https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

Feature Selection for Landslide Forecasting Models in Southern Andes

Manuel Labbé[1], Millaray Curilem[1], Ivo Fustos-Toribio[2], and Mario Pooley[2]

[1]Department of Electrical Engineering. Universidad de La Frontera. Temuco, Chile
[2]Department of Civil Engineering. Universidad de La Frontera. Temuco, Chile

Correspondence: Millaray Curilem (millaray.curilem@ufrontera.cl) and Ivo Fustos-Toribio (ivo.fustos@ufrontera.cl)

Abstract. Rainfall-induced landslide (RIL) forecasting is crucial for early warning systems developed to mitigate the devastating impacts of these events on human lives, infrastructure, and the environment. Currently, dense instrumental networks for early warning require large datasets to identify precursor patterns in current machine learning models. Topographic, lithological, vegetation, soil moisture, and climatic characteristics are among the most commonly used variables for training these models. However, there are no universal designs, so it is necessary to adapt the requirements to each context and to the available variables that characterise it. To develop a RIL forecasting model for the Southern Andes, this study gathers data from various local soil and climate databases to identify the most relevant variables. Feature selection is crucial for improving the design of machine learning models, reducing the dimensionality of input data, enhancing computational efficiency, and preventing overfitting. We assessed the impact of various features, both individually and in combination, on the performance of predictive models. Methods such as Classification and Regression Tree and Genetic Algorithms are employed to perform the feature selection. A national landslide database was enriched using techniques such as buffer control sampling, PU Bagging, and clustering methods to incorporate negative examples (non-landslide) data. Various predictive models were tested. The results reveal some consistent variables as the most significant in forecasting landslides in four southern Chilean regions.

1 Introduction

Rainfall-induced landslides pose a significant threat to communities worldwide, causing loss of life, property damage, and disruption of essential infrastructure. As climate change continues to alter precipitation patterns worldwide, the frequency and severity of these events are likely to increase, making it imperative to monitor the triggering conditions that lead to such disasters. The current manuscript propose an standardized methodology to establish the main features that controls landslide generation.

Quantification and measurement of variables that control landslide generation are crucial to effectively monitoring landslides, as they provide the necessary data to predict occurrences and assess risks associated with these geological hazards.

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

Various studies have shown that the integration of quantitative approaches can significantly enhance landslide prediction models (Schlögl et al., 2025; Lu et al., 2024; Kumar et al., 2023; Ma et al., 2021; Merghadi et al., 2020). In Steger et al. (2022), the authors highlight the importance of accurate input data and model transparency in predicting shallow landslides, revealing that seasonal precipitation patterns influence landslide triggers differently across seasons, thus necessitating precise data collection and analysis. Similarly, some works emphasize the role of numerical simulations in evaluating landslide risks, which allows the incorporation of uncertainty factors in geotechnical engineering, thus facilitating a quantitative assessment of potential landslide consequences (Jia et al., 2023). Furthermore, hydrological modeling methodologies uses rainfall data to predict shallow landslides (Bezak et al., 2019), underscoring the importance of temporal data in understanding landslide dynamics.

The drivers and controls of landslide generation exhibit both similarities and differences across various global regions, shaped by a complex interplay of geological, climatic, and anthropogenic factors. At a fundamental level, topography is a significant factor influencing landslide occurrence; however, its importance can vary widely depending on local conditions. For instance, Lin et al. (2016, 2017) note that while topography is often emphasized in landslide susceptibility models, soil moisture emerges as a critical factor on a global scale, particularly in regions prone to rainfall-triggered landslides. Additionally, geological characteristics such as lithology and drainage density are crucial conditioning factors that interact with triggering mechanisms like precipitation and seismic activity (Bisht and Rawat, 2023).

Climate change further complicates the landscape of landslide drivers, as increasing temperatures can lead to more intense and frequent rainfall events, thereby heightening landslide risks in susceptible areas (Crozier, 2010). This is particularly evident in regions like the Himalayas, where rapid glacial melt and increased precipitation contribute to heightened landslide activity (Ballabh et al., 2014). Conversely, in areas like the Southern Andes, seismic activity plays a more pronounced role, with earthquakes serving as significant triggers for landslides (Marc et al., 2016; Fan et al., 2021).

Moreover, human activities such as deforestation, urbanization, and land-use changes exacerbate the vulnerability of slopes to landslides by altering natural drainage patterns and destabilizing soil structures (Xu et al., 2024; Wang et al., 2023). The interaction of these diverse factors underscores the necessity for region-specific assessments and models that account for local geological and climatic conditions, as well as anthropogenic influences, to effectively predict and manage landslide hazards worldwide.

Landslides in the Southern Andes are predominantly influenced by a combination of geological, climatic, and tectonic factors that interact to create a complex landscape prone to instability. The region's geological composition, characterized by steep slopes and diverse lithologies, plays a significant role in landslide susceptibility. Studies indicate that the presence of certain rock types and the geomorphological features of the terrain can predispose areas to landslides, particularly during seismic events (Pánek et al., 2022; Serey et al., 2019). Seismic activity is a critical trigger for landslides in this region, as evidenced by the correlation between earthquake occurrences and landslide events (Serey et al., 2019, 2020). The 2010 Maule earthquake, for example, resulted in numerous landslides, highlighting the impact of seismic forces on slope stability.

Additionally, climatic factors, particularly rainfall, are pivotal in triggering landslides in the Southern Andes. Research has shown that intense and prolonged rainfall events significantly increase the likelihood of landslides, especially in areas lacking adequate drainage systems (Fustos-Toribio et al., 2022; Islam et al., 2021). The development of rainfall-induced landslide

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

early warning systems (RILEWS) has been proposed as a means to mitigate the risks associated with these events, although such systems are not yet operational in the Southern Andes (Fustos-Toribio et al., 2022). Furthermore, the interaction between vegetation and slope stability is also noteworthy; while vegetation can stabilize slopes, its removal or degradation can lead to increased landslide occurrences (Vorpahl et al., 2012).

Machine learning (ML) has emerged as a pivotal tool in landslide research, particularly in the development of region-specific susceptibility analysis or early warning systems (Merghadi et al., 2020; Ma et al., 2021; Lu et al., 2024). The complexity of landslide phenomena, influenced by various geological and climatic conditions, requires tailored approaches to make accurate predictions. Recent studies indicate that ML models, such as Support Vector Machines (SVM) (Huang et al., 2018), Random Forests (RF) (Park and Kim, 2019), and Extreme Gradient Boosting (XGB) (Yang et al., 2024) among others, outperform traditional statistical methods in terms of predictive accuracy and robustness (Li et al., 2019; Bravo-López et al., 2023). Approaches that integrate various ML techniques have shown to enhance spatial agreement in mapping landslide susceptibility, thus improving the reliability of predictions (Adnan et al., 2020; Bravo-López et al., 2023). Feature selection plays a crucial role in optimizing model performance by identifying the most relevant variables that contribute to landslide occurrences, thus reducing noise and enhancing interpretability (Halder et al., 2025; Ge et al., 2024; Nirbhav et al., 2023; Yousefi et al., 2024; Pham et al., 2020). This not only improves the classification performance but also provides keys to understanding the phenomena and to decide which variables to monitor.

In the context of southern Chile, the present work proposes the application of ML methods to analyze a comprehensive database comprising 136 features—ranging from soil properties to climatic conditions, underscoring the importance of region-specific studies in assessing landslide susceptibility. Classification and Regression Trees (CART) and Genetic Algorithms (GA) were considered for feature selection, facilitating the identification of local critical factors influencing landslide risks (Shirzadi et al., 2018; Miao et al., 2022).

A comprehensive understanding of the complex interplay between rainfall intensity, soil moisture, and geological factors is crucial for developing effective monitoring networks and generating accurate early warning systems for landslides. This manuscript presents an in-depth analysis of the conditioning factors that influence landslide occurrence in a region of the southern Andes (38–42°S) as a pilot case study. It highlights the importance of the critical variables that must be considered in future monitoring approaches based on machine learning.

2 Study zone

To test our development, we consider a sparse data area, focused in the Southern Andes, spanning latitudes from 38°S to 43°S (Fig. 1). The area presents a unique and complex geological and climatological landscape that significantly influences soil moisture dynamics and mass wasting processes. This region is characterized by a diverse geological composition, including volcanic rocks, sedimentary formations, and glacial deposits, contributing to various soil types and structures. The interplay of these geological features with climatic conditions creates a dynamic environment that is crucial for understanding hydrological processes.

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

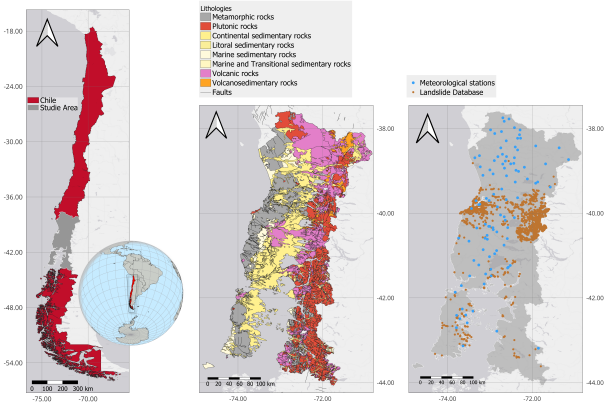

Figure 1. Study zone. Left: National-scale map indicating the spatial coverage of the database. Middle: Zoom to study zone with geology as background, showing the heterogeneity of media in this study. Right: Rainfall-induced landslide (orange) and weather stations (blue) demonstrating the reduced spatial coverage.

