Discussion

The spatial prediction of landslides is considered to be one of the most complex tasks in natural hazard risk assessment. Despite the fact that numerous methodologies have been proposed, the accuracy of the predictions is still a controversial issue. The development in the field of machine learning and the GIS platforms has led to the development of many new techniques, and methods. However, the further exploration of new methods is still necessary.

This research addresses this issue by evaluating and comparing four machine learning techniques. In general, RF outperform the other models in terms of classification effectiveness. In terms of hyperparameter calibration, the available computational resources have been used to perform a grid search. In the case of RF and Xgboost, these algorithms need to adjust a larger set of parameters.

The machine learning models are suitable for solving the studied problem, since they are able to handle the complex relationships between LCFs and removal susceptibility and to be robust in noisy environments (He et al. (2012), Wu et al. (2013), Huang et al. (2020)). The algorithms presented in this paper have been widely used in the literature for the generation of landslide susceptibility maps. SVM has obtained AUROC values ranging from 0.768 to 0.946 (Abedini et al. (2019b), Zhao et al. (2022), Huang et al. (2020), Huang et al. (2022), Chen and Guestrin (2016)). Logistic regression, which is mainly used as a benchmark with which it is possible to make a comparison with other models, has obtained AUROC values ranging from 0.792 to 0.934 (Zhao et al. (2022), Tsangaratos and Ilia (2016), Bruzón et al. (2021), Zhu et al. (2020), Huang et al. (2022)). On the other hand, XGBoost, although it has been used in fewer publications than the other algorithms, has obtained promising results: in Can et al. (2021) it obtained an AUROC of 0.96, while in Bruzón et al. (2021) it obtained an AUROC of 0.979. Finally, RF, which almost always obtains outstanding results in this problem, has an AUROC ranging from 0.9 to 0.985 (Zhao et al. (2022), Tsangaratos and Ilia (2016), Bruzón et al. (2021), Arabameri et al. (2020), Huang et al. (2022)), which is consistent with the results obtained, which are also supported by findings from previous studies (Pourghasemi and Rahmati (2018), Ali et al. (2021), Zhao et al. (2022)). One of the advantages that RF has in conjunction with XGBoost is that both are immune to multicollinearity that can occur due to the presence of multiple topographic derivatives as conditioning factors (Kotsiantis (2013), Piramuthu (2008), Can et al. (2021)) and has the ability to handle large data sets and its resistance to overfitting. Other advantages of RF are that it does not require assumptions on the statistical distribution in the conditioning factors, it takes into account interactions and nonlinear characteristics among the variables, and the ability to provide information on the influence of each variable in the final model (Catani et al. (2013), Pourghasemi and Rahmati (2018), Tsangaratos and Ilia (2016)). The differences between the models lie mainly in the fact that the principles they use to generate predictions are different. The SVM is able to map low-dimensional features to high-dimensional spaces using a kernel to find a characteristic hyperplane to maximize the categorical space. The problem with this method is that the corresponding mapping may be poor for the prediction in question, and also, if the data is noisy or overlapping, the performance of SVM may decrease. The RL characterizes the spatial relationship between the landslide events and the conditioning factors looking for the best fitting algorithm. However, it is very sensitive to multicollinearity, which limits its performance (Huang et al., 2022) and also, as the amount of data increased, it may not have been able to effectively model the relationship between variables, resulting in a decrease in accuracy. In the case of XGBoost, it can lead to overfitting if the number of trees is not carefully controlled (Abedi et al., 2022).

This study is novel in respect from a machine learning perspective in that a 5-fold cross-validation with 100 replicates is used to calculate the prediction metrics, while most studies use a static data partition to then calculate the indexes of interest. This methodology does not deal with the stochastic nature of the problem, so applying cross-validation with repetitions allows obtaining more robust results.

The choice of conditioning factors is a key aspect that influences the quality of susceptibility models (Costanzo et al., 2014). Although various methodologies for selecting factors have been proposed, including linear correlation (Irigaray et al., 2007) and the Kolmogorov-Smirnov test (Costanzo et al., 2014), there is still no universal criterion for making these selections, and the issue remains a topic for debate (Tien Bui et al., 2017). In general, topographic, geologic, soil, hydrologic, geomorphologic, and anthropogenic factors have been accepted in the literature for most susceptibility models. In some cases, factors that do not have predictive capability cause noise, affecting the quality of the model. In addition, it is important to eliminate those factors that have a high correlation index between them, to be able to apply cross-validation.

