

COMPARATIVE ANALYSIS OF SPATIAL LANDSLIDE SUSCEPTIBILITY MODELS USING THE SZ PLUGIN IN QGIS, UNDER DATA-LIMITED CONDITIONS. CASE STUDY: SĂCELE RESERVOIR IN BRAȘOV, ROMANIA

Sorin-Gabriel VATAMANU¹ 

ABSTRACT. – **Comparative Analysis of Spatial Landslide Susceptibility Models Using the SPATIAL-ZONING Plugin in QGIS, Under Data-Limited Conditions. Case Study: Săcele Reservoir in Brașov, Romania.** Landslide susceptibility in data limited environments represents a methodological challenge. This study investigates spatial landslide susceptibility at Săcele reservoir in Brașov, Romania, by using an open-data reproducible approach implemented by the use of the Spatial-Zoning plugin in QGIS. Slope units were used as the primary mapping units for modeling, where independent parameters were derived freely from available geospatial datasets by using an extraction script in GEE (Google Earth Engine). The landslide inventory was created through in situ survey and satellite image analysis, identifying a total of 33 landslides within the study area. In order to evaluate robustness under small sample and imbalanced conditions, three approaches were tested, a) machine learning (ML) algorithms such as: Random Forest (RF), Support Vector Machine (SVM), and b) statistical tools as Generalized Additive Model (GAM). Firstly, model fittings were performed with k fold = 1, after which 5 fold cross-validation yielded comparable mean AUC values for RF and SV, (≈ 0.70), but with substantial variability (RF: 0.50-0.93, SVM: 0.52-0.86). GAM achieved a higher and more stable performance ($AUC \approx 0.75$, range 0.69-0.85), indicating superior generalization capacity. Considering predictive accuracy and stability, the GAM model was selected to produce the final susceptibility index map by using a confusion matrix. The findings of this study demonstrate that semi-parametric models may outperform more complex machine-learning approaches in small sample susceptibility studies.

Keywords: *GIS, Spatial-Zoning plugin, landslide inventory, machine learning, generalized additive model, AUC, slope units*

¹ Babeș-Bolyai University, Faculty of Environmental Sciences and Engineering, , 30 Fântânele Str., 400294 Cluj-Napoca, Romania, e-mail: sorin.vatamanu@stud.ubbcluj.ro.



1. INTRODUCTION

Landslide susceptibility represents a standard practice when designing hazard mitigation plans for vulnerable communities. In this context, Săcele reservoir in Braşov County has been identified as increasingly vulnerable to landslides. This growing vulnerability is largely driven by continuous deforestation carried out by the local community from within the valley where the Târlung River develops. This has led to the loss of vegetation fixating the soil, resulting in large hill areas that are uncovered and face continuous erosion processes.

Previous research such as the study “Landslide study at Săcele reservoir” conducted by EGU (European Geosciences Union), highlights the presence of landslide risks as being present in this area (J. Cannone, C. Scavia, M. Topa, 2010). The study starts with a qualitative and quantitative analysis of the terrain, to describe the features of the relief in the format of thematic maps. The terrain analysis reveals information such as the slope inclination and its exposure, the topographic limits and the type of relief from the study area.

The relief of the study area is karst and the geological substratum varies across the study area where neocamian siliceous sandstones, argillites and limestones are mostly predominant. However, around the eastern edges of the study area, lower Holocene gravel is mostly predominant (Geological Institute of Romania, 2022). These sedimentary rocks are well known for having easy erodibility properties when in contact with acidic or slightly acidic ph water levels, further confirming the karst relief that is present in the studied area (D. C. Ford, P. Williams, 2007).

For the landslide susceptibility models, a landslide inventory was created through in situ survey and direct observation from satellite images, identifying a total of 33 landslides. The landslide inventory represents the dependent parameter of the models, and is conditioned by several independent parameters. The study area is rather small, following the catchment of the Târlung River that supplies the reservoir, resulting in small training samples for different folds across the dataset. Because of this, several modeling methods were tested in order to determine the best performing method to generate the final susceptibility map considering both predictive accuracy and stability.

The independent parameters that were used for the models, were derived freely from available geospatial datasets. Independent parameters data was extracted using a code that was made available through a public repository by the creators of the Spatial-Zoning plugin (G. Titti et al., 2022). The code was introduced into GEE (Google Earth Engine) together with an asset representing the slope units. Small adjustments were made to the code in order to ensure that all functions deal with one task at a time and also extract data from the 2015-2025 timeframe, for the parameters that do not originate from the used DEM, such as the precipitation and NDVI parameters.

These independent parameters represent environmental factors mostly. Among them, one can find NDVI (Normalized Difference Vegetation Index), that represents vegetation density and health, serving as an indicator for slope stability. Areas with lower NDVI values represent sparse or degraded vegetation, which is associated with a higher susceptibility to landslides (I. Poddar, R. Roy, 2026). In addition, the standard deviation of NDVI was also used as a parameter in order to capture spatial variability in vegetation coverage, reflecting landscape heterogeneity and potential unstable areas that are present (G. Shang et al., 2024).

Topographic factors were also used as independent parameters for the models. Slope gradient directly influences gravitational geomorphological processes, with steeper slopes being more prone to failure (Z. Deng et al, 2022). The aspect, that represents slope exposure, was also used as a parameter but had to be transformed first in order to avoid circularity issues. The aspect was transformed into its sin and cos and it captures the influence of slope orientation on solar radiation absorption, moisture retention and vegetation distribution (A. W. Western et al., 1999). Relative relief was used as a parameter and it represents local elevation variability indicating terrain ruggedness.

As a climatic parameter, the average amount of precipitation between 2015 and 2025 was used for the models, at a spatial resolution of 5 Km with a cloud cover under 20%. Precipitation plays as a triggering factor for landslides, as it affects soil saturation and overall slope stability (F. Guzzetti, C. Vennari, 2007). By integrating a multi-annual average, the models capture the long-term hydrological conditions.

A time frame for the identified landslides was not available; therefore, the susceptibility models are only able to determine which areas are more prone to landslide occurrence based on spatial and environmental factors without estimating when the landslides could occur.

2. MATERIALS AND METHODS

2.1. Case study area

Brașov County is located in the central part of Romania, having the characteristics of a mountain basin, as it is surrounded by the central Southern Carpathian Mountains (A. Cioacă, M. Dinu, 2002), in the geographical and historical region of Transylvania and it has an area of approximatively 5351 km². It borders to the North with the counties of Mureș and Harghita, to the East with Covasna, to the South with Prahova, Dâmbovița, Argeș, Buzău, and in the West with Sibiu (fig. 1, a). The study area is located in the administrative unit of Săcele. The administrative unit of Săcele is located in the South-East of Brașov County, and covers the catchment of Tărlung River (fig. 1, b).

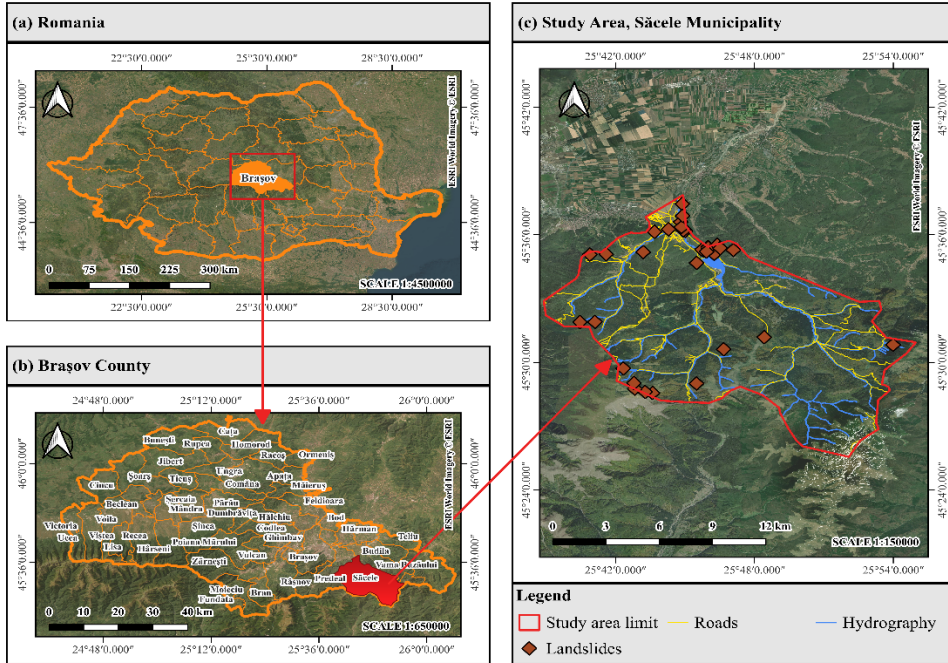


Fig. 1. **a.** Location of Brașov County in relation to neighboring counties, **b.** Location of the study area within Brașov County in relation to neighboring administrative units, **c.** Location of the study area within the administrative unit of Săcele in relation to Târlung River and surrounding infrastructure. *Source: the author*

The study area boundaries follow the catchment of Târlung River (fig. 1, c), within which a few infrastructure elements can be found. The study area was chosen based on the relevance of key infrastructure points that could be affected by landslides, for example the Săcele dam and reservoir, which serves as the main fresh water supply for the inhabitants of the city of Brașov.