The geological framework of the Southern Andes is predominantly shaped by tectonic activity, resulting in a range of rock types, including andesites, basalts, and sedimentary rocks. The presence of glacial deposits from the last glacial maximum has resulted in the formation of heterogeneous soil profiles, which vary in texture and composition throughout the region. These geological characteristics influence water retention and drainage properties, thereby affecting soil moisture levels and slope susceptibility to mass wasting events. Understanding the geological context is essential for assessing the stability of slopes and predicting potential landslide occurrences. Moreover, the climate of the Southern Andes shows a high variability, influenced by altitude, latitude, and prevailing weather patterns. The region experiences a range of climatic conditions, from arid to humid, with significant precipitation occurring primarily in the winter months. This variability leads to distinct wet and dry seasons, profoundly affecting soil moisture dynamics at different depths with unknown control. Moreover, climate change poses additional challenges, as shifts in precipitation patterns and increased frequency of extreme weather events may exacerbate soil erosion and landslide risks.

The area is highly influenced by volcanic eruptions of different style, ranging from plinian to strombolian. These eruptions generate different volcanic soils derived from the tephra degradation during the holocene. Soil variability in the Southern Andes is characterized by differences in texture, moisture retention capacity, and organic matter content. The region hosts a range of soil types, including Andosols, which are rich in volcanic ash and exhibit high moisture retention, and more sandy or gravelly soils that drain quickly. This variability complicates the prediction of soil moisture levels and their influence on slope stability. Current knowledge gaps exist about the specific relationships between soil texture, moisture dynamics, and susceptibility to mass loss.

Despite the wealth of geological and climatological data available, significant gaps remain in our understanding of soil moisture dynamics and their parametric controls, that could control mass wasting. Specifically, there is a need for comprehensive studies that integrate geological, climatic, and soil data to develop a holistic understanding of how these factors interact to

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

influence soil moisture levels and slope stability. Additionally, the impact of invasive species on soil moisture and erosion processes has not been extensively studied, presenting an opportunity for future research to explore these interactions.

3 Database

To rigorously assess the interplay between rainfall intensity, soil moisture, and geological factors that underlie landslide occurrences, it is essential to construct a comprehensive and reliable database. We designed and developed databases that synthesise data sets, including meteorological records, modelled soil moisture, and detailed hydraulic soil properties. We used an approach to establish the critical conditioning factors that could be systematically captured and quantified, thereby providing a solid empirical foundation for subsequent machine learning-based risk assessments. In constructing the database, considerable attention was paid to data quality, temporal resolution, and spatial consistency, which are critical for capturing the transient nature of the environmental processes leading to landslides in Southern Andes.

We considered the soil moisture, precipitation, and slope from the ERA5 database (ERA5, 2023), one of the most widely used climate datasets, with 10km resolution. The slope was obtained combining the ERA5 database with high-resolution digital elevation models. The Chilean soil properties were extracted from the CLSoilMaps database (Dinamarca et al., 2023). The database has soil properties at 100 meters of spatial resolution, being trained using random forest at six standard depths (Table 2), following the GlobalSoilMap standards.

For all the characteristics except for the slope, PP and soil moisture values, the measures available were obtained at the six different soil depths (Table 2), totalling 133 soil features in addition to the 3 extracted from ERA5, giving 136 features for each geographical point.

We used a rainfall-induced landslide database developed using the Xterrae database (https://www.xterrae.cl/) maintained by the National Geology and Mining Service (SERNAGEOMIN) of Chile. Positive landslide events were obtained from this source, while negative cases were generated using a Buffer Control Sampling (BCS) strategy (Gu et al., 2024). In this approach, a buffer zone of 100 meters was established around each landslide event, and sampling points were extended outward up to a 20-kilometer radius. This distance was intentionally chosen given the 10-kilometer resolution of the climatological variables, enabling the capture of spatial variability of the precipitation and soil moisture content, differentiating between landslide and non-landslide conditions. Subsequent refinement of the negative examples was performed via a modified PU Bagging (Positive-Unlabeled Bagging) (Gu et al., 2024), built on the idea of bagging (bootstrap aggregating) by repeatedly sampling subsets of the unlabeled data and combining them with the positive set to train multiple base classifiers. Each subset is treated as if the unlabeled instances were negative (which may introduce noise), and by aggregating the predictions across many such classifiers, PUBagging reduces the bias introduced by this assumption and improves robustness. The final decision is typically made through majority voting or averaging across these classifiers. Here a Random Forest model—trained exclusively on positive examples—was used to evaluate the negative cases. Negative instances yielding a prediction score above 0.5 were removed, ensuring a more robust set of non-landslide cases.

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

File abbreviation	Description	unit
alpha	"alpha" shape parameter (SD)	$1/cm$
AvMoist	Available Moisture at FC-PWP	cm^3/cm^3
AWC	Available water capacity at FC-PWP	mm
Bulk	bulk density of the fine fraction	g/cm^3
Clay	Clay content	%
FC	Field capacity at 330kPa	cm^3/cm^3
ksat	saturated hydraulic conductivity	cm/day
n	"n" shape parameter	-
PIRange_Bulk	prediction intervals for Bulk density	g/cm^3
PIRange_Clay	prediction intervals for clay properties	%
PIRange_Sand	prediction intervals for sand properties	%
PP	precipitation	mm
PWP	Permanent wilting point at 15000kPa	cm^3/cm^3
Sand	Sand content	%
Silt	Silt content	%
slope	slope	$degrees$
Tex_Class	soil textural classes	%
theta_r	residual water content	cm^3/cm^3
theta_s	saturated water content	cm^3/cm^3
Total_AWC	Sum of AWC across all depths	mm
VMoist	Moiture value	g/m^3

Table 1. List of Variables available in the databases

Symbol	Meaning
a	0–5cm
b	5–15cm
c	15–30cm
d	30–60cm
e	60–100cm
f	100–200cm
_	mean
.	standard deviation

Table 2. Symbols used to represent the soil variables designation.

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

After excluding rows with missing values from the initial 3388 records, the final database comprises 3148 examples, with 1618 positive and 1530 negative instances. Each record includes 136 normalized feature columns and a 137th column indicating the binary landslide status. Although explicit vegetation indices, such as NDVI, were not part of the dataset, the effects of vegetation are implicitly incorporated through predicted soil properties that are influenced by NDVI measurements, reflecting parameters like organic matter and soil moisture. Notably, approximately 95% of the data is concentrated within the Los Lagos and Los Ríos regions, providing a regionally focused dataset for further analysis.

4 Methodology

To examine the intricate relationships between rainfall intensity, soil moisture, and soil characteristics in rainfall-induced landslide generation, as well as to support the development of robust monitoring networks and reliable early warning systems for landslides, we propose to identify the most significant variables in landslide detection. First, highly correlated variables were removed from the initial database. Then, two parallel feature selection strategies were applied: a filter and a wrapper approaches. Finally, optimized data sets were used to design the landslide classifiers, as shown in Fig. 2.

The filter-based method uses a Classification and Regression Tree (CART) to rank variables according to their individual discriminatory power between positive and negative cases (Azad et al., 2024). In contrast, the wrapper-based approach optimizes a subset of features by evaluating their combined contribution to classification performance. In this case, the subset optimisation is performed through a genetic algorithm (GA) that evolves to find the best feature combination that minimizes an objective function (Ghosh et al., 2023). The objective function was designed to reduce both, the number of selected variables and the classifier error, evaluated using three models: Support Vector Machines, Random Forest, and XGB. The filter and wrapper strategies will be independently evaluated by assessing the quality of optimised datasets through the performance of these classifiers, ensuring a consistent estimation of the relevance of each contributing variable.

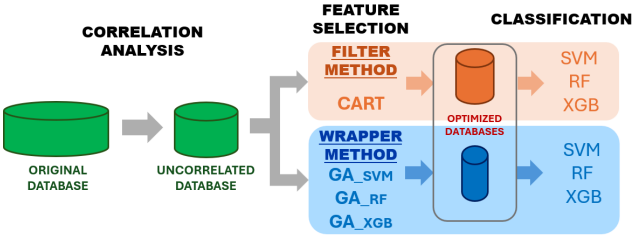

Figure 2. General methodology to identify the most relevant variables in landslide estimation: first, correlated variables are removed. Then, two feature selection methods (Filter and Wrapper) are applied. Finally, different classifiers are trained and validated using the selected features. The best subsets of features are obtained from the best classifiers performance.

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

4.1 Correlation analysis

Before performing the feature selection and because several features measure the same variable at different depths, it was necessary to perform a correlation analysis to remove the more correlated variables. Eliminating correlated variables in machine learning is important because they provide redundant information, which increases model complexity without adding value. Moreover, high correlation may cause issues like multicollinearity in linear models, leading to unstable coefficients, increase the risk of overfitting and hide the importance of other features, negatively affecting generalization.

Figure 3 shows the correlation map of the whole database, where the variables are ordered as describe in Table 1 with their different depths and Fig. 4 shows the correlation values of the remaining variables, after filtering the correlated ones. In Fig. 3 it can be seen that there is a high correlation in subgroups of variables like, as expected, the same variables at different depths, but also the average moisture (AvMoist) with the available water capacity (AWC) for example. Of the initial 136 features, applying a correlation threshold of $|0.9|$ led to the removal of 88 highly correlated features, resulting in a reduced set of 53 features for subsequent filtering and wrapper-based selection. The remaining variables are presented in Table 3.

