

Influence of DEM Resolution on Geomorphometric Representation, Explainable Artificial Intelligence and Temporal Transferability of Machine Learning Models for Landslide Susceptibility Mapping in the Nepal Himalaya

Nirmal Kafle^{1,2*}, Dinesh Pathak² and Ranjan Kumar Dahal²

¹Khwopa College of Engineering, Libali, Bhaktapur, Nepal

²Central Department of Geology, Tribhuvan University, Kirtipur, Kathmandu, Nepal

(*Corresponding E-mail: nirmalkafle1917@gmail.com)

Abstract: Landslide susceptibility mapping (LSM) is essential for risk mitigation in the Nepal Himalaya, yet the effects of digital elevation model (DEM) resolution on geomorphometric representation, model discrimination, explainability, and temporal transferability remain insufficiently resolved. This study evaluates 5, 12.5, 30, and 90 m DEMs using Random Forest (RF) and Extreme Gradient boosting (XGBoost) with spatially grouped 10-fold cross-validation. The modelling dataset comprised 54,952 samples from 486 landslide polygons (1955–2010), with temporal validation using 135 landslides (2011–2025). Results show resolution sensitivity in topographic ruggedness index (TRI), stream power index (SPI), slope, relief, and elevation, whereas plan and profile curvature were robust. Mean spatial cross-validated AUROC remained stable (0.779–0.792), with no significant resolution effect (Friedman $\chi^2 = 2.76$, $p = 0.430$), but threshold-dependent metrics degraded at 90 m. SHapley Additive exPlanations (SHAP) showed stable feature rankings dominated by rainfall, elevation, lithology, and distance to roads, with rainfall–lithology as the strongest interaction. Multi-criteria scoring placed 12.5–30 m within the preferred range. Temporal validation preserved susceptibility ranking (AUROC 0.809–0.846) but revealed under-confidence, with Youden-optimal thresholds of 0.13–0.17. The results indicate that DEM coarsening changes terrain representation more strongly than environmental relationships learned by ensemble models, while probability calibration requires temporal updating.

Keywords: Landslide susceptibility mapping, DEM resolution, Random forest, XGBoost, SHAP, Temporal transferability, Nepal Himalaya, Geomorphometric analysis

Introduction

Landslides are recurrent hazards in the Nepal Himalaya because of steep terrain, fragile geology, intense monsoon rainfall, active tectonics, and infrastructure development. Ensemble machine-learning methods such as random forest (RF) and extreme gradient boosting (XGBoost) are widely used in landslide susceptibility mapping (LSM) because they capture non-linear relationships among conditioning factors (Merghadi et al., 2020; Usta et al., 2024). Yet digital elevation model (DEM) resolution is a fundamental modelling choice because many terrain predictors are derived directly from elevation data, and differences in

input data quality and resolution can affect susceptibility results (Gaidzik & Ramírez-Herrera, 2021).

Previous studies report scale-dependent and sometimes inconsistent effects of DEM resolution on susceptibility modelling (Chang et al., 2019; Wubalem, 2022; Yan et al., 2023; Zhao et al., 2024). Spatially random validation may also overestimate performance where observations are spatially autocorrelated, while temporal validation remains comparatively underused (Roberts et al., 2017; Nocentini et al., 2024). This study therefore evaluates the full resolution-to-transferability chain.

Study area and methods

The study area is an axis-aligned rectangular corridor surrounding the eastern segment of the national highway (NH3) within the Okhaldhunga tectonic window, eastern Nepal. A 20 km buffer on either side of NH3 captures highway and feeder-road influences. Elevation ranges from about 200 to >2,500 m, median slope is 26–33°, and annual rainfall is approximately 1,200–2,000 mm, with about 80% falling during June–September. The setting comprises structurally complex Lesser Himalayan terrain and recurrent slope instability. The spatial extent and temporal separation of the landslide inventories are shown in Figure 1.

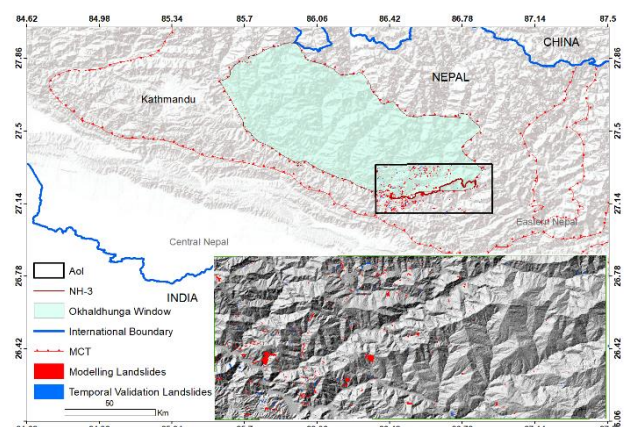


Figure 1, Study area and landslide inventories used for model development (1955–2010) and independent temporal validation (2011–2025).