2.2. Terrain analysis

The terrain analysis was carried out by using the morphometric and morphographic methods, which assess the qualitative and quantitative features of the terrain. They are presented in the format of thematic maps (fig. 2), and explain features such as the landforms and their evolution, altitudinal indices based on topographic steps, slope aspect and inclination level, as well as the terrain's hillshade (T. Morariu, V. Velcea, 1971).

COMPARATIVE ANALYSIS OF SPATIAL LANDSLIDE SUSCEPTIBILITY MODELS USING THE SZ PLUGIN IN QGIS, UNDER DATA-LIMITED CONDITIONS. CASE STUDY: SĂCELE RESERVOIR IN BRAȘOV, ROMANIA

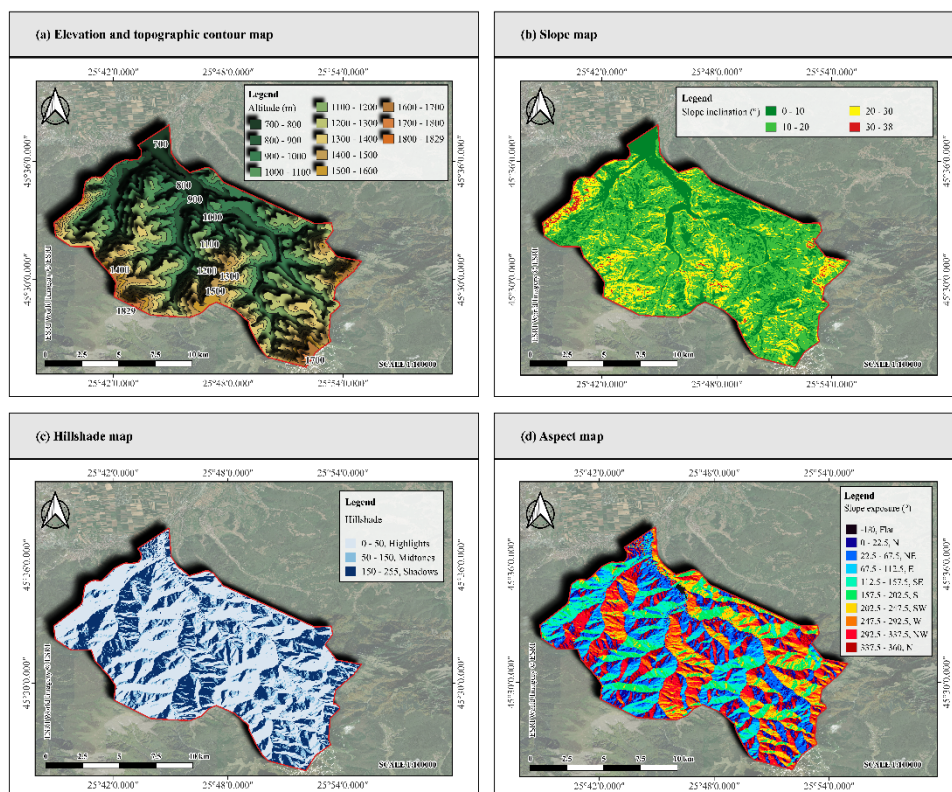


Fig. 2. Terrain analysis thematic maps representing **a)** Elevation and topographic contour, **b)** Slope in degrees, **c)** Hillshade, **d)** Aspect. *Source: the author*

a) The first map (fig. 2. a) presents the topographic boundaries and the elevation of the study area, providing perspective on the relief and terrain ruggedness. The elevation contour has an index interval of 100 m. The terrain is mixed and rugged with a clear definition of a steep valley in the proximity of Tărlung River and Săcele reservoir. The highest elevation point is situated in the South-West of the study area with a maximum elevation of 1829 m, while the lowest elevation point is situated in the North, downstream of the river at around 682 m.

b) The second map (fig. 2. b) presents the slope values and refers to the level of inclination at a given point. It can be measured in degrees or percentages, also providing insight about the relief and it plays an important

role in landslide susceptibility (G. Grelle, A. Donnarumma, P. Revellino, 2013). The slope values were derived from the DEM with a standard resolution of 30 m where generally low to moderate slope values are present, with exceptions around peaks and along the valley walls, formed by the drainage network. In the proximity of the reservoir, one can notice that high slope values are present, with over 30° inclination, in the East.

c) The third map (fig. 2. c) visually represents the relief that simulates the way in which light and solar radiation would fall on the surface of the terrain at a certain azimuth angle (ArcGis User manual). The terrain mostly favors the absorption of light, having a preferential direction facing the West with a Northern tendency, therefore favorizing the absorption of light especially during the warm months, May-August.

d) The fourth map (fig. 2. d) shows the cardinal direction in which a slope is inclined, also called slope exposure or aspect. The legend and the map were created according to the values and steps provided by the ARCGIS user manual, a similar software to QGIS. The aspect can influence microclimates and the distribution of vegetation through the variation of received solar radiation (G. G. Sokolova, 2016).

2.3. Datasets

2.3.1. Landslides inventory

The landslides inventory was created by combining two approaches, direct observation from satellite images and in-situ survey. For the in-situ survey, several visits were conducted in order to identify reachable landslides. The landslides are visible and accessible by car, where several roads have a direct viewpoint on them, mostly in the proximity of Săcele dam and the walls of the valley where the river flows, summing up a total of 12 identified landslides. By direct observation, with the help of satellite images, a number of 21 landslides were identified, but those are not accessible by roads and they mostly develop next to high slopes which are positioned around higher elevation levels.

The landslides vary in size, and the most significant sizes are mainly present within the river's valley, where steep slopes are also present. The landslides inventory has a point resolution and every identified landslide corresponds with a marker on the map (fig. 3. g). The vectorial layer representing the inventory was then merged in QGIS with the slope units, where they have the following property: 0 for absence and 1 for presence, representing a binomial variable.

2.3.2. Geomorphometry

Geomorphometric parameters were derived from the used DEM, with a standard resolution of 30 m. Shuttle Radar Topography Mission or in short SRTM is a NASA mission that mapped the Earth's land surface elevation in order to create a digital elevation model (ESA 2025). The mission took place in the early 2000's and still provides valuable information for studies today. Based on the DEM, a series of parameters can be derived such as aspect, slope, elevation and relative relief. In Google Earth Engine (GEE), the extraction code that was used in order to extract the data for the parameters, used the USGS/SRTMGL1_003 product, which is a 1 arc-second (30 m) DEM. The independent parameters that were derived are the terrain slope, relative relief and aspect, represented in Table 1. The geomorphometric parameters thematic maps using the slope units can be observed in Figure 3, a, b, and c.

2.3.3. Normalized difference vegetation index (NDVI)

The normalized difference vegetation index comes from the Sentinel-2 mission, that consists of two polar-orbiting satellites Sentinel2A and Sentinel2B. Both of the satellites are equipped with a Multispectral Instrument (MSI), recording imagery in 13 spectral bands, where the MSI is an optical imaging high-resolution sensor that supports multiple applications that include vegetation coverage, vegetation monitoring and water quality monitoring. The satellites range from 10 m to 60 m and the radiometric resolution is 12-bit, with an aggregate revisit time (ESA 2025). The extraction code that was used instructs Google Earth Engine (GEE) to extract only Sentinel imagery that has a cloud coverage under the 20% threshold, that is a standard value (L. Illarionova et al., 2025).