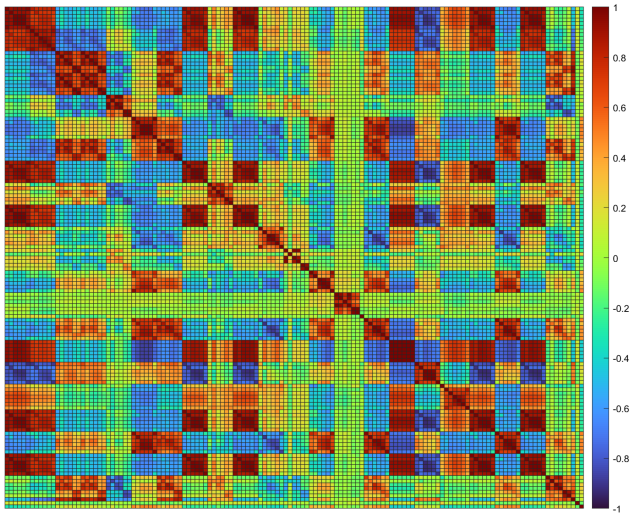

Figure 3. Correlation Matrix for the 136 variables. It can be observed as groups of variables are redundant.

4.2 Feature Selection

The core of this study lies in feature selection, aiming to identify the most influential variables involved in landslide generation in the Southern Andes. Feature selection reduces model complexity, accelerates training, enhances generalization, and improves interpretability while mitigating overfitting (Ge et al., 2024; Ebrahimi et al., 2025; Zheng et al., 2021). To ensure robustness and consistency, we adopt redundant feature selection procedures considering both filter and wrapper methods (Song et al., 2025). The filter method is based on CART while the wrapper method assesses through GA feature subsets collectively, leveraging classifier performance to guide selection.

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

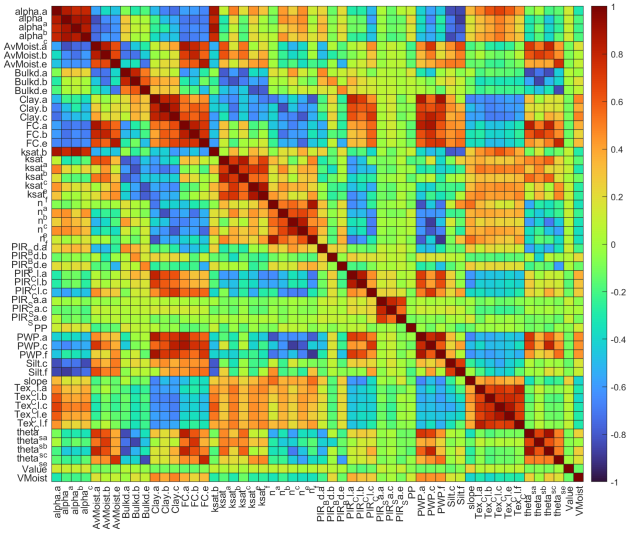

Figure 4. Correlation matrix for the 53 less correlated variables (correlation under the threshold of $|0.9|$).

alpha.a	ksat_b	PP
alpha_a	ksat_c	PWP.a
alpha_b	ksat_e	PWP.c
alpha_c	ksat_f	PWP.f
AvMoist.a	n_a	Silt.c
AvMoist.b	n_b	Silt.f
AvMoist.e	n_c	slope
Bulkd.a	n_e	Tex_Cl.a
Bulkd.b	n_f	Tex_Cl.b
Bulkd.e	PIR_Bd.a	Tex_Cl.c
Clay.a	PIR_Bd.b	Tex_Cl.e
Clay.b	PIR_Bd.e	Tex_Cl.f
Clay.c	PIR_Cl.a	theta_s_a
FC.a	PIR_Cl.b	theta_s_b
FC.b	PIR_Cl.c	theta_s_c
FC.e	PIR_Sa.a	theta_s_e
ksat.b	PIR_Sa.c	VMoist
ksat_a	PIR_Sa.e	

Table 3. List of the 53 less correlated variables that remained after the correlation filtering process. The correlation threshold was $|0.9|$.

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

4.2.1 CART filter method

This method was applied to filter out less significant variables in the estimation model by recursively partitioning the data set using a decision tree algorithm (Fig. 5). Each node in the resulting tree is split based on the feature that best separates the data into distinct classes, ranking variables according to their individual contribution to accurate classification. In the CART-based filtering strategy, variables that appear deeper in the decision tree are considered less significant, as they split fewer samples than those selected near the root. Consequently, these lower-ranked features can be discarded to retain only the most informative ones. However, while CART effectively identifies and ranks individual features, it does not assess the combined contribution of multiple variables.

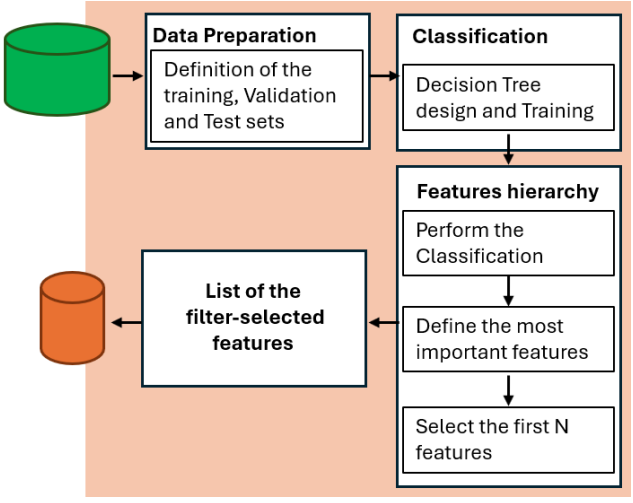

Figure 5. Filter method: general structure of the CART feature selection process. First the database is prepared and the the Decision Tree is trained using the global database. The features are sorted based on their importance scores obtained in the training process. A threshold is defined to retain the top features. At the end of the process a new dataset is created, contains only the CART selected top features.

To perform the split, the Gini Index (or Gini Impurity) metric was used (Eq. 1), measuring how "pure" is the dataset at a node during the tree-building process. It supports the split at each step selecting the features that reduce the impurity in the decision nodes.

$$Gini = 1 - \sum_{k=1}^{C} p_k^2 \tag{1}$$

where, p_k represents the proportion of samples in a node that belong to class k, and C is the number of classes (two in this case). The Gini Index ranges from 0 (a pure node, where all samples belong to one class) to 0.5 (maximum impurity for binary classification, where samples are evenly distributed between the two classes). Finally, the importance of each feature is calculated as the average of the impurity reductions it achieves across all the nodes in the model.

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

The CART design involves tuning seven hyperparameters. The maximum depth of the tree, is adjusted with discrete values: $max\_depth \in \{3,4,5,6,7,8,10,12,15\}$. The minimum number of samples to divide the data set, and the minimum number of leaves of the separation, are defined with $min\_samples\_split \in \{2,3,4,5,7,10,15,20\}$ and $min\_samples\_leaf \in \{1,2,3,4,5,7,10\}$, respectively. The maximum number of features considered at each split, *max_features*, is chosen from $\{sqrt, log2, none\}$. If None, then nodes are expanded until all leaves are pure or until all leaves contain less samples than $min\_samples\_split$. Additionally, *class_weight* balances class weights, the minimum fraction of a leaf is set as $min\_weight\_fraction\_leaf \in \{0.0, 0.1, 0.2, 0.3\}$, and the minimum impurity reduction is defined as $min\_impurity\_decrease \in \{0.0, 0.01, 0.05, 0.1\}$.

4.2.2 GA wrapper method

Genetic algorithms are optimization techniques inspired by natural selection, where an evolution process improves a set of solutions over successive generations. The process begins with a random initial population of potential solutions represented as "chromosomes", each formed by a set of encoded parameters called genes. These solutions are evaluated using a fitness function, which measures how well they perform on the target task. After evaluation, the fittest solutions of each generation have a greater chance of being selected to "reproduce", using genetic operators like crossover (combination of the chromosomes) and mutation (randomly altering a gene). These operators introduce variation, which allows the algorithm to explore new parts of the solution space. Over multiple generations, less optimal solutions are naturally discarded and the population evolves toward the best solution.

In the present study, the chromosome is defined as a sequence of M bits, where M is the number of features remaining after correlation (53 features) and each location in the chromosome corresponds to a specific feature. A 1 in location m indicates that this feature will be considered in the classifier design, while a zero indicates the absence of this feature. At the beginning the values of the chromosomes of the first population were random, so each chromosome proposes a random combination of the features. The evaluation of each chromosome (that is, the evaluation of each combination of features) is carried out using the fitness function, which assigns a fitness value. In our case the fitness values is composed of the classification error and a value that penalizes the number of enabled features, so that the GA seeks the best performance with a small number of variables, to simplify the classification models and to define the most relevant features. The classifiers were implemented using SVM, RF and XGB. The general structure of the GA feature selector is presented in Fig. 6.

The fitness value assigned to each chromosome is given by Eq. 2.

$$FitnessValue = Error + \frac{N_{SF}}{100} \tag{2}$$

where N_{SF} is the number of selected features for a chromosome. The $Error$ index is calculated as:

$$Error = \frac{FP + FN}{N} \tag{3}$$

where FP are the false positives landslides, while FN are the false negative, that is the non detected landslides and N the number of examples.