Four independently sourced DEMs were analysed: 5 m from the Department of Survey, Nepal; 12.5 m ALOS PALSAR; 30 m SRTM GL1; and 90 m SRTM GL3. Twenty-one conditioning factors were prepared at every resolution, comprising geomorphometric, hydrological, climatic, vegetation, geological, land-cover and infrastructure variables. The inventory included 486 landslides for model development (1955–2010) and 135 later landslides for independent temporal validation (2011–2025). Positive observations were generated by area-dependent landslide sampling; stable observations were drawn outside a 50 m exclusion buffer. The final modelling dataset comprised 54,952 observations, while the temporal holdout comprised 3,480 balanced points (1,740 landslide and 1,740 stable).

RF and XGBoost were trained separately at each resolution using the same spatially grouped 10-fold partitions, with all observations from a landslide cluster constrained to a single fold. Area under receiver operating characteristics (AUROC) and area under the precision recall curve (AUPRC) assessed threshold-independent discrimination, while F1-score and Cohen's Kappa were evaluated at a 0.50 cutoff. A Friedman test was applied to matched fold-level AUROC values, with post-hoc Wilcoxon tests where required (Friedman, 1937). Permutation importance, recursive feature elimination cross validation (RFECV), and TreeSHAP evaluated feature relevance and interpretability (Lundberg and Lee, 2017). Predictor relevance was examined with out-of-fold permutation importance RFECV. SHapley Additive exPlanations (SHAP) was used to assess global importance, non-linear dependence and pairwise interactions. For temporal transfer, the historical models were applied without refitting to the 2011–2025 holdout; RF and XGBoost probabilities were also averaged into an ensemble. Operational DEM selection used a six-criterion composite score combining predictive performance, threshold robustness, explainability stability, model parsimony, data quality and processing efficiency. Youden's J statistic was used to identify operating thresholds on the temporal holdout (Youden, 1950).

Results and discussion

Geomorphometric response and predictor redundancy

DEM coarsening produced derivative-specific changes. Median topographic ruggedness index (TRI) changed from 5.978 at 5 m to 0.494 at 90 m (–91.73%), stream power index (SPI) declined from 5.796 to 1.479 (–74.48%), and median slope decreased from 32.79° to 26.42° (–19.44%). Plan and profile curvature were comparatively robust, while topographic wetness index

(TWI) and topographic position index (TPI) were statistically sensitive but had negligible effect sizes. Strong predictor redundancy increased from five $|p| \geq 0.70$ pairs at 5 m to 12 at 90 m; Slope–TRI reached $|p| = 0.994$ at 30 m. The cross-resolution model performance metrics are summarized in Table 1.

Table 1, Spatially grouped cross-validation performance across DEM resolutions.

Resolution	RF AUROC	RF F1	RF κ	XGB AUROC	XGB F1	XGB κ
5 m	0.785	0.565	0.322	0.781	0.611	0.342
12.5 m	0.791	0.533	0.298	0.789	0.601	0.333
30 m	0.792	0.517	0.287	0.785	0.599	0.331
90 m	0.779	0.398	0.208	0.779	0.576	0.314

Model performance and explainability

Mean spatial AUROC ranged from 0.779 to 0.792. RF reached 0.792 at 30 m and XGBoost 0.789 at 12.5 m, but the Friedman test found no significant resolution effect (Chi-square, $\chi^2 = 2.76$, $p = 0.430$). In contrast, RF F1-score declined from 0.565 at 5 m to 0.398 at 90 m and Kappa from 0.322 to 0.208. Thus, stable discrimination did not imply stable probability calibration.

SHAP and permutation-importance analyses showed strong cross-resolution agreement. Rainfall, elevation, lithology, and distance to roads consistently ranked highest; Kendall's W (coefficient of concordance) was 0.960 for RF and 0.946 for XGBoost ($p < 1.4 \times 10^{-8}$). Rainfall–lithology remained the strongest pairwise interaction, indicating that dominant environmental controls were comparatively insensitive to DEM scale.

Temporal transferability and resolution selection

When historical models were transferred without refitting to the 2011–2025 holdout, AUROC remained 0.809–0.846, indicating that relative susceptibility ranking was preserved across time. Nevertheless, reliability analysis showed systematic under-confidence. Youden-optimal thresholds were 0.168 for RF, 0.131 for XGBoost and 0.149 for the probability-averaged ensemble, all substantially below the conventional 0.50 cutoff. At the optimized thresholds, RF sensitivity/specificity were 0.774/0.731 (Kappa = 0.502), XGBoost 0.812/0.718 (Kappa = 0.524), and the ensemble 0.793/0.725 (Kappa = 0.518). Thus, the models transferred better in relative ranking than in absolute probability calibration.

The balanced multi-criteria scores were 0.641 (5 m), 0.828 (12.5 m), 0.846 (30 m) and 0.200 (90 m). Alternative performance-focused and practicality-focused weighting schemes preserved the advantage of the 12.5–30 m range. The small balanced-score gap between 12.5 and 30 m (0.018) indicates that both are

defensible, while 30 m offers a practical compromise between predictive quality, computational burden and broad data availability. The 90 m option was disadvantaged by threshold-metric degradation, weaker explainability stability and greater sensitivity to coarse terrain representation. The balanced composite scores are summarized in Table 2.