NDVI is computed from Sentinel-2 imagery as the normalized difference between near-infrared/NIR (Band 8) and red (Band 4) reflectance (L. Illarionova et al., 2025). Both the mean and standard deviation of NDVI were used as input parameters in order to represent vegetation density in spatial variability, providing additional information on landscape heterogeneity relevant to landslide susceptibility (J. Riera et al., 1998). The thematic maps representing the NDVI mean and standard deviation using the slope units can be viewed in Figure 3. d and e.

2.3.4. Daily precipitation – Annual average amount

The precipitation data was obtained from a CHIRPS dataset, which is a quasi-global and blended rainfall product. CHIRPS is made up of several components, where precipitation estimates are derived from geostationary satellites that calculate rainfall intensity based on cloud-top temperatures and cold-cloud duration (C. Funk et al., 2015). The satellite derived estimates are then combined with a high-resolution precipitation climatology (CHPclim) in order to compute rainfall anomalies. The last component of CHIRPS is made up of ground-based gauge rain observations in order to bias-correct the computed

rainfall anomalies (C. Funk et al., 2015). The final product represents a spatially and temporally consistent precipitation dataset at approximately 5 km resolution. This blended approach improves accuracy compared to satellite only estimates. A summary of all datasets used in the study, including satellite-based and in-situ based observations is provided in the Table 1. The thematic map using the slope units representing the precipitation data can be viewed in Figure 3. f.

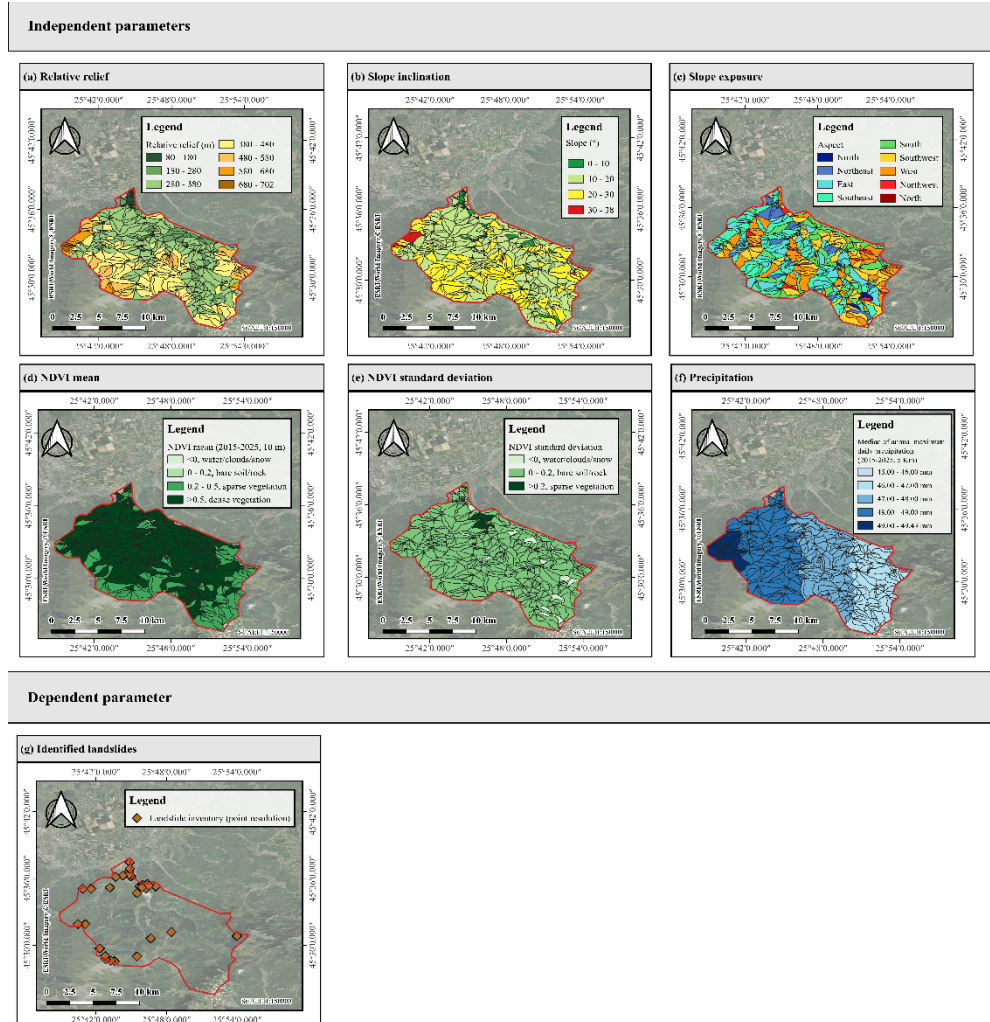


Fig. 3. Thematic maps of the used parameters per slope units, **a)** relative relief, **b)** Slope inclination, **c)** Slope exposure, **d)** NDVI mean, **e)** NDVI standard deviation, **f)** Precipitation, **g)** landslide inventory. *Source: the author*

Table 1. Summary of datasets used in this study.

Dataset/Product	Parameter	Spatial resolution	Temporal coverage	Time resolution	Cloud cover	Data source
SRTMGL1_003	Slope, Aspect, Relief	30 m	~ 2000	Static	N/A	USGS Earth Explorer
Sentinel-2 SR	NDVI (mean, standard deviation)	10 m	2015-2025	5 days	< 20%	Copernicus
CHIRPS	Precipitation	5 Km	2015-2025	Daily	N/A	UCSB Climate Hazard Center
Landslide inventory	Landslides	Point	N/A	N/A	N/A	Satellite/ in-situ observations

Source: the author

As one can remark in table 1, there is a resolution difference between the datasets, as they represent different physical processes. Because of this, slope units were selected as the primary mapping unit where the data was aggregated to polygons in GEE by using the function “reduceRegions”. This means that each slope unit/ polygon gets the average/median values of the underlying raster layer regardless of the original resolution. Aggregation effectively harmonizes the datasets so that the models use one feature table with consistent values per slope units (T. R. Bhattarai, N. P. Bhandary, 2026).

2.4. Data preprocessing

2.4.1. Data integration

After running the extraction code in GEE together with the asset that represents the vector layer of the slope units, the independent parameters data was exported into QGIS under three layers representing the NDVI, precipitation and geomorphometry. Data integration deals with using the parameters under the same layer in order to be able to use the Spatial-Zoning plugin. For that, the joins function from under the slope units layer properties was used where the target field selected was DN.

2.4.2. Data cleaning

After the data integration, the attribute table of the final layer was examined for missing or inconsistent values. In the precipitation parameter derived from CHIRPS, several cells corresponding to some slope units contained the value of 0. Given the coarse resolution of 5 km relative to the slope unit scale, these values were interpreted as artefacts resulting from spatial mismatch during the extraction process, representing a partial pixel coverage or no data assignment, rather than true 0 precipitation.

Although slope units were used as the primary mapping units specifically in order to avoid resolution mismatch errors and to harmonize spatial representation, the integration of datasets with differing spatial resolution may still introduce artefacts during the aggregation process (E. B. Jo, J. H. Lee, H. J. Park, 2025), and these values can also be a result of the raster to polygon conversion. To address this issue, a mean imputation approach was applied, where the value of 0 was replaced with the average of valid observations within the dataset.

In addition to the data cleaning, a variable transformation was applied for the raw aspect parameter to account for its circular nature. Since aspect values range from 0° to 360°, direct use in statistical models may introduce artificial discontinuities. To overcome this limitation, aspect was decomposed into its sin and cos components (H. Liu et al., 2026). This transformation preserves directional information while ensuring continuity in the predictor space. The transformation was made in QGIS by using the raster calculator with the following formulas “sin(radians(“Asp_mean”)), cos(radians(“Asp_mean”))”.