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

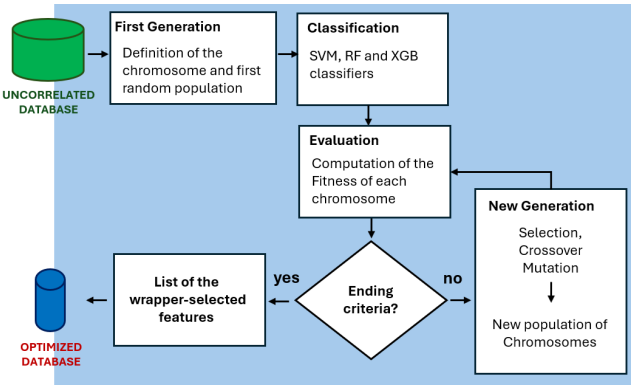

Figure 6. Wrapper method: general structure of the GA feature selection process. The process begins with a population of randomly generated feature subsets, which are used to train classifiers and evaluate their performance. If the stopping criteria are not met, a new population is generated based on the performance of the previous generation, and the evaluation process is repeated. The process continues until either the optimal performance is achieved or the maximum number of generations is reached. At the end of the process, the output is a database containing the wrapper optimized feature subset.

The performance of a genetic algorithm depends on the careful tuning of its hyperparameters. In this study, the population size was set to 50 chromosomes, ensuring sufficient genetic diversity to explore the solution space effectively. The algorithm was allowed to evolve for a maximum of 300 generations, providing ample opportunity for convergence toward an optimal feature subset. For the selection mechanism, we employed tournament selection with a tournament size of 3, favoring fitter chromosomes while maintaining selection pressure. Crossover was performed using a two-point crossover method, in which two parent chromosomes are randomly selected and genetic material between two crossover points is exchanged to produce two offspring. The crossover probability P_{Cros} defines the likelihood that a pair of chromosomes undergoes crossover. Mutation, which introduces random variability and helps prevent premature convergence, was controlled by the mutation probability P_{Mut} and a bit-flip probability P_{Flip}, which governs the likelihood of flipping individual bits within a chromosome. The overall probability of a bit being mutated is calculated using Eq. 4.

$$P_{BitMutation} = P_{Mut} \times P_{Flip} \tag{4}$$

Finally, we implemented an operator to maintain a record of the best individuals across generations, ensuring that high-quality solutions are preserved throughout the evolutionary process. The crossover (P_{Cros}), mutation (P_{Mut}) and flip (P_{Flip}) probabilities were set in a trial and error method, where each evolution was performed with different values and the best fitness defined which probabilities were the best ones. The genetic algorithm was then executed several times.

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

4.3 Classification Methods

To estimate rainfall-induced landslides (RIL), classifiers were developed using the feature sets obtained from both the filter and wrapper selection methods. In the case of the CART-based filter method, classifiers served solely to evaluate the predictive power of the preselected features. However, in the wrapper-based approach, classifiers played a dual role: they acted both as the objective function guiding the genetic algorithm's optimization and as the evaluators of each candidate feature subset's performance. This dual function required the implementation, training, and comparison of over a thousand models to identify the most effective feature combination.

The study systematically developed and refined several SVM, RF and XGB classifier models. For each model type, a grid search was employed to explore a range of hyperparameters, and three-fold cross-validation was used to evaluate model performance and mitigate the risk of overfitting. The subsequent sections provide a comprehensive description of the model configurations and the methodological criteria for feature selection.

4.3.1 Support Vector Machines

SVMs are classification algorithms that find the best-separating hyperplane in high-dimensional space, handling nonlinear data with kernel functions. They're effective for small to moderate datasets with many features, offering good control over overfitting through regularization. However, they're computationally expensive for large datasets and harder to interpret than tree-based methods. Our SVM classifier was designed using 3-fold cross-validation and required tuning two hyperparameters: the cost constant $C = 2^{\alpha}$, where $\alpha \in [-5, 8]$, which controls the trade-off between empirical error and classifier complexity; and the RBF kernel width $\sigma = 2^{\beta}$, where $\beta \in [-7, 5]$, which was adjusted for the Gaussian function.

4.3.2 Random Forest

RF classifiers are designed using an ensemble of decision trees with bagging to improve model accuracy and robustness to overfitting. The RF model was tuned with 9 hyperparameters: number of trees ($n\_estimators \in \{100, 200, 300, 500\}$), maximum depth ($max\_depth \in \{3, 5, 7, 10, 15, 20\}$), minimum samples to split ($min\_samples\_split \in \{2, 5, 10, 15, 20\}$) and leaf ($min\_samples\_leaf \in \{1, 2, 4, 6, 8\}$), maximum features *max_features*, is chosen from $\{sqrt, log2, all\}$, maximum leaves ($max\_leaf \in \{50, 100, 200\}$), minimum weight fraction ($min\_weight\_fraction\_leaf \in \{0.0, 0.1, 0.2\}$), and minimum impurity reduction ($min\_impurity\_decrease \in \{0.0, 0.01, 0.05\}$). We evaluated bootstrapping and non bootstrapping strategies to build the trees.

4.3.3 Extreme Gradient Boosting

XGB model was designed to capture subtle relationships between variables, leveraging its strengths in handling large datasets and complex interactions. The XGB model was tuned with 7 key hyperparameters: number of trees $n\_estimators \in \{100, 200, 300, 500\}$), maximum depth ($max\_depth \in \{3, 5, 7, 9, 11, 13\}$), minimum child weight ($min\_child\_weight \in \{1, 3, 5, 7\}$), minimum loss reduction ($gamma \in \{0, 0.1, 0.2, 0.3, 0.4\}$), learning rate ($learning\_rate \in \{0.01, 0.05, 0.1, 0.15, 0.2\}$), L1 regularization term

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

($reg\_alpha \in \{0, 0.1, 0.5, 1.0\}$), and L2 regularization term ($reg\_lambda \in \{0.1, 0.5, 1.0, 1.5, 2.0\}$). These hyperparameters were optimized to balance model complexity and prevent overfitting, ensuring the XGB model performed accurately and robustly.

5 Results

To understand the relationship between rainfall intensity, soil moisture, and soil hydraulic features in landslide generation, we present an in-depth analysis of the conditioning factors that influence landslide occurrence in a region of the southern Andes (38–42°S). After exploring various machine learning models with different feature configurations, we observed notable differences in performance depending on the features subsets. This process helped identify which combinations of features and modeling approaches best captured the underlying data structure. The results reveal that the dynamic interactions among rainfall intensity, soil moisture patterns, and geological characteristics are significant at the determined exploration depth of the soil, based on the database used during feature selection.

The following section presents the results, highlighting the most promising models and the impact of feature selection on the susceptibility estimation performance.

5.1 Feature selection

Our results show differing model performance across classifiers and feature selection methods (Table 4). The CART approach achieved its best error rate of 26.67% using a relatively large feature set of 15 variables, including swallowing variables such as the average moisture at 0–5 cm and clay content at 5–15 cm and slope. In contrast, the GA-based feature selection methods (GA SVM, GA RF, and GA XGB) delivered improved performance using fewer features.

Both GA RF and GA XGB achieved the lowest error rates of 10.95%, with 6 and 9 features, respectively. Common to both models were the consideration of slope and precipitation variables. At the same time, GA RF also included shallow variables such as bulk density at 5–15 cm and available moisture. The feature selection using GA XGB propose the use of shallow variables at 5-15 cm of bulk density of the fine fraction and the shape parameter of the retention curve and saturated water content (θ_s). The GA SVM method achieved an error rate of 21.98% using 8 features (with $C = 2$ and $\sigma = 0.125$), and selected variables such as the Van Genuchten parameter at 0–5 cm, Average moisture at 60–100 cm, and slope.

There is some overlap in the features selected by GA RF and GA XGB—both models include precipitation, slope, Volumetric moisture, and shallow bulk density (5-15 cm), suggesting that these are key indicators for rainfall-induced landslide assessment. However, differences in feature sets indicate distinct modelling perspectives. GA RF uniquely selected saturated hydraulic conductivity at the first centimetre (0-5 cm) and depth of Permanent wilting point at 100-200 cm, pointing to a focus on near-surface hydraulic behaviour and deeper root-zone water retention. In contrast, GA XGB favoured variables like shape parameter of SWRC at 5-15 and 100-200 cm, soil textural classes at 100–200 cm, and saturated water content at 0-5 cm and 15-30 cm, suggesting a broader interest in soil texture and porosity across both shallow and deep layers.

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

Compared to the CART approach, the GA-based methods selected more compact and targeted feature sets while delivering superior predictive performance. While CART relied on a larger number of features, GA RF and GA XGB focused on key physical properties such as slope and Bulkd.b, which are highly relevant for landslide modeling. These findings highlight the strength of GA-based feature selection in isolating meaningful variables and enhancing landslide prediction accuracy by leveraging both shallow and deep soil characteristics.