Table 2, Multi-criteria composite scores under balanced weighting.

DEM	5 m	12.5 m	30 m	90 m
Balanced score	0.641	0.828	0.846	0.200

Conclusion

DEM resolution substantially changes selected geomorphometric derivatives and predictor redundancy, but it does not fundamentally change the discrimination capacity or the dominant environmental relationships learned by RF and XGBoost. TRI, SPI and slope were the most scale-sensitive variables, while plan and profile curvature were comparatively robust. The disagreement between stable AUROC and degraded RF F1 and Kappa at 90 m demonstrates that discrimination and probability calibration should not be treated as interchangeable properties of an LSM model. Stable SHAP rankings and the persistent Rainfall × Lithology interaction further indicate that the principal environmental controls are comparatively resolution-invariant.

For the investigated Himalayan corridor, 12.5–30 m is the defensible operational range under the evaluated criteria, with 30 m providing the strongest overall balance. Temporal validation further shows that a historical susceptibility model can retain relative ranking across time while its raw probabilities become under-confident. Accordingly, operational LSM should combine spatially grouped validation, threshold-dependent metrics and explainability analysis with an independent temporal holdout and probability recalibration using recent inventory data. This integrated approach provides a more reliable basis for DEM selection than AUROC alone.

Important limitations remain. The landslide inventory may contain differences in mapping completeness between historical and recent periods, and the temporal holdout represents one study region rather than a network of independent Himalayan basins. The multi-criteria resolution score is also decision-oriented: different users may assign different weights to predictive performance, interpretability and processing cost. These limitations motivate the next stage of work toward calibration transfer, external validation and testing whether the observed 12.5–30 m behaviour persists across additional physiographic and geological settings.

The temporal experiment also separates two commonly mixed properties of susceptibility models. Discrimination describes whether higher-risk locations

are ranked above lower-risk locations; calibration concerns whether a predicted probability corresponds to observed frequency. Here, the historical models retained useful ranking ability in 2011–2025, but the default 0.50 probability cutoff was poorly aligned with the later inventory. A susceptibility surface used for planning may therefore be interpreted primarily as a relative propensity surface, whereas binary operational decisions require a threshold derived from recent data and quantitative probability applications require recalibration. The use of an independent temporal holdout is important because it reveals distributional shifts that are not visible in spatial cross-validation alone.

References

- Chang, K.-T., Merghadi, A., Yunus, A. P., Pham, B. T., and Dou, J. (2019), “Evaluating scale effects of topographic variables in landslide susceptibility models using GIS-based machine learning techniques,” *Scientific Reports*, 9, 12296.
<https://doi.org/10.1038/s41598-019-48773-2>
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200), 675–701.
<https://doi.org/10.1080/01621459.1937.10503522>
- Gaidzik, K., and Ramírez-Herrera, M. T. (2021). The importance of input data on landslide susceptibility mapping. *Scientific Reports*, 11, 19334.
<https://doi.org/10.1038/s41598-021-98830-y>
- Lundberg, S. M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
<https://doi.org/10.48550/arXiv.1705.07874>
- Merghadi, A., Yunus, A. P., Dou, J., Whiteley, J., ThaiPham, B., Bui, D. T., Avtar, R., and Abderrahmane, B. (2020). Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. *Earth-Science Reviews*, 207, 103225.
10.1016/j.earscirev.2020.103225
- Nocentini, N., Rosi, A., Piciullo, L., Liu, Z., Segoni, S., and Fanti, R. (2024). Regional-scale spatiotemporal landslide probability assessment through machine learning and potential applications for operational warning systems. *Landslides*, 21, 2369–2387.
10.1007/s10346-024-02287-9
- Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillerá-Aroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schroeder, B., Thuiller, W., Warton, D. I., Wintle, B. A., and Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913–929. 10.1111/ecog.02881

- Usta, Z., Akıncı, H., and Akın, A. T. (2024). Comparison of tree-based ensemble learning algorithms for landslide susceptibility mapping in Murgul (Artvin), Turkey. *Earth Science Informatics*, 17, 1459–1481. 10.1007/s12145-024-01259-w
- Wubalem, A. (2022). The impact of DEM resolution on landslide susceptibility modeling. *Arabian Journal of Geosciences*, 15, 967. <https://doi.org/10.1007/s12517-022-10241-z>
- Yan, G., Tang, G., Li, S., Lu, D., Xiong, L., and Liang, S. (2023). Uncertainty in regional scale assessment of landslide susceptibility using various resolutions. *Natural Hazards*, 117, 399–423. <https://doi.org/10.1007/s11069-023-05865-7>
- Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35. [https://doi.org/10.1002/1097-0142\(1950\)3:1](https://doi.org/10.1002/1097-0142(1950)3:1)
- Zhao, X., Chen, W., Tsangaratos, P., and Ilia, I. (2024). Evaluating landslide susceptibility: The impact of resolution and hybrid integration approaches. *Geomatics, Natural Hazards and Risk*, 15(1), 2409198.