2.4.3. Multicollinearity analysis

Before the models development by using both ML algorithms and statistical tools, after obtaining the final layer and cleaning the missing data, a correlation analysis was employed to assess potential multicollinearity among the independent variables. For this the correlation plot from the Spatial-Zoning plugin was used from under the data preparation section. The correlation plot uses the Pearson correlation coefficient that quantifies the linear relationship between pairs of variables (H. Akoglu, 2018). The coefficient is defined as:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (\text{K. Pearson, 1895})$$

Where r represents the strength and direction of the linear relationship between two variables. This analysis was used in order to identify highly correlated variables and reduce redundancy among the predictors prior to model training. For this a threshold of $|r| > 0.7$ was selected to represent parameters that describe too similar information, everything under this threshold being considered acceptable for the models (K. P. Vatcheva et al., 2016). After running the correlation plot the following relationships between the parameters were established and can be observed in Table 2.

Table 2. Correlation plot between the independent variables

Relief	1.00	0.60	-0.08	0.18	-0.03	0.12	-0.11
Slope	0.60	1.00	0.10	-0.01	0.11	0.07	-0.26
NDVI	-0.08	0.10	1.00	0.11	-0.26	0.02	-0.26
Precipitation	0.18	-0.01	0.11	1.00	0.15	0.27	0.10
NDVI sd	-0.03	0.11	-0.26	0.15	1.00	0.18	-0.08
Aspect sin	0.12	0.07	0.02	0.27	0.18	1.00	-0.03
Aspect cos	-0.11	-0.26	-0.24	0.10	-0.08	-0.03	1.00
	Relief	Slope	NDVI	Precipitation	NDVI sd	Aspect sin	Aspect cos

Source: the author

The correlation plot indicates low to moderate correlations, with most coefficients ranging from between -0.26 to 0.24. The highest correlation ($r=0.60$) was remarked for the slope and relief parameters, and this is mainly because of their common derivation, as they both originate from the same digital elevation model that was used, SRTMGL1_003. However, this value remains below the commonly accepted threshold of $|r| > 0.70$ for multicollinearity (E. Putriani et al. 2026).

2.5. Machine learning (ML) and statistical models

2.5.1. Support Vector Machine (SVM)

SVM represents a supervised learning algorithm that finds the optimal decision boundary, also called a hyperplane, maximizing the margin between classes, such as landslide presence and absence, representing a binomial format (B. T. Pham et al., 2017). It identifies support vectors, the training points closest to the decision boundary, which are essential for defining its position. SVM uses the kernel trick, which maps input data into a higher-dimensional feature space where linear separation becomes possible (B. E. Boser et al., 1992). The decision function can be expressed as:

$$f(x) = \sum_{i=1}^n \alpha_i y_i K(x_i, x) + b \quad (\text{B. E. Boser et al., 1992})$$

Where $K(x_i, x)$ represents the kernel function, α_i represents Lagrange multipliers and y_i the class labels. In landslides susceptibility studies, SVM can be an effective machine learning algorithm due to its ability to model complex boundaries between stable and unstable areas. It performs best in high-dimensional spaces and it can capture subtle patterns in the data (M. H. Huh, 2015). SVM performance can also be affected by the size of training samples, where few landslide samples can overfit the model or even show up with mediocre metric values (S. Salcedo-Sanz et al., 2014).

2.5.2. Random Forest (RF)

RF represents a ML algorithm that constructs a large number of individual decision trees and then combines their predictions in order to improve the overall model performance and robustness (H. A. Salman, A. Kalakech, A. Steiti, 2024). Each tree is trained by using a bootstrap sample of the original dataset, meaning that observations are randomly sampled with replacement. At each node split, only a subset of predictor parameters is considered, which reduces correlation among trees and enhances model generalization (L. Breiman, 2001). The final prediction is obtained by aggregating the outputs of all individual trees, typically through majority voting for classification problems and averaging for regression and is formulated as:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T f_t(x) \quad (\text{L. Breiman, 2001})$$

Where T is the number of trees and $f_t(x)$ is the prediction of the t -th tree. The RF algorithm can handle complex nonlinear relationships between environmental variables and landslide occurrences. It is also robust to noise, outliers and multicollinearity among predictors, which are common issues in geospatial datasets (G. Biau, E. Scornet, 2016). However, because RF is a ML algorithm, similar to SVM, it also requires larger datasets with more samples. The use of the RF is strongly dependent on the samples size, where few samples can overfit the model.

2.5.3. Generalized Additive Model (GAM)

GAM represents a semi-parametric extension of the generalized linear models that allows more flexible, nonlinear relationships between predictor variables and the response variable (T. Hastie, R. Tibshirani, 1986). Instead of assuming a strictly linear relationship, GAM models the response as the sum of smooth functions of the predictors as:

$$g[E(Y)] = \beta_0 + \sum_{i=1}^p f_i(X_i) \quad (\text{T. Hastie, R. Tibshirani, 1986})$$

Where g is the link function, β_0 is the intercept, f_i are the smoothing functions typically represented by splines and X_i are the predictor variables. For binary outcomes such as landslide occurrence/absence, a logistic link function is used and is represented as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \sum f_i(X_i) \quad (\text{T. Hastie, D. Pregibon, 1992})$$

GAM represents a statistical tool that can capture nonlinear relationships between environmental factors such as the independent parameters that were used for the models and landslide probability while maintaining interpretability (S. N. Wood, 2006). Compared to the ML algorithms, GAM usually provides a balance between predictive performance and interpretability, which explains its superior performance and stability in this study.

2.6. Evaluation metrics

The Spatial-Zoning plugin from QGIS offers a couple of evaluation metrics, such as AUC (The Area Under the Receiver Operating Characteristic Curve), F1 and K (Cohen's Kappa). Models performance was primarily evaluated using the AUC values, which provides a threshold-independent measure of model discrimination. Additionally, F1 and K were used as supplementary metrics only, providing further insight into classification performance.

AUC is used to evaluate the ability of a model to distinguish between landslide presence and absence and it represents the probability that a randomly selected landslide pixel/ presence is assigned a higher susceptibility value than a randomly selected non-landslide pixel/ absence (D. W. Hosmer, S. Lemeshow, 2000). The ROC curve is generated by plotting the true positive rate (sensitivity) against the false positive rate across all possible classification thresholds (T. Fawcett, 2006). In this study, AUC was used during both model fitting ($k=1$) and cross-validation ($k=5$) to compare the predictive performance of the models.

The F1 score represents the precision and recall of the models, providing a balanced measure of classification performance, particularly in the presence of imbalanced datasets (M. Sokolova, N. Japkowicz, S. Szpakowicz, 2006), which is the case of this study taking into account the open-data sources of the datasets that were used. It reflects the proportion of correctly predicted landslides among all predicted landslides, while recall measures the proportion of actual landslides correctly identified by the model (M. Sokolova, G. Lapalme, 2009).

Cohen's Kappa coefficient measures the agreement between observed and predicted classifications while accounting for agreement beyond chance (A. Ben-David, 2008). Unlike overall accuracy, K provides a more robust assessment of model performance. It is mostly used in conjunction with the F1 score to support the selection of the optimal classification threshold (J. Cohen, 1960). The evaluation metrics are mathematically represented in table 3.

Table 3. The list of statistical evaluation metrics

Statistical metrics	Formulation
AUC (Area Under the ROC Curve)	$AUC = \int_0^1 TPR(FPR) d(FPR)$
F1-score	$2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$
K (Cohen's Kappa)	$\frac{P_o - P_e}{1 - P_e}$

Source: the author

2.7. Susceptibility classification and statistical validation

2.7.1. Confusion matrix and classification threshold

Following the model selection based on AUC performance, a confusion matrix was employed to generate the final susceptibility index map. The confusion matrix compares predicted and observed values and is defined by four components: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) (T. Fawcett, 2006). Based on these components, classification performance is expressed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall (sensitivity) = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

2.7.2. Statistical validation

To further evaluate the model's ability to discriminate between landslide presence and absence, a statistical validation was conducted on the susceptibility index values after running the model through the confusion matrix. First, a boxplot was used to visualize the distribution of SI values for the two classes. The boxplot summarizes the data using the following statistical components: first quartile (Q1), median (Q2), third quartile (Q3), following the interquartile range:

$$IQR = Q_3 - Q_1 \quad (\text{J. W. Tukey, 2008})$$

These elements allow for a visual comparison of the central tendency and dispersion between presence and absence groups (J. W. Tukey, 2008). To statistically test whether the two distributions differ, a Mann-Whitney U test was employed. The U statistic is expressed as:

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

Where n_1 and n_2 are the sample size of the two groups, and R_1 is the sum of ranks for the first group (H. B. Mann, D. R. Whitney, 1947). The test evaluates whether one group tends to have higher values than the other, without assuming normality. The statistical significance of the result is determined using the p-value, which represents the probability of observing the obtained difference under the null hypothesis, H_0 : the two groups have identical distributions (W. J. Conover, 1999). A significance threshold of $p < 0.05$ was used. The obtained p-value ($p = 9.87 \times 10^{-5}$) indicates a statistically significant difference between presence and absence of landslides (S. C. Pearce, 1992).