In our model, the NDGI and EVI spectral indices were used, instead of the NDVI, which has been widely used. However, it has important limitations, such as its dependence on the daily time in which the aerial images are taken, since it does not correct for changes in the angle of solar incidence. Therefore, this index produces inaccurate results. In this sense, EVI, which is calculated similarly to the NDVI, uses additional wavelengths to correct the NDVI inaccuracies. This corrects for variations in the solar angle, atmospheric distortions caused by airborne particles and land cover signals under vegetation. On the other hand, the NDGI, which has mainly used for glacier characterization, has a high predictive value for the susceptibility estimation, given by the IGR, so it also replaces the NDVI. NDGI uses spectral bands corresponding to green and red, so this would imply that landslide and non-landslide areas create contrast between these wavelengths. Therefore, it is suggested to use these indices in areas similar to the studied in this work.

Among the factors studied in this work, two stand out with respect to the others in terms of their influence on the model: the valley depth index (VD) and the TWI. A high valley depth index may be related to a high susceptibility to landslides due to the steep topography and abrupt relief present in the study area, which may favor the occurrence of gravitational processes and increase the erosion rate on the slopes, while a high TWI indicates a saturated soil, which implies an increase in the susceptibility to landslides.

It is also novel that the “Valley Depth” (VD) index is the one that provides the most information for the model. The variable VD (Valley Depth) in the study refers to the vertical distance to a base level of the hydrographic network. This index is calculated using an algorithm that involves interpolating the elevation of the base level of the hydrographic network and then subtracting this base level from the original elevations. This characteristic corresponds to the vertical distance to a base level of the hydrographic network. The algorithm that calculates this index consists of two steps, which involve the interpolation of the elevation of the base level of the hydrographic network, and the subsequent subtraction of this base level from the original elevations (Conrad, 2015). A high valley depth index may be related to a high susceptibility to landslides due to the steep topography and abrupt relief present in the study area, which favors the occurrence of gravitational processes and increases the rate of erosion on the slopes. This implies that the landslide and non-landslide sites in the area share similar values of VD respectively. This aspect is important for morphologies such as that of the Salado River basin, which has a marked slope at the geographic transition as it crosses from the foothills to the intermediate depression and has a “funnel” shape (González, 2018).

In summary, in terms of the the novelty of this study consists of applying repeated cross-validation to obtain the metrics of the models, and the use of Valley depth index, NDGI and EVI to construct the susceptibility models. Also, other novelty is the use of the MLR3 package in solving the machine learning problem, and the combination with other geospatial packages in R in order to produce the susceptibility maps. Also, The data sources used in the construction of the model proposed in our article come exclusively from satellite images and digital elevation models, unlike other studies, which consider sources of information with a greater number of data and are therefore more difficult for disaster risk management analysts to apply in practice. The approach has the advantage that it can allow the generation of systems that create susceptibility maps based on the periodic updating of satellite images, which can contribute to the creation of a susceptibility monitoring system that can be implemented by technical agencies in the disaster area.

After an extensive and updated literature review, we found few publications linked to susceptibility assessment in the Andes. In this regard, we found that in (Ospina-Gutiérrez, 2021) a susceptibility mapping was performed in a different Andean area in terms of geomorphology and climate, but like our study, the most successful algorithm corresponds to Random Forest. In (Brenning, 2015) they use GAM models for the calculation of susceptibility in areas near roads, and here they note the importance of curvature, like our study, as an important factor in the calculation of susceptibility. In (Lizama, 2022), they also found the relevance in the curvature. Finally, in (Buecchi, 2019) found that they can build useful and effective landslide susceptibility maps using only the DEM of the zone, holding the results obtained in this work, which also uses satellite imagery. Also, they use a logistic regression model to calculate susceptibility in the Cordillera Blanca, achieving an AUC of 0.75. The region in question presents topographic similarities with the Salado Basin, so the model built in this study may have promising results in that area.

The applicability of the proposed model is determined by the climatic, topographic and the morphometric characteristics of the study area. Under that perspective, the model can be expected to be suitable in areas worldwide that is a semi-arid zone, with a variable topography and a Mediterranean climate with a prolonged dry season, in addition to having narrow and deep valleys, where the maximum susceptibility is concentrated. Examples of these zones are the following:

- Colca Valley, Peru: This region is located in southern Peru and has a rugged topography with narrow and deep valleys. The climate is semi-arid with a prolonged dry season and has geomorphological characteristics similar to those of the Salado Basin.

- Indo Valley, Pakistan: This valley is located in northern Pakistan and is a mountainous region with deep, narrow valleys. The climate is arid with a prolonged dry season and the region has a geomorphology similar to the study zone.

- Colorado River Valley, United States: This region is located in the southern part of the state of Colorado and in northern New Mexico. It is a semi-arid area with a rugged and mountainous topography, and has narrow and deep valleys similar to those of the Salado Basin.