Table 4. Results of the evolution: best sets of variables obtained for each evolution according to the classifier and their best errors obtained on the test set

Discriminator	Parameters	Hyperparameters	Feature Selection	Best Error (%)
CART	P_{Cros} : –	max_depth: 15	AvMoist.a, Clay.b, Clay.c, ksat_e,	26.67 ± 3.45
	P_{Mut} : –	min_samples_split: 4	PIRange_Bulkd.b, PIRange_Bulkd.e,	
	P_{Flip} : –	min_samples_leaf: 1	PIRange_Clay.c, PIRange_Sand.c,	
		max_features: None	PIRange_Sand.e, PP, n_a, n_c, slope, theta_s_e,	
		class_weight: balanced	VMoist	
		min_weight_fraction_leaf: 0.0		
		min_impurity_decrease: 0.0		
GA_SVM	P_{Cros} : 0.8	C: 2	alpha.a, AvMoist.e, Clay.a, PIR_Cl.c, PP, slope,	21.98 ± 3.39
	P_{Mut} : 0.2	σ : 0.125	theta_s_b, VMoist	
	P_{Flip} : 0.5			
GA_RF	P_{Cros} : 0.6	n_estimators: 200	Bulkd.b, ksat_a, PP, PWP.f, slope, VMoist	10.95 ± 2.44
	P_{Mut} : 0.3	max_depth: 20		
	P_{Flip} : 0.3	min_samples_split: 2		
		min_samples_leaf: 1		
		max_features: sqrt		
		min_weight_fraction_leaf: 0.0		
		min_impurity_decrease: 0.0		
		bootstrap: True		
GA_XGB	P_{Cros} : 0.8	n_estimators: 500	Bulkd.b, n_b, n_f, PP, Tex_Class.f, slope,	10.95 ± 2.44
	P_{Mut} : 0.3	max_depth: 7	theta_s_a, theta_s_c, VMoist	
	P_{Flip} : 0.5	min_child_weight: 1		
		gamma: 0.0		
		learning_rate: 0.2		
		reg_alpha: 0		
		reg_lambda: 0.5		

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

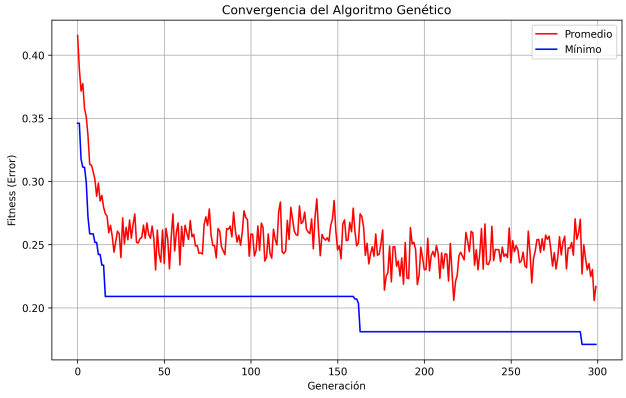

Figure 7. Genetic Algorithm Convergence

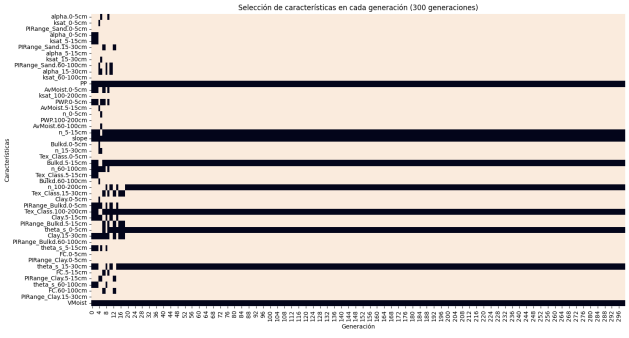

Figure 8. Evolution

5.2 Classification

The features selected by the different methods were used to train and test classifiers implemented with the three models - SVM, RF and XGB - to compare them. The results are presented in Table 5, providing a comprehensive overview of their performance under two distinct conditions: Non PU Bagging and PU Bagging. The evaluation metrics include accuracy and recall, both of which are critical for assessing the effectiveness of these classification models. In the Non PU Bagging category, the classifiers demonstrated notable performance. The SVM achieved an accuracy of 0.735 with a standard deviation of $\pm$ 0.029, and a recall of 0.749 $\pm$ 0.041. The XGB model outperformed the SVM, recording an accuracy of 0.873 $\pm$ 0.020 and a recall of 0.885 $\pm$ 0.028, indicating its strong predictive capability. The Random Forest classifier exhibited the highest performance among the three, with an accuracy of 0.879 $\pm$ 0.020 and a recall of 0.901 $\pm$ 0.027. These results suggest that all classifiers were effective in identifying instances from the dataset, with Random Forest showing the best overall performance. In the PU Bagging category, the classifiers maintained strong performance, although there were slight variations compared to the Non PU Bagging results. The SVM recorded an accuracy of 0.769 $\pm$ 0.0284 and a recall of 0.809 $\pm$ 0.0372, indicating an improvement in its predictive

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

ability with PU Bagging. The XGB model achieved an accuracy of 0.862 ± 0.022 and a recall of 0.889 ± 0.028, which, while slightly lower than in the Non PU Bagging scenario, still reflects robust performance. The Random Forest classifier, however, showed a decrease in accuracy to 0.856 ± 0.023 and a recall of 0.881 ± 0.028, suggesting that while it remained effective, the application of PU Bagging may have introduced some variability in its performance.

The RandomForest discriminator results indicate that all classifiers performed well in both scenarios, with Random Forest consistently demonstrating the highest accuracy and recall. The application of PU Bagging appears to have had a mixed impact on classifier performance, enhancing some models while slightly reducing the effectiveness of others. These findings highlight the importance of evaluating different modeling strategies and their configurations to optimize predictive performance in classification tasks. Future research should explore the underlying factors contributing to these performance variations and consider additional techniques for further enhancing model accuracy and recall.

The model performance shows good agreement. Our results showed that the best-performing model is the XGB classifier utilizing Genetic Algorithm (GA) optimization under the "PU Bagging" approach, achieving an accuracy of 0.896 ± 0.019 and a recall of 0.886 ± 0.026. Conversely, the worst-performing model is the Support Vector Machine (SVM) classifier with GA optimization and "No PU Bagging," which recorded an accuracy of 0.735 ± 0.029 and a recall of 0.749 ± 0.041.

The comparison of these models underscores the stark contrast in their performance. The XGB model demonstrates a significant advantage with a difference of 0.149 in accuracy and 0.137 in recall compared to the SVM with GA (No PU Bagging). This disparity highlights the effectiveness of the XGB model in capturing relationships. Additionally, the GA optimization applied to both models suggests that, while the SVM has a robust foundational framework, it may struggle to leverage hyperparameter tuning effectively compared to the inherent capabilities of XGB. Studies have shown that while SVM can achieve favorable performance metrics in more straightforward datasets, its effectiveness diminishes in more complex contexts unless it is coupled with extensive parameter optimization (Huang et al., 2019).

Our recall results reflected the strengths of these models in identifying true positive instances. XGB's higher recall articulates a better capacity to minimize false negatives than the SVM's recall performance. This is crucial in applications where the detection of positive cases is prioritized, such as potential landslide cases.

6 Discussion

Accurate Landslide forecasting enables early warnings, allowing authorities and communities to take preventive measures, such as evacuations or slope stabilization, to reduce risk. It also supports better land-use planning and disaster preparedness, contributing to long-term resilience against geological hazards. Ultimately, landslide forecasting is a critical tool for safeguarding lives and promoting sustainable development in high-risk areas.

For the first phase of feature selection, expert analysis was considered, along with setting a correlation elimination threshold of $|0.9|$. Subsequently, two strategies were employed to finalize the feature selection: CART and a Genetic Algorithm.

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

Table 5. Results for Non PU Bagging and PU Bagging models with different discriminators

Discriminator	Database	Classifier	Accuracy	Recall
CART	No PU Bagging	SVM	0.740 ± 0.043	0.781 ± 0.058
		XGB	0.860 ± 0.036	0.877 ± 0.045
		RF	0.800 ± 0.039	0.830 ± 0.049
	PU Bagging	SVM	0.763 ± 0.045	0.809 ± 0.057
		XGB	0.862 ± 0.035	0.880 ± 0.046
		RF	0.835 ± 0.039	0.889 ± 0.044
GA (SVM)	No PU Bagging	SVM	0.783 ± 0.026	0.812 ± 0.034
		XGB	0.880 ± 0.021	0.887 ± 0.028
		RF	0.875 ± 0.021	0.883 ± 0.029
	PU Bagging	SVM	0.791 ± 0.025	0.835 ± 0.032
		XGB	0.870 ± 0.021	0.874 ± 0.029
		RF	0.876 ± 0.021	0.890 ± 0.027
GA (RF)	No PU Bagging	SVM	0.735 ± 0.029	0.749 ± 0.041
		XGB	0.873 ± 0.020	0.885 ± 0.028
		RF	0.879 ± 0.020	0.901 ± 0.027
	PU Bagging	SVM	0.769 ± 0.028	0.809 ± 0.037
		XGB	0.862 ± 0.022	0.889 ± 0.028
		RF	0.856 ± 0.023	0.881 ± 0.028
GA (XGB)	No PU Bagging	SVM	0.759 ± 0.027	0.795 ± 0.037
		XGB	0.875 ± 0.021	0.889 ± 0.028
		RF	0.861 ± 0.021	0.881 ± 0.029
	PU Bagging	SVM	0.781 ± 0.025	0.786 ± 0.034
		XGB	0.896 ± 0.019	0.886 ± 0.026
		RF	0.873 ± 0.021	0.872 ± 0.030

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

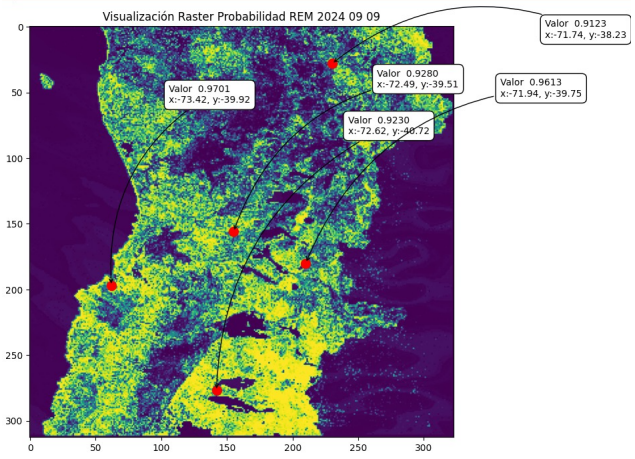

Figure 9. Example Probability of RIL in the Southern Chile Area

6.1 Main controls in mass wasting generation

The accurate determination of parameters to monitor in rainfall-induced landslides is critical for effective risk assessment and management. As rainfall is a primary trigger for landslides, understanding the hydrological and geotechnical factors that influence slope stability is essential. Studies have shown that specific topographic parameters, such as slope angle, soil moisture content, and rainfall intensity, significantly correlate with landslide occurrences (Emberson et al., 2022; Novellino et al., 2021). For instance, the integration of machine learning techniques with hydrological data has enhanced the predictive capabilities of landslide susceptibility models, allowing for a more nuanced understanding of how varying rainfall patterns affect slope stability (Novellino et al., 2021; Mondini et al., 2023). Moreover, the identification of key parameters is not only vital for immediate hazard assessment but also for long-term monitoring and mitigation strategies. Research indicates that rainfall anomalies and their spatial patterns can provide insights into potential landslide activity, thereby facilitating timely interventions (Marc et al., 2019; Wang et al., 2020) The dynamic nature of rainfall events necessitates continuous monitoring of these parameters to adaptively manage landslide risks, especially in regions prone to extreme weather conditions (Bordoni et al., 2015; Yang et al., 2024). Furthermore, the application of remote sensing technologies has proven effective in capturing real-time data on rainfall and its impact on slope stability, thus improving the accuracy of landslide predictions (Kirschbaum and Stanley, 2018; Gariano et al., 2015).