3. RESULTS

3.1. Models validation

The performance of the models was evaluated using 5-fold cross-validation. The value of 5 was determined by considering the number of landslides each training fold gets, as the study has its limitations, meaning that at 5 training folds, each model gets at least 3 to 4 landslide samples per training fold. The results are summarized as mean values and standard deviation across the folds in order to represent the stability of the model.

The GAM model achieved the highest predictive performance based on the models graphs presented in fig. 4 and table 4, with an AUC of 0.75 ± 0.06 , indicating both good discrimination ability and relatively low variability. In contrast, the ML models, RF and SVM respectively, yielded lower mean AUC values (0.70 ± 0.17 and 0.70 ± 0.14), accompanied by higher standard deviations, suggesting less stable performance across folds.

The F1 score and Cohen's Kappa values were low across all models, reflecting on the nature of the imbalanced dataset, characterized by a limited number of landslide occurrences. Given that the AUC is threshold-independent and less sensitive to class imbalance, it was considered the primary metric for model comparison in this study. Based on these results, the GAM model was selected for the generation of the final landslide susceptibility map.

Table 4. Computed model results (mean + standard deviation)

Model	AUC	F1	K
SVM	0.70 ± 0.14	0.13 ± 0.03	0.06 ± 0.04
RF	0.70 ± 0.17	0.12 ± 0.01	0.05 ± 0.01
GAM	0.75 ± 0.06	0.12 ± 0.02	0.06 ± 0.02

Source: the author

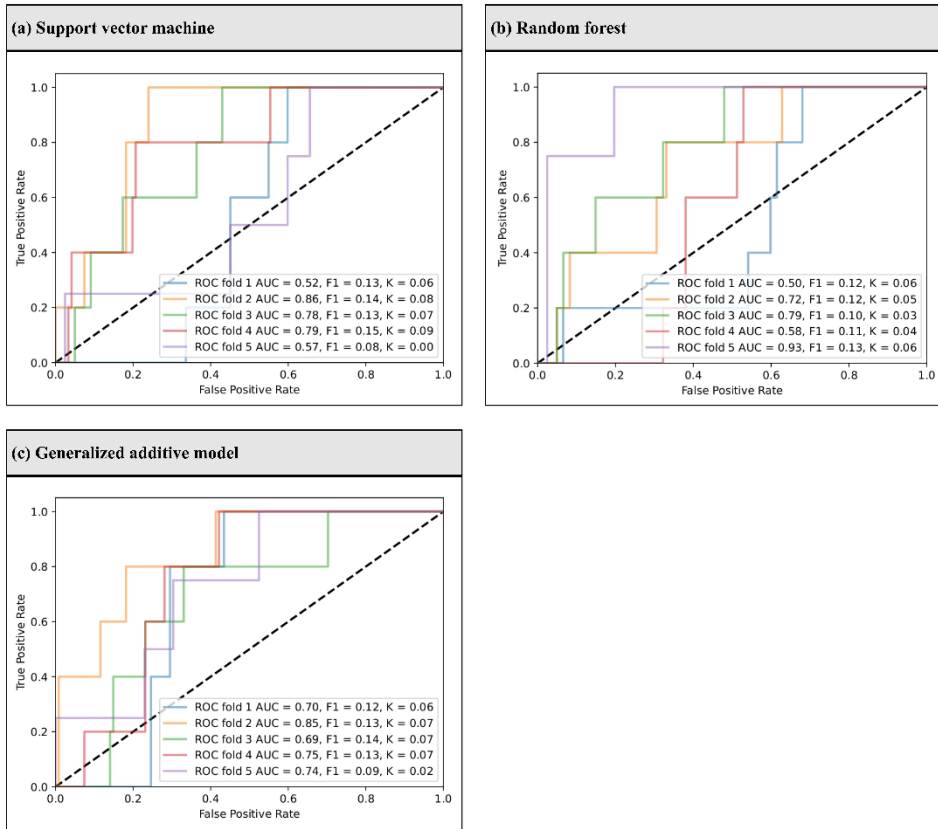


Fig. 4. Comparison of model performance **a)** SVM/**b)** RF/**c)** GAM, using 5-fold cross-validation based on metrics performance graphs: AUC, F1 score and Cohen's Kappa.

Source: the author

3.2. Influence of conditioning factors for the GAM model

Unlike the ML algorithms, under statistical tools in the Spatial-Zoning plugin from QGIS, when computing the GAM model, NDVI mean and Slope mean were selected as interacting variables. This option is meant to show how two variables combined effect can influence landslide susceptibility (G. Titti et al., 2025). The partial effect is represented by $f(NDVI, Slope)$, and it measures the interaction surface between the variables. The interaction between the parameters is scaled, resulting in a pattern (fig. 5). Sparse vegetation together with higher slope values can result in a higher susceptibility while dense vegetation with gentle slopes results in more stable areas indicating a lower susceptibility (R. C. Sidle, T. A. Bogaard, 2016). The figure also shows how the vegetation effect is highly dependent on the slope values. This also represents an advantage of using the GAM model approach as the ML algorithms cannot produce a scaled interaction between two sets of variables.

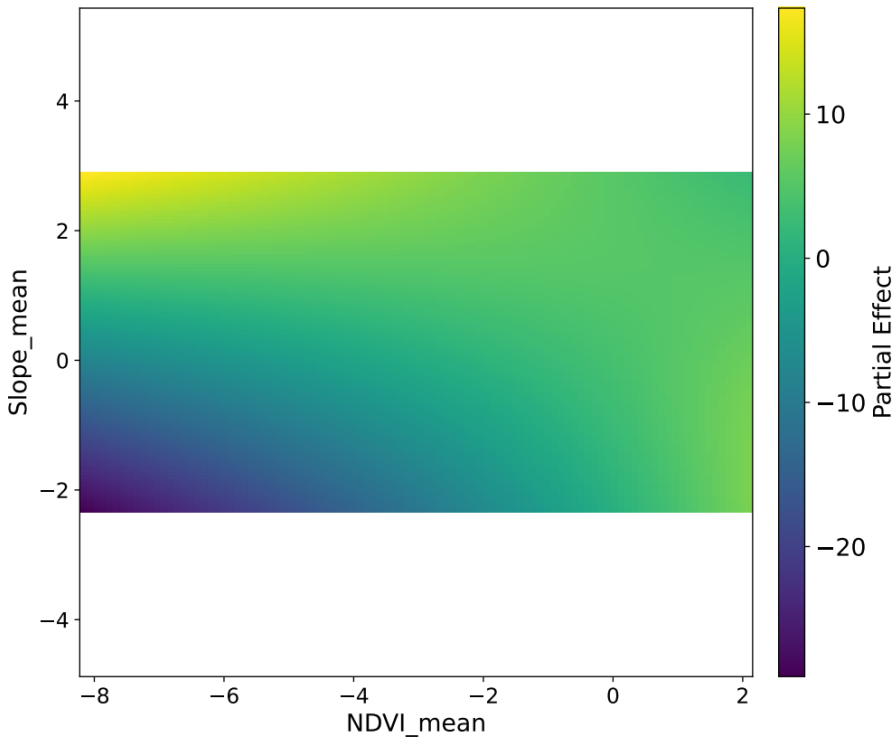


Fig. 5. Scaled partial effect between the NDVI_mean and Slope_mean parameters.
Source: the author

Partial effects plots were generated for the other used variables in order to assess their contribution to susceptibility for the GAM model (fig. 6). They illustrate the individual influence of conditioning factors on landslide susceptibility. One may notice that Rainfall (RnMax_mean) exhibits a clear positive relationship, indicating that higher precipitation levels significantly increase susceptibility (R. C. Sidle, T. A. Bogaard, 2016). Similarly, the standard deviation of NDVI (NDVI_sd) shows a strong positive effect, suggesting that areas characterized by heterogeneous or disturbed vegetation are more prone to slope instability. In contrast, relief (Rlf_mean) demonstrates a weak negative trend, implying a limited influence on landslide occurrence. The aspect related variables (asp_sin, asp_cos) display minimal and inconsistent effects, indicating that slope orientation does not play a significant role in controlling landslide susceptibility. Overall, the results highlight rainfall and vegetation variability as the most influential factors, while topographic orientation and relief exert comparatively minor effects in this specific study.

