensemble models (IVM-DT, IVM-FR, and IVM -GBDT) were designed, generated, and implemented to obtain susceptibility maps. Both the progressive ML models and the ensemble models utilized 70% of the landslide grid cells for training and 30% for validation. Notably, equal numbers of non-landslide samples for individual models were randomly selected across the entire study area. For ensemble models, non-landslide samples were specifically chosen from very low and low susceptibility zones, as delineated in the initial IVM model-based susceptibility maps. This methodology enabled a comparative evaluation of susceptibility outcomes based on different non-landslide sample selection strategies.

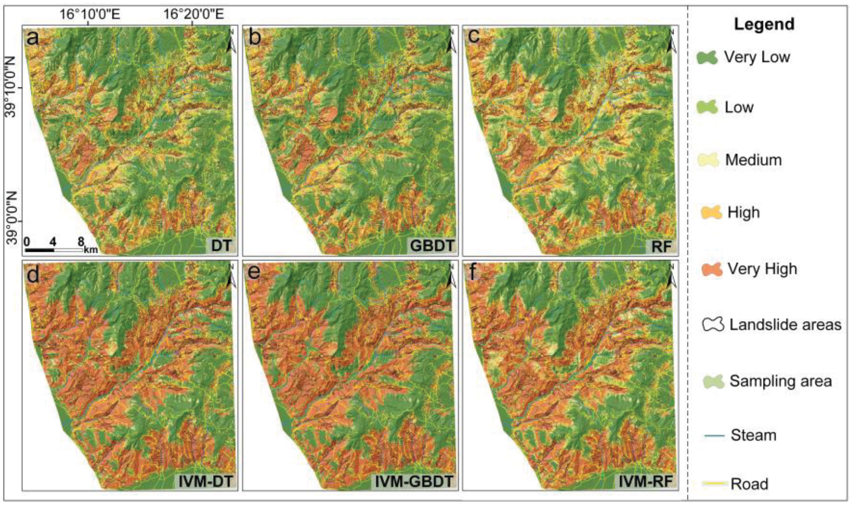

Fig. 3. Landslide susceptibility maps **generated** by DT, GBDT, RF, IVM-DT, IVM-GBDT and IVM-RF

The resulting susceptibility maps were categorized into five classes, from very low to very high (Fig. 3). Overall, the progressive ML models exhibit similar characteristics in the distribution of high-susceptibility zones, which are predominantly located in the higher-elevation areas of the western region and along both sides of the waterways in the northeastern region. In comparison, the ensemble models predicted a larger area of high-susceptibility zones; however, the overall distribution pattern of susceptibility remains consistent with the predictions of the individual models. Additionally, the AUC values indicated that the IVM-RF model achieved the highest predictive accuracy of 0.97, followed by IVM-GBDT with 0.96. The predictive accuracies of other models were as follows: IVM-DT (0.91) > GBDT (0.89) > RF (0.82) > DT (0.80).

5. Discussion and conclusion

The influence of various predisposing factors on regional landslide susceptibility models (LSM) varies significantly, as evidenced by the differing performance levels of individual models. To assess the relative contributions of these factors in the Calabria region, the average feature importance values were calculated across six distinct models. Figure 4 highlights that elevation (0.321) has the highest overall importance for landslide occurrence, although its influence varies considerably among the models. Slope (0.184) emerges as the second most influential factor, followed by lithology (0.136) and distance to roads (0.114), both of which show lower variability in importance across the models. Furthermore, the influence of distance to water (0.08), aspect (0.071), profile curvature (0.052), and terrain roughness (0.042) on landslide occurrence progressively decreases. These results emphasize the dominant role of topographic and geological factors, while also recognizing the contribution of anthropogenic elements like road networks in triggering landslides in the Calabria region.