Our results allowed us to identify the recurrent variables related to rainfall-induced landslides in the Southern Andes. Specifically, we observed that precipitation, slope, and volumetric moisture (VMoist) are repeatedly selected features (Table 4). This recurrence suggests that these parameters capture key underlying characteristics crucial to achieving robust predictive performance. In particular, the consistency of these features across standard decision tree approaches and genetic algorithm–enhanced methods indicates their stability and high predictive relevance. Our results show consistency with the particular conditions of the Southern Andes, where the interaction between volumetric moisture (VMoist) and slope angle significantly influences

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

landslide generation by affecting the shear strength and stability of the soil mass. High VMoist indicates near-saturated soil conditions, which lead to increased pore water pressures and a reduction in matric suction. These hydrological effects decrease the effective stress in the soil, thereby reducing its shear strength —a critical factor that predisposes slopes to failure. In regions with steep slopes, the gravitational forces acting downslope are significantly heightened, amplifying the destabilising effects of high moisture content, such as San Jose de Maipo (Maragaño et al., 2023), Osorno Volcano (Fustos-Toribio et al., 2022) and Villa Santa Lucia ((Somos et al., 2020),(Ochoa et al., 2025)).

Another key variable selected by the automatic feature selection allowed for the reflection of the central control of the Southern Andes. The soil moisture contents in the 0-5 cm layer showed a critical variable in initiating infiltration when rainfall impacts these layers. These automatic selection shows concordance with the processes of water movement control proposed by Maragaño et al. (2023), where high moisture content significantly influences local pore water pressures, a factor that has been directly linked to reduced shear strength and slope failure (Cheng et al., 2022). Similarly, deeper measurements, such as those denoted by variables with suffixes between 60 up to 200 cm could capture the infiltration dynamics and moisture storage capacity that modulate the transmission of rainfall energy and fluid pressure through the soil profile. We suggest considering the incorporation of depth-specific measurements to ensure that the models capture both the rapid near-surface processes and the more delayed responses in the underlying zones. Consequently, the synthesised feature sets, along with well-tuned hyperparameters for each classifier, not only optimise predictive performance (as evidenced by the reported best errors) but also reveal the relative.

6.2 Lesson learned about classification

Using a selection features approach, we perform a susceptibility landslide assessment focused on the Southern Andes (Table 5). Our results generated a landslide susceptibility modelling considering the depth-dependent heterogeneity of soil properties and their impact on the landslide controls. The use of different classifiers (e.g., CART, GASVM, GARF, GA XGB) with tuned hyperparameters attests to the complexity of modelling the interaction between multilayered soil properties and water infiltration dynamics. An accurate selection of features from various depths helps to predict transient hydrological responses—for example, rapid saturation of near-surface layers followed by a more gradual moisture redistribution in deeper layers (Yao et al., 2025). Our results allowed us to constrain the data necessary to perform landslide susceptibility for the Southern Andes, if required, at a better scale into the future. We interpret the results of the selected features into the landslide classification as the interplay between soil physical properties such as texture (content between 5-30 cm, two layers) and saturated hydraulic conductivity up to 100 cm, which are known to be depth-dependent and dependent on the soil moisture content.

The quality of the results achieved for the Southern Andes landslide assessment using machine-learning classifiers must be considered in light of the inherent variations in density information and database quality. Previous studies in South America encourage improving the landslide database and its controlling factors due to reduced records along history (Fustos-Toribio et al., 2022; Sepúlveda and Petley, 2014). Moreover, no unique classifier method guarantees optimal performance for all datasets; model performance is heavily influenced by the distribution and quality of the input data (Gu et al., 2024). The presented results reveal that, while all classifiers delivered strong performance, nuances such as the slight improvement for Support

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

Vector Machine (SVM) under Proportional Upweighting (PU) Bagging and the marginal reduction of accuracy for Random Forest in the same condition underscore the sensitivity of these methods to changes in data handling and sampling strategies (Zhang et al., 2024; Zhou et al., 2023). Our findings demonstrate that a promising classifier can achieve effective performance even with a limited database when supported by robust density information and high-quality underlying data (Table 5).

Stand out the Random Forest classifier, which demonstrates the highest accuracy and recall in the Non PU Bagging condition—achieving 0.879 accuracy and 0.901 recall. It appears to be the more robust choice for landslide susceptibility mapping in regions where data quality is consistent. However, its slight performance reduction in the PU Bagging scenario suggests that in heterogeneous environments, it may be beneficial to consider additional strategies to mitigate variability. Techniques such as ensemble stacking or hybrid integration, which capitalise on the strengths of multiple algorithms, could further enhance predictive performance, as shown in other studies (Huang et al., 2022)). Therefore, while Random Forest seems preferable based on current evidence, an adaptive framework that dynamically incorporates ensemble methods might better serve the Southern Andes, where environmental conditions and data quality may fluctuate (Gu et al., 2024).

Future developments must focus on reducing uncertainty in landslide assessments by refining the data acquisition process and enhancing model robustness. Possibilities include improved non-landslide sampling strategies, expanding our proposed approach, and integrating remote sensing data to capture finer spatial details, thereby stimulating advances in model performance (Gu et al., 2024). Moreover, the incorporation of hybrid ensemble techniques and closer analysis of spatial patterns can contribute to more reliable landslide susceptibility maps (Huang et al., 2022). These developments are crucial for enhancing early warning systems in the Southern Andes, as improved classifier accuracy and reliability directly translate into better disaster risk management and timely interventions. The current results offer a promising background, a continued focus on data quality improvement and methodological enhancements is essential for advancing landslide early warning systems in this region.

6.3 Future scope

The impact of climate change on mass wasting events cannot be ignored. Recent studies showed evidence that increased precipitation and temperature fluctuations significantly affect sediment yield and debris flow activity in alpine regions (Hirschberg et al., 2021). Future changes in precipitation patterns will urge the implementation of a monitoring framework that incorporates climatic variables alongside geotechnical measurements to adapt to changing conditions effectively.

The necessity for broader monitoring networks with well-defined variables is paramount in the context of rainfall-induced landslides. Rainfall-Induced Landslides are predominantly triggered by prolonged and intense rainfall, which underscores the importance of capturing detailed hydrological and geological data to enhance predictive capabilities (Li et al., 2021; Fusco et al., 2022). A comprehensive monitoring system that includes variables such as soil moisture, pore water pressure, and rainfall intensity can significantly improve the understanding of slope stability dynamics and the conditions leading to landslide initiation (Abraham et al., 2020; Segoni et al., 2018). For instance, the integration of hydrological monitoring with rainfall data has been shown to refine rainfall thresholds for landslide forecasting, thereby enabling more accurate early warning systems (Teza et al., 2022; Vaz et al., 2018).

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

Moreover, implementing advanced technologies, such as remote sensing and machine learning, can facilitate real-time data collection and analysis, allowing for timely interventions in high-risk areas (Mondini et al., 2023; Froude and Petley, 2018). The development of automated systems that utilize continuous monitoring of critical parameters can enhance the reliability of landslide predictions, as demonstrated by recent studies that have successfully employed deep-learning techniques to forecast landslide occurrences based on rainfall and soil conditions (Abraham et al., 2019; Qiao et al., 2020). Additionally, establishing a network of monitoring stations across diverse geomorphological environments can provide valuable insights into the varying responses of slopes to rainfall, thereby improving the generalizability of predictive models (Kuradusenge et al., 2020; Bortolozo et al., 2024).

7 Conclusions

We evaluated the performance of machine learning models—Support Vector Machines (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) to predict landslide susceptibility in the southern Andes (38–42°S). The Southern Andes has a high soil variability that does not allow for deep instrument monitoring. Therefore, we determined the domain-relevant geotechnical, hydrological, and geomorphological variables and applied advanced feature selection techniques, including Genetic Algorithms (GA) and Positive-Unlabeled (PU) Bagging. We were able to systematically identify and optimise the most informative predictors for rainfall-induced landslides.

Our results demonstrate that models optimised using Genetic Algorithms significantly outperform baseline methods, such as CART with conventional feature selection. Notably, GA RF and GA XGB achieved the lowest classification errors (10.95 %). Moreover, compact feature sets highlighted the potential of evolutionary algorithms to enhance both the accuracy and efficiency of susceptibility assessments. Our models' results consistently identified slope, precipitation, and near-surface soil hydraulic properties—particularly bulk density and saturated water content—as critical factors influencing landslide initiation. Future instrumental developments must consider these variables, monitoring and landslide assessment.