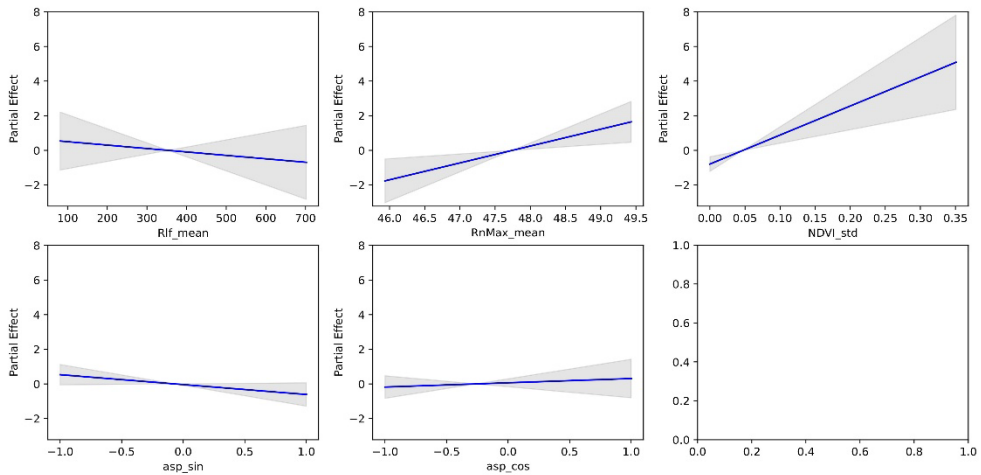


Fig. 6. Partial effect plots derived from the GAM model, showing conditioning factor influences on landslide susceptibility. *Source: the author*

3.3. Confusion matrix and susceptibility index map

Following the model evaluation, the Generalized Additive Model (GAM) was selected as the best performing model based on its superior and more consistent AUC values compared to the ML models. The selected model was subsequently applied within a confusion matrix framework to generate the final landslide susceptibility map.

The confusion matrix enables the assessment of classification performance by comparing observed landslide occurrences with predicted susceptibility outcomes, quantifying true positives, true negatives, false positives and false negatives (T. Fawcett, 2006). For the confusion matrix a conservative threshold representing a cutoff percentile was selected to the 90th percentile, ensuring that only the highest susceptibility values were classified as landslide prone areas (F. Guzzetti et al., 2006).

This approach prioritizes reliability by reducing false positives and focusing on the most critical zones of potential instability, ultimately supporting a robust and interpretable susceptibility assessment.

After applying the confusion matrix with the GAM model, the final slope-unit based landslide susceptibility map was generated. For this map the continuous susceptibility index (SI) was classified into five categories from very low to very high by using both a quantile-based approach and a natural breaks approach. The quantile-based approach ensures an equal distribution of slope units across classes and facilitates comparison between susceptibility levels (T. A. Slocum et al., 2009), while the natural breaks approach displays only the most critical slope units that are prone to landslide occurrence (G. F. Jenks, 1967). The final landslide susceptibility map is presented in fig. 7, and explained below the figure.

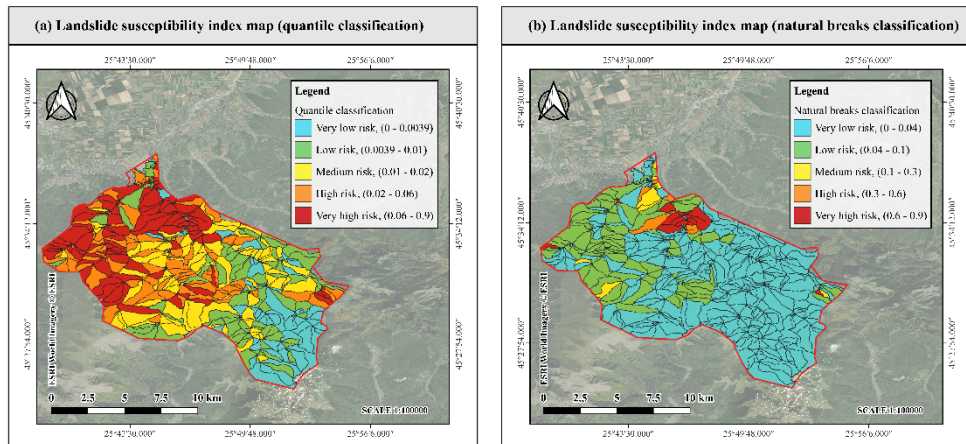


Fig. 7. Landslide susceptibility map based on the susceptibility index (SI), classified into five categories across slope units. *Source: the author*

The final slope-unit-based landslide susceptibility map indicates that slope units under the non-landslide label show up as highly susceptible. This means that those areas represent unidentified landslides or that those areas, based on the selected parameters, are areas that can represent future landslides. The

model managed to correctly identify and classify slope units with landslide presence as highly susceptible areas. The distribution of the high-risk slope units varies. They are mostly present around the accumulation basin, where previous studies identified slopes around the basin as being unstable. There are also several high-risk slope units around the escarpments of Tărlung River valley.

3.4. Statistical validation

To further validate the predictive performance of the selected model, a non-parametric statistical analysis was employed using the Mann-Whitney U test accompanied by a boxplot visualization. The boxplot from Figure 8., illustrates the distribution of the susceptibility index (SI) values for landslide presence and absence, highlighting differences in median, interquartile range, and the presence of outliers between the two groups.

This visual comparison provides an indication of the model's ability to discriminate between stable and unstable slope units. The Mann-Whitney U test was applied to quantitatively assess whether the SI values associated with landslide occurrences are statistically significantly higher than those of non-landslide areas, without assuming a normal distribution of the data.

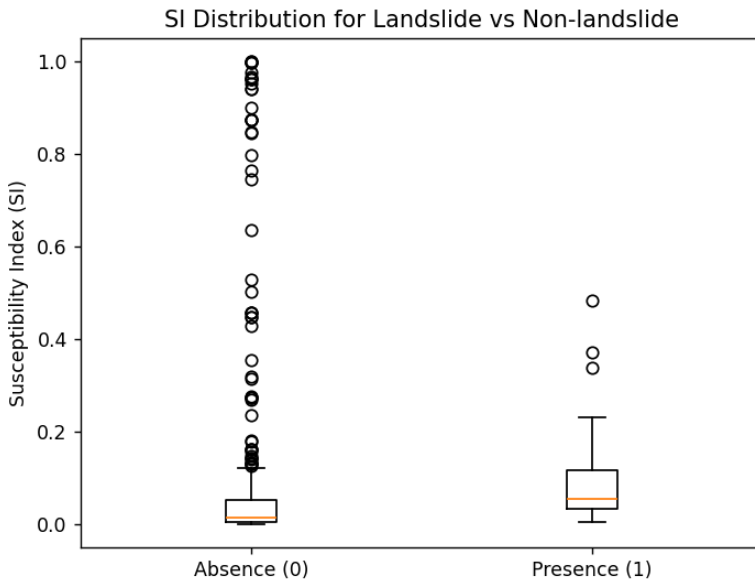


Fig. 8. Boxplot of susceptibility index (SI) for landslide and non-landslide areas.