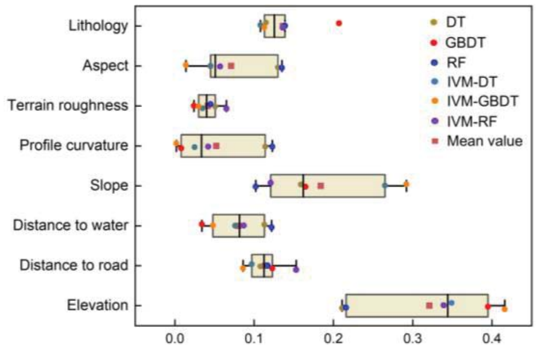

Fig. 4. Boxplots that express the evaluation of the importance of predisposing factors for the different models

In conclusion, this study addressed the challenge of subjectively generating non-landslide samples by implementing optimization strategies aimed at enhancing the overall performance of LSM. The progressive ML models (DT, RF, and GBDT) and the ensemble models (IVM-DT, IVM-RF, and IVM-GBDT) exhibit similar patterns in the distribution of high-susceptibility zones. These zones are predominantly concentrated in higher-elevation areas of the western region and along both sides of the waterways in the northeastern region. The LSM results indicate that the IVM-RF model achieved the highest predictive accuracy of 0.97, followed by IVM-GBDT with 0.96. The average feature importance values indicate that elevation, slope, and lithology are the most significant factors influencing landslide occurrence. Additionally, the impact of roads is also notable in the Calabria region. The development of a reliable landslide susceptibility map is essential for accurate risk assessment, providing critical spatial insights into potential hazard zones and informing mitigation strategies.

Acknowledgement

This study was carried out within the RETURN Extended Partnership and received funding from the European Union Next-GenerationEU (National Recovery and Resilience Plan – NRRP, Miss. 4, Comp. 2, Inv. 1.3 – D.D. 1243 2/8/22, PE0000005)

Reference

Brezzi, L., Carraro, E., Pasa, D., Teza, G., Cola, S., & Galgaro, A. (2021). Post-collapse evolution of a rapid landslide from sequential analysis with FE and SPH-based models. Geosciences, 11(9). https://doi.org/10.3390/geosciences11090364

Chen, T., et al. (2020). Mapping landslide susceptibility at the Three Gorges Reservoir, China, using gradient boosting decision tree, random forest and information value models. Journal of Mountain Science, 17(3), 670–685. https://doi.org/10.1007/s11629-019-5839-3

Dou, J., et al. (2019). Evaluating Gis-based multiple statistical models and data mining for earthquake and rainfall-induced landslide susceptibility using the lidar dem. Remote Sensing, 11(6). https://doi.org/10.3390/rs11060638

Dou, J., Yunus, A. P., Merghadi, A., Shirzadi, A., Nguyen, H., Hussain, Y., Avtar, R., Chen, Y., Pham, B. T., & Yamagishi, H. (2020). Different sampling strategies for predicting landslide susceptibilities are deemed less consequential with deep learning. Science of the Total Environment, 720. https://doi.org/10.1016/j.scitotenv.2020.137320

Hu, Q., Zhou, Y., Wang, S., & Wang, F. (2020). Machine learning and fractal theory models for landslide susceptibility mapping: Case study from the Jinsha River Basin. Geomorphology, 351. https://doi.org/10.1016/j.geomorph.2019.106975

Huang, F., Cao, Z., Jiang, S. H., Zhou, C., Huang, J., & Guo, Z. (2020). Landslide susceptibility prediction based on a semi-supervised multiple-layer perceptron model. Landslides, 17(12), 2919–2930. https://doi.org/10.1007/s10346-020-01473-9

Xiang, Z., Dou, J., Yunus, A. P., Zhang, L., Wang, X., & Luo, W. (2023). Vegetation-landslide nexus and topographic changes post the 2004 Mw 6.6 Chuetsu earthquake. Catena, 223. https://doi.org/10.1016/j.catena.2023.106946

Zhou, X., Wen, H., Zhang, Y., Xu, J., & Zhang, W. (2021). Landslide susceptibility mapping using hybrid random forest with GeoDetector and RFE for factor optimization. Geoscience Frontiers, 12(5). https://doi.org/10.1016/j.gsf.2021.101211