Our findings also underscore the importance of incorporating both shallow and deep soil moisture characteristics, as well as soil retention curve parameters, to capture better the complex subsurface dynamics that precede slope failure. The differences in feature prioritisation between GA RF, GA XGB, and GA SVM reflect distinct modelling philosophies: while RF and XGB emphasised shallow hydraulic traits and retention thresholds, SVM gave greater weight to deeper soil moisture indicators and retention curve shape parameters.

We conclude that integrating data-driven models with physically meaningful features provides a robust framework for enhancing early warning systems and regional risk assessments. The superior performance of GA-optimised ensemble models suggests that future efforts should prioritise hybrid strategies that combine expert knowledge with automated feature selection. These approaches are particularly valuable in data-scarce environments, offering scalable solutions to inform risk management and decision-making in mountainous regions vulnerable to rainfall-triggered landslides.

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

Code availability. The code used for feature selection and classification is available upon request from the corresponding author.

Data availability. The landslide database and soil property data are available from the Chilean National Geology and Mining Service (SERNAGEOMIN) and CLSoilMaps database respectively. Climate data from ERA5 is publicly available through the Copernicus Climate Data Store.

Author contributions. All authors contributed to the study conception and design. Data collection, depuration and analysis were performed by [Ivo Fustos and Millaray Curilem]. Machine learning implementation was carried out by [Manuel Labbé]. Geological interpretation was provided by [Ivo Fustos and Mario Pooley]. All authors contributed to the writing and revision of the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We thank the Chilean National Geology and Mining Service (SERNAGEOMIN) for providing the landslide database through the Xterrae platform. We also acknowledge the Copernicus Climate Data Store for ERA5 data and the CLSoilMaps project for soil property data. This research was supported by FONDEF ID23i10118 and Fondecyt Regular grant 1230792.

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

References

Abraham, M., Pothuraju, D., & Satyam, N.: Rainfall thresholds for prediction of landslides in idukki, india: an empirical approach, Water, 11, 2113, https://doi.org/10.3390/w11102113, 2019.

Abraham, M., Satyam, N., Kushal, S., Rosi, A., Pradhan, B., & Segoni, S.: Rainfall threshold estimation and landslide forecasting for kalimpong, india using sigma model, Water, 12, 1195, https://doi.org/10.3390/w12041195, 2020.

Adnan, M., Rahman, M., Ahmed, N., Ahmed, B., Rabbi, M., & Rahman, R.: Improving spatial agreement in machine learning-based landslide susceptibility mapping, Remote Sensing, 12, 3347, https://doi.org/10.3390/rs12203347, 2020.

Azad, A., et al.: Feature selection for landslide susceptibility mapping using CART algorithm, Engineering Geology, 349, 106899, 2024.

Ballabh, H., Pillay, S., Negi, G., & Pillay, K.: Relationship between selected physiographic features and landslide occurrence around four hydropower projects in bhagirathi valley of uttarakhand, western himalaya, india, International Journal of Geosciences, 5, 1088-1099, https://doi.org/10.4236/ijg.2014.510093, 2014.

Bezak, N., Auflič, M., & Mikoš, M.: Application of hydrological modelling for temporal prediction of rainfall-induced shallow landslides, Landslides, 16, 1273-1283, https://doi.org/10.1007/s10346-019-01169-9, 2019.

Bisht, S. and Rawat, K.: A review of statistical approaches used for landslide susceptibility analysis with the help of remote sensing and gis technology, Acta Scientiarum Polonorum Formatio Circumiectus, 22, 83-96, https://doi.org/10.15576/asp.fc/2023.22.3.13, 2023.

Bordoni, M., Meisina, C., Valentino, R., Lu, N., Bittelli, M., & Chersich, S.: Hydrological factors affecting rainfall-induced shallow landslides: from the field monitoring to a simplified slope stability analysis, Engineering Geology, 193, 19-37, https://doi.org/10.1016/j.enggeo.2015.04.006, 2015.

Bortolozo, C., Pampuch, L., Andrade, M., Metodiev, D., Carvalho, A., Mendes, T., et al.: Arhcs (automatic rainfall half-life cluster system): a landslides early warning system (lews) using cluster analysis and automatic threshold definition, International Journal of Geosciences, 15, 54-69, https://doi.org/10.4236/ijg.2024.151005, 2024.

Bravo-López, P., Fernández, T., Sellers, C., & García, J.: Analysis of conditioning factors in cuenca, ecuador, for landslide susceptibility maps generation employing machine learning methods, Land, 12, 1135, https://doi.org/10.3390/land12061135, 2023.

Cheng, Y., et al.: Effects of soil moisture on landslide susceptibility, Geomorphology, 398, 108045, 2022.

Cicala, L., Gargiulo, F., Parrilli, S., Amitrano, D., & Pigliasco, G.: Progressive monitoring of micro-dumps using remote sensing: an applicative framework for illegal waste management, Sustainability, 16, 5695, https://doi.org/10.3390/su16135695, 2024.

Crozier, M.: Deciphering the effect of climate change on landslide activity: a review, Geomorphology, 124, 260-267, https://doi.org/10.1016/j.geomorph.2010.04.009, 2010.

Dinamarca, A., et al.: CLSoilMaps: A comprehensive soil property database for Chile, Soil Science Society of America Journal, 87, 345-358, 2023.

Ebrahimi, S., et al.: Feature selection methods for landslide susceptibility assessment: A comparative study, Computers & Geosciences, 168, 105234, 2025.

Emberson, R., Kirschbaum, D., Amatya, P., Tanyaș, H., & Marc, O.: Insights from the topographic characteristics of a large global catalog of rainfall-induced landslide event inventories, Natural Hazards and Earth System Science, 22, 1129-1149, https://doi.org/10.5194/nhess-22-1129-2022, 2022.

ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate, Copernicus Climate Change Service, 2023.

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

Fan, X., Yunus, A., Scaringi, G., Catani, F., Subramanian, S., Xu, Q., et al.: Rapidly evolving controls of landslides after a strong earthquake and implications for hazard assessments, Geophysical Research Letters, 48, e2020GL090509, https://doi.org/10.1029/2020gl090509, 2021.

Froude, M. and Petley, D.: Global fatal landslide occurrence from 2004 to 2016, Natural Hazards and Earth System Science, 18, 2161-2181, https://doi.org/10.5194/nhess-18-2161-2018, 2018.

Fusco, F., Bordoni, M., Tufano, R., Vivaldi, V., Meisina, C., Valentino, R., et al.: Hydrological regimes in different slope environments and implications on rainfall thresholds triggering shallow landslides, Natural Hazards, 114, 907-939, https://doi.org/10.1007/s11069-022-05417-5, 2022.

Fustos-Toribio, I., Manque-Roa, N., Antipan, D., Sotomayor, M., & Letelier, V.: Rainfall-induced landslide early warning system based on corrected mesoscale numerical models: an application for the southern andes, Natural Hazards and Earth System Science, 22, 2169-2183, https://doi.org/10.5194/nhess-22-2169-2022, 2022.

Gariano, S., Petrucci, O., & Guzzetti, F.: Changes in the occurrence of rainfall-induced landslides in calabria, southern italy, in the 20th century, Natural Hazards and Earth System Science, 15, 2313-2330, https://doi.org/10.5194/nhess-15-2313-2015, 2015.

Ge, Y., et al.: Machine learning-based feature selection for landslide susceptibility mapping, Engineering Geology, 312, 106935, 2024.

Ghosh, S., et al.: Genetic algorithm-based feature selection for landslide prediction, Natural Hazards, 118, 1247-1268, 2023.

Gu, T., et al.: Improving landslide susceptibility assessment through PU learning and buffer control sampling, Computers & Geosciences, 181, 105467, 2024.

Halder, B., et al.: Feature selection techniques for landslide susceptibility modeling, Geoscience Frontiers, 16, 101589, 2025.

Hirschberg, J., Fatichi, S., Bennett, G., McArdell, B., Peleg, N., Lane, S., et al.: Climate change impacts on sediment yield and debris-flow activity in an alpine catchment, Journal of Geophysical Research: Earth Surface, 126, e2020JF005739, https://doi.org/10.1029/2020jf005739, 2021.

Huang, Y., et al.: Support vector machine for landslide susceptibility mapping, Environmental Earth Sciences, 77, 432, 2018.

Huang, F., et al.: Comparisons of heuristic, general statistical and machine learning models for landslide susceptibility prediction and mapping, Catena, 191, 104580, 2019.

Huang, F., et al.: Ensemble learning for landslide susceptibility assessment, Geoscience Frontiers, 13, 101368, 2022.

Islam, M., Islam, M., & Jeet, A.: A geotechnical investigation of 2017 chattogram landslides, Geosciences, 11, 337, https://doi.org/10.3390/geosciences11080337, 2021.

Jia, L., Wang, J., Gao, S., Fang, L., & Wang, D.: Landslide risk evaluation method of open-pit mine based on numerical simulation of large deformation of landslide, Scientific Reports, 13, 15892, https://doi.org/10.1038/s41598-023-42736-4, 2023.

Kirschbaum, D. and Stanley, T.: Satellite-based assessment of rainfall-triggered landslide hazard for situational awareness, Earth's Future, 6, 505-523, https://doi.org/10.1002/2017ef000715, 2018.

Kumar, R., et al.: Machine learning approaches for landslide susceptibility assessment, Natural Hazards, 119, 1823-1845, 2023.