Source: the author

The Mann–Whitney U test yielded a U value of 10,695.0 and a p-value of 9.88×10^{-5} ($p < 0.05$), indicating a statistically significant difference in the rank distributions of susceptibility values between landslide and non-landslide locations. This finding provides evidence that the two groups are characterized by different susceptibility-value distributions. Further assessment using appropriate validation and accuracy measures is required to determine the overall performance of the susceptibility model.

4. DISCUSSIONS

This study compared statistical and ML methods for landslide susceptibility mapping in a data-limited environment using the Spatial-Zoning plugin from QGIS. Support Vector Machine (SVM), Random Forest (RF), and Generalized Additive Model (GAM) were evaluated with slope units as the mapping unit and an inventory of 33 landslides.

The results show that the GAM model outperformed the ML models in terms of stability and predictive performance based on the AUC values. Under limited sample sizes, flexible and statistically regularized models like GAM may generalize more reliably than complex ML algorithms. The poor and unstable performance of SVM and RF likely results from overfitting on small, imbalanced datasets, where too few landslide occurrences prevent accurate class boundary definition.

Complementary metrics such as F1 score and Cohen's Kappa were also applied but remained low for all models. This strongly reflects on the class imbalance in the dataset, a common issue in landslide susceptibility when using open-access inventories. The imbalance limits the usefulness of classification metrics beyond the non-landslide class, highlighting the need to interpret AUC in a relative rather than absolute sense.

The set of parameters, including the mean and standard deviation of NDVI, slope angle, relative relief and transformed aspect, provides a meaningful representation of the terrain and vegetation. Using slope units as spatial entities improved geomorphological consistency over pixel-based approaches and aligns with established practices in terrain susceptibility analysis. There were no land-use or lithologic parameters that were used and this is mainly because of the data scarcity conditions, at least for lithology. The land use was not considered to be a relevant parameter that adds information to the models mostly because the study area is covered in vegetation, where few structures are present, not covering a consistent part of the study area in order to be taken into consideration.

By integrating open geospatial datasets through the help of GEE, it is demonstrated that this approach has the potential for reproducible and scalable landslide susceptibility mapping workflows. However, the small landslide inventory remains a significant constraint on model robustness.

These results support and extend previous findings, including the 2010 European Geosciences Union study, which identified landslide risks in the study area. This analysis confirms that these patterns remain valid in 2026, suggesting temporal stability of the controlling geomorphological factors under current conditions.

5. CONCLUSIONS

This study evaluated the performance of ML algorithms, specifically SVM and RF in comparison with the statistical GAM, for landslide susceptibility mapping under limited data conditions using slope units and open-access geospatial inputs. The primary finding is that the GAM model exhibited the most stable and reliable performance relative to the evaluated ML methods, particularly when applied to a highly constrained dataset containing only a few landslide training samples. These results suggest that, in data-limited environments, statistically driven models may outperform more complex ML approaches due to reduced sensitivity to overfitting and class imbalance.

The study further demonstrates that data imbalance significantly affects classification based performance metrics, resulting in low F1 and K values across all models. However, the AUC results indicate that meaningful predictive patterns can still be identified from limited data when appropriate modeling strategies are employed.

These findings also corroborate previous research describing the study area as being increasingly vulnerable to landslide risks, especially in proximity of Săcele dam, which represents a key infrastructure point for the inhabitants of the city of Braşov. The study emphasizes that model selection for landslide susceptibility analysis should be primarily guided by data availability, with simpler statistical approaches offering a robust alternative in data limited contexts.

REFERENCES

1. Akoglu, H. (2018), *User's guide to correlation coefficients*, Turkish Journal of Emergency Medicine, Volume 18, Issue 3, Pages 91-93, ISSN: 2452-2473, <https://doi.org/10.1016/j.tjem.2018.08.001>.

2. Ben-David, A. (2008), *Comparison of classification accuracy using Cohen's Weighted Kappa*, Expert Systems with Applications, 34, 2, p. 825-832, ISSN 0957-4174, <https://doi.org/10.1016/j.eswa.2006.10.022>.
3. Bhattarai T.R., Bhandary N.P. (2026), *Comparative analysis of slope and grid units for co-seismic landslide susceptibility mapping using machine learning methods*, Discover Geoscience; 4:121. Doi:10.1007/s44288-026-00489-3.
4. Biau, G., Scornet, E. (2016), *A random forest guided tour*, TEST, 25, p. 197-227. Doi:10.1007/s11749-016-0481-7.
5. Boser, B.E., Guyon, I.M., and Vapnik, V.N. (1992), *A training algorithm for optimal margin classifiers*, Proceedings of the 5th Annual Workshop Computer Learning Theory, Association for Computing Machinery, Doi:10.1145/130385.130401.
6. Breiman. L. (2001), *Random forests*, Machine Learning, 45, p. 5-32. Doi:10.1023/A:1010933404324,
7. Cannone, J., Scavia, C., Topa, M., and Cotesta, L. (2010), *Landslide hazard assessment at the Săcele reservoir (Romania)*, in: EGU General Assembly 2010; May 2-7; Vienna, Austria. Copernicus Pub; 2010. p. 14655. (EGU Gen Assem. Conference Abstracts; vol. 12). Bibcode: 2010EGUGA.1214655C.
8. Cioacă, A. and Dinu, M. (2002), *Pliocene-Quaternary evolution in the contact area between Brașov Depression and the surrounding mountains (Romania)*, Geografia Fisica e Dinamica Quaternaria. Comitato Glaciologico Italiano, Torino, 24, pp. 3-14..
9. Cohen, J., (1960), *A coefficient of agreement for nominal scales*, Educational and Psychological Measurement, 20, 37-46. <https://doi.org/10.1177/001316446002000104>.
10. Conover, W.J. (1999), *Practical nonparametric statistics*. 3rd ed., Wiley, New York.
11. Deng, Z., Pan, M., Niu, J., and Jiang, S.H. (2022), *Full probability design of soil slopes considering both stratigraphic uncertainty and spatial variability of soil properties*, Bulletin of Engineering Geology and the Environment, 81 (5), 195, Doi:10.1007/s10064-022-02702-2.
12. Esri (2025), *ArcGIS Pro documentation*, available from: <https://pro.arcgis.com/> (last accessed 28th April 2026).
13. Esri (2026), *World Imagery basemap*, Available from: https://services.arcgisonline.com/ArcGIS/rest/services/World_Imagery/MapServer/tile/{z}/{y}/{x} (last accessed 18th May 2026)
14. European Space Agency (2025), *Sentinel-2 mission overview*, Available from: https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-2 (last accessed 26th April 2026).
15. Fawcett, T. (2006), *An introduction to ROC analysis*, Pattern Recognition Letters, Vol. 27, Issue 8, p. 861-874, ISSN: 0167-8655, <https://doi.org/10.1016/j.patrec.2005.10.010>.
16. Ford, D.C. and Williams, P. (2007), *Karst hydrogeology and geomorphology*, Wiley, Chichester, Doi:10.1002/9781118684986.
17. Funk, C., Peterson, P., Landsfeld, M., Pedreros, H.D., Verdin, J., Shukla, S., Husak, J.G., Rowland, J., Harrison, L., Hoell, A., and Michaelsen J. (2015), *The climate hazards infrared precipitation with stations (CHIRPS): A new environmental record for monitoring extremes*, Scientific Data, 2, 150066, Doi:10.1038/sdata.2015.66.

18. Geological Institute of Romania (1968), *Spatial geological map data*. Available from: <https://geoportal.igr.ro/viewgeol1M> (last accessed 5th May 2026).
19. Global Administrative Areas (GADM) (2026), *GADM database for Romania*, available from <https://gadm.org> (last accessed on 29th April 2026).
20. Grelle, G., Donnarumma, A., and Revellino, P. (2013), *Slope angles as an indicator parameter of landslide susceptibility in a geologically complex area*, in Proceedings of the 2nd World Landslide Forum, pp. 425-433, Doi:10.1007/978-3-642-31325-7_56.
21. Guzzetti, F., Galli, M., Reichenbach, P., Ardizzone, F., and Cardinali, M. (2006), *Landslide hazard assessment in the Collazzone area, Umbria, Central Italy*, Natural Hazards and Earth System Sciences, 6, pp. 115-131, Doi:10.5194/nhess-6-115-2006.
22. Guzzetti, F. and Vennari C. (2007), *Rainfall threshold for possible occurrence of landslides in Italy*, Meteorology and Atmospheric Physics, 98(3), pp. 239-267, Doi:10.1007/s00703-007-0262-7.
23. Hastie, T. and Tibshirani, R. (1986), *Generalized additive models*, Statistical Science, 1(3), pp. 297-310, Doi:10.1214/ss/1177013604.
24. Hastie, T.J. and Pregibon, D. (1990), *Generalized Linear Models*, in Chambers J.M. and Hastie T.J. (eds), *Statistical Models in S. Pacific Grove (CA)*, Wadsworth & Brooks/Cole, Chapter 6, Doi:10.1201/9780203738535-6
25. Hosmer, D.W. and Lemeshow, S. (2000), *Applied logistic regression*, 2nd ed., Wiley, New York, Doi:10.1002/0471722146.
26. Huh, M.H. (2015), *Kernel-trick regression and classification*, Communications for Statistical Applications and Methods, 22(2), pp. 207-221, Doi:10.5351/CSAM.2015.22.2.201.
27. Illarionova, L., Dubrovin, K., Fomia, E., Stepanov, A., and Verkhoturov, A. (2025), *Multi-sensor NDVI time series for crop fallow land classification in Khabarovsk Krai, Russia*, Sensors, 25(18), 5746, Doi:10.3390/s25185746.
28. Jenks, G.F. (1967), *The data model concept in statistical mapping*, International Yearbook of Cartography, 7, pp. 186-190.
29. Jo, E.B., Lee, J.H., and Park, H.J. (2025), *Analysis of the impact of mapping units on landslide susceptibility: A comparative study of grid units and slope units*, EGU General Assembly, Vienna, Austria, EGU25-7844, Doi:10.5194/egusphere-egu25-7844.
30. Liu, H., Song, Y., Yi, W., and Zhang P. (2026), *Degree of geocomplexity for diagnosing spatial pattern loss beyond prediction accuracy*, GIScience & Remote Sensing, 63 (1), Doi:10.1080/15481603.2026.2657087.
31. Mann, H.B., and Whitney D.R. (1947), *On a test of whether one of two random variables is stochastically larger than the other*, Annals of Mathematical Statistics, 18 (1), pp. 50-60. Doi:10.1214/aoms/1177730491.
32. Morariu T., and Velcea V. (1971), *Principii și metode de cercetare în geografia fizică*, Edit. Academiei, București.
33. OpenStreetMap contributors (2026), *OpenStreetMap data extract for Study area*, downloaded via Geofabrik portal, available from <https://download.geofabrik.de> (last accessed on 18th May 2026)

34. Pearce, S.C. (1992), *Introduction to Fisher (1925). Statistical Methods for Research Workers*, in Kotz. S. and Johnson, N.L. (eds), *Breakthroughs in statistics*. Springer Series in Statistics, Springer, New York, Doi:10.1007/978-1-4612-4380-9_5.
35. Pearson, K. (1895), *Note on regression and inheritance in the case of two parents*, Proceedings of the Royal Society of London, 31, 58 (347-352), pp. 240-242, Doi:10.1098/rspl.1895.0041.
36. Pham, B.T., Bui, D.T., Prakash, I., Nguyen, L.H., and Dholakia, M.B. (2017), *A comparative study of sequential minimal optimization-based support vector machines, vote feature intervals, and logistic regression in landslide susceptibility assessment using GIS*, Environmental Earth Sciences, 76, 371, Doi:10.1007/s12665-017-6689-3.
37. Poddar, I. and Roy, R. (2026), *Assessing landslide susceptibility mapping in the Sikkim Himalayas using an ensemble machine learning approach*, Discover Geoscience. 4, 14, Doi:10.1007/s44288-025-00375-4.
38. Putriani, E., Wu, Y.M., and Chen C.W. (2026), *Multivariate statistical assessment for mapping landslide susceptibility: A case study of the Yilan area, Taiwan*, Geoenvironmental Disasters, 13, 31, Doi:10.1186/s40677-026-00384-6.
39. QGIS Development Team (2026), *QGIS Geographic Information System software*, available from: <https://qgis.com> (last accessed on 5th March 2026).
40. Riera, J., Magnuson, J., Vande Castle, J., and MacKenzie, M. (1998), *Analysis of Large-Scale Spatial Heterogeneity in Vegetation Indices among North American Landscapes*, Ecosystems, 1, pp. 268-282, Doi:10.1007/s100219900021.
41. Salcedo-Sanz, S., Rojo-Álvarez, J.L., Martínez-Ramón, M., and Camps-Valls, G. (2014), *Support vector machines in engineering: An overview*, WIREs Data Mining and Knowledge Discovery, 4, pp. 234-267, Doi:10.1002/widm.1125.
42. Salman, H.A., Kalakech A., and Steiti A. (2024), *Random Forest Algorithm*, Babylonian Journal of Machine Learning, 2024, pp. 69-79. Doi:10.58496/BJML/2024/007.
43. Shang, G., Yuan, Q., Zhang, X., Sun, M., Liu, S., Hu, Y., Yan, Z., Gao, Y., and Zhang C. (2024), *Remote sensing inversion of land surface temperature for cloud coverage areas based on NDVI in the North China Plain*, International Journal of Remote Sensing, 45 (19-20), 7361-7376, Doi:10.1080/01431161.2023.2216854.
44. Sidle, R.C. and Bogaard, T.A. (2016), *Dynamic earth system and ecological controls of rainfall-initiated landslides*, Earth Science Reviews, 159, pp. 275-291, Doi:10.1016/j.earscirev.2016.05.013.
45. Slocum, T.A., McMaster, R.B., Kessler, F.C., and Howard, H.H. (2022), *Thematic Cartography and Geovisualization*, 4th edition, CRC Press, Doi:10.1201/9781003150527.
46. Sokolova, G.G. (2016), *The influence of terrain latitude, slope exposure, and slope degree on plant spatial distribution*, Acta Biologica Sibirica, 2(3), pp. 34-45, Doi:10.14258/abs.v2i3.1453.
47. Sokolova M., Japkowicz N., and Szpakowicz S. (2006), *Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation*, in: Sattar, A.

- and Kang, B.H. (eds) *Advances in Artificial Intelligence. AI 2006. Lecture Notes in Computer Sciences*, Vol. 4304, Springer, Berlin, Heidelberg.
Doi:10.1007/11941439_114.
48. Sokolova M. and Lapalme G. (2009), *A systematic analysis of performance measures for classification tasks*, Information Processing & Management, 45, 4, pp. 427-437, Doi:10.1016/j.ipm.2009.03.002.
 49. Titti, G., Hu, L., Festi, P., Elia, L., Borgatti, L., and Lombardo L. (2025), *An updated version of the spatial-zoning plugin: From space to space-time data-driven modeling in QGIS*. International Journal of Applied Earth Observation and Geoinformation, Vol. 142, 104679, Doi:10.1016/j.jag.2025.104679.
 50. Titti, G., Sarretta, A., Lombardo, L., Crema, S., Pasuto, A., and Borgatti, L. (2022), *Mapping susceptibility with open-source tools: A new plugin for QGIS*, Frontiers in Earth Science, 10, 842425. Doi:10.3389/feart.2022.842425.
 51. Tukey, J.W. (2008), *Exploratory data analysis*, in *The concise encyclopedia of statistics*, Springer, New York, Doi:10.1007/978-0-387-32833-1_136.
 52. Vatcheva, K.P., Lee, M., McCormick, J.B., and Rahbar M.H. (2016), *Multicollinearity in regression analysis conducted in epidemiologic studies*, Epidemiology (Sunnyvale), 6 (2), 227, Doi:10.4172/2161-1165.1000227
 53. Western, A.W., Grayson, R.B., Blöschl, G., Willgoose, G.R., and McMahon, T.A. (1999), *Observed spatial organization of soil moisture and its relation to terrain indices*, Water Resources Research, 35 (3), pp. 797-810, Doi:10.1029/1998WR900065.
 54. Wood, S.N. (2006), *Generalized additive models: An introduction with R*, Chapman & Hall/CRC, Boca Raton, Doi:10.1201/9781420010404.