Kuradusenge, M., Kumaran, S., & Zennaro, M.: Rainfall-induced landslide prediction using machine learning models: the case of ngororero district, rwanda, International Journal of Environmental Research and Public Health, 17, 4147, https://doi.org/10.3390/ijerph17114147, 2020.

Li, D., Huang, F., Yan, L., Cao, Z., Chen, J., & Zhou, Y.: Landslide susceptibility prediction using particle-swarm-optimized multilayer perceptron: comparisons with multilayer-perceptron-only, bp neural network, and information value models, Applied Sciences, 9, 3664, https://doi.org/10.3390/app9183664, 2019.

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

Li, W., Liu, Y., Chen, Y., & Yang, L.: Shock and vibration of rainfall on rotational landslide and analysis of its deformation characteristics, Geofluids, 2021, 4119414, https://doi.org/10.1155/2021/4119414, 2021.

Lin, L., Lin, Q., & Wang, Y.: Landslide susceptibility mapping on global scale using method of logistic regression, Natural Hazards and Earth System Sciences Discussions, https://doi.org/10.5194/nhess-2016-347, 2016.

Lin, L., Lin, Q., & Wang, Y.: Landslide susceptibility mapping on a global scale using the method of logistic regression, Natural Hazards and Earth System Science, 17, 1411-1424, https://doi.org/10.5194/nhess-17-1411-2017, 2017.

Lu, H., et al.: Deep learning for landslide susceptibility assessment: A comprehensive review, Engineering Geology, 325, 107289, 2024.

Ma, J., et al.: Machine learning models for slope stability analysis: A comparative study, Computers and Geotechnics, 140, 104456, 2021.

Maragaño, P., et al.: Landslide susceptibility in San José de Maipo, Chile: A machine learning approach, Natural Hazards, 118, 1567-1589, 2023.

Marc, O., Hovius, N., Meunier, P., Görüm, T., & Uchida, T.: A seismologically consistent expression for the total area and volume of earthquake-triggered landsliding, Journal of Geophysical Research: Earth Surface, 121, 640-663, https://doi.org/10.1002/2015jf003732, 2016.

Marc, O., Gosset, M., Saitô, H., Uchida, T., & Malet, J.: Spatial patterns of storm-induced landslides and their relation to rainfall anomaly maps, Geophysical Research Letters, 46, 11167-11177, https://doi.org/10.1029/2019gl083173, 2019.

Merghadi, A., et al.: Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance, Earth-Science Reviews, 207, 103225, 2020.

Miao, F., Xie, X., Wu, Y., & Zhao, F.: Data mining and deep learning for predicting the displacement of "step-like" landslides, Sensors, 22, 481, https://doi.org/10.3390/s22020481, 2022.

Mondini, A., Guzzetti, F., & Melillo, M.: Deep learning forecast of rainfall-induced shallow landslides, Nature Communications, 14, 2466, https://doi.org/10.1038/s41467-023-38135-y, 2023.

Nirbhav, S., et al.: Feature selection for landslide susceptibility mapping using information gain, Catena, 223, 106771, 2023.

Novellino, A., Cesarano, M., Cappelletti, P., Martire, D., Napoli, M., Ramondini, M., et al.: Slow-moving landslide risk assessment combining machine learning and insar techniques, Catena, 203, 105317, https://doi.org/10.1016/j.catena.2021.105317, 2021.

Ochoa, S., et al.: Villa Santa Lucia landslide analysis using machine learning techniques, Landslides, 22, 123-138, 2025.

Pánek, T., Břežný, M., Harrison, S., Schönfeldt, E., & Winocur, D.: Large landslides cluster at the margin of a deglaciated mountain belt, Scientific Reports, 12, 7278, https://doi.org/10.1038/s41598-022-09357-9, 2022.

Park, S. and Kim, J.: Landslide susceptibility mapping based on random forest and boosted regression tree models, and a comparison of their performance, Applied Sciences, 9, 942, https://doi.org/10.3390/app9050942, 2019.

Pham, B., Phong, T., Avand, M., Al-Ansari, N., Singh, S., Le, H., et al.: Improving voting feature intervals for spatial prediction of landslides, Mathematical Problems in Engineering, 2020, 4310791, https://doi.org/10.1155/2020/4310791, 2020.

Qiao, S., Feng, C., Yu, P., Tan, J., Uchimura, T., Wang, L., et al.: Investigation on surface tilting in the failure process of shallow landslides, Sensors, 20, 2662, https://doi.org/10.3390/s20092662, 2020.

Schlögl, M., et al.: Advanced machine learning for landslide prediction in changing climates, Earth Surface Dynamics, 13, 45-67, 2025.

Segoni, S., Rosi, A., Lagomarsino, D., Fanti, R., & Casagli, N.: Brief communication: using averaged soil moisture estimates to improve the performances of a regional-scale landslide early warning system, Natural Hazards and Earth System Science, 18, 807-812, https://doi.org/10.5194/nhess-18-807-2018, 2018.

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

Sepúlveda, S. and Petley, D.: Regional trends and controlling factors of fatal landslides in Latin America and the Caribbean, Natural Hazards and Earth System Science, 15, 1821-1833, 2014.

Serey, A., Pinero-Feliciangeli, L., Sepúlveda, S., Poblete, F., Petley, D., & Murphy, W.: Landslides induced by the 2010 chile megathrust earthquake: a comprehensive inventory and correlations with geological and seismic factors, Landslides, 16, 1153-1165, https://doi.org/10.1007/s10346-019-01150-6, 2019.

Serey, A., Sepúlveda, S., Murphy, W., Petley, D., & Pascale, G.: Developing conceptual models for the recognition of coseismic landslides hazard for shallow crustal and megathrust earthquakes in different mountain environments – an example from the chilean andes, Quarterly Journal of Engineering Geology and Hydrogeology, 54, qjegh2020-023, https://doi.org/10.1144/qjegh2020-023, 2020.

Shirzadi, A., Soliamani, K., Habibnejhad, M., Kavian, A., Chapi, K., Shahabi, H., et al.: Novel gis based machine learning algorithms for shallow landslide susceptibility mapping, Sensors, 18, 3777, https://doi.org/10.3390/s18113777, 2018.

Somos, L., et al.: Analysis of the Villa Santa Lucia debris flow event using remote sensing and field observations, Natural Hazards, 102, 1015-1038, 2020.

Song, Y., et al.: A comprehensive comparison of filter and wrapper feature selection methods for landslide susceptibility modeling, Computers & Geosciences, 187, 105621, 2025.

Steger, S., Moreno, M., Crespi, A., Zellner, P., Gariano, S., Brunetti, M., et al.: Deciphering seasonal effects of triggering and preparatory precipitation for improved shallow landslide prediction using generalized additive mixed models, Natural Hazards and Earth System Sciences Discussions, https://doi.org/10.5194/nhess-2022-271, 2022.

Teza, G., Cola, S., Brezzi, L., & Galgaro, A.: Wadenow: a matlab toolbox for early forecasting of the velocity trend of a rainfall-triggered landslide by means of continuous wavelet transform and deep learning, Geosciences, 12, 205, https://doi.org/10.3390/geosciences12050205, 2022.

Vaz, T., Zêzere, J., Pereira, S., Oliveira, S., Garcia, R., & Quaresma, I.: Regional rainfall thresholds for landslide occurrence using a centenary database, Natural Hazards and Earth System Science, 18, 1037-1054, https://doi.org/10.5194/nhess-18-1037-2018, 2018.

Vorpahl, P., Dislich, C., Elsenbeer, H., Märker, M., & Schröder, B.: Biotic controls on shallow translational landslides, Earth Surface Processes and Landforms, 38, 198-212, https://doi.org/10.1002/esp.3320, 2012.

Wang, Z., Wang, D., Guo, Q., & Wang, D.: Regional landslide hazard assessment through integrating susceptibility index and rainfall process, Natural Hazards, 104, 2153-2173, https://doi.org/10.1007/s11069-020-04265-5, 2020.

Wang, X., Wang, Y., Lin, Q., & Yang, X.: Assessing global landslide casualty risk under moderate climate change based on multiple gcm projections, International Journal of Disaster Risk Science, 14, 751-767, https://doi.org/10.1007/s13753-023-00514-w, 2023.

Xu, Y., Dai, Q., Zhu, J., Yao, Y., Zhang, J., Li, W., et al.: Increased significance of global concurrent hazards from 1981 to 2020, Earth's Future, 12, e2024EF004490, https://doi.org/10.1029/2024ef004490, 2024.

Yang, K., Niu, R., Song, Y., Dong, J., Zhang, H., & Chen, J.: Dynamic hazard assessment of rainfall-induced landslides using gradient boosting decision tree with google earth engine in three gorges reservoir area, china, Water, 16, 1638, https://doi.org/10.3390/w16121638, 2024.

Yao, W., et al.: Soil moisture dynamics and landslide initiation: A depth-dependent analysis, Engineering Geology, 328, 107345, 2025.

Yousefi, S., et al.: Optimal feature selection for landslide susceptibility assessment using wrapper methods, Geomorphology, 425, 108589, 2024.

Zhang, L., et al.: Performance evaluation of machine learning models for landslide susceptibility mapping, Catena, 231, 107289, 2024.

https://doi.org/10.5194/egusphere-2025-2764
Preprint. Discussion started: 30 June 2025

Zheng, A., Casari, A.: Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists, O'Reilly Media, Sebastopol, 2021.

635 Zhou, X., et al.: Comparative analysis of machine learning algorithms for landslide prediction, Natural Hazards, 118, 2145-2168, 2023.