Introduction

Geological hazards constitute one of the greatest impacts that global and local economies, as well as human settlements, may face (Li et al. (2020), Wang et al. (2019)). In particular, landslides, which are defined as movements of soil, mud, debris, or rock, are the most common geological hazard in the world (Dang et al., 2020). Landslides are commonly induced by natural events such as earthquakes or heavy rainfall, which occur in specific geological, geomorphological, and hydrological environments. In mountainous areas, landslides can have significant effects on the topographic features, forests, soil, as well as on infrastructure such as roads and farming land.

The extent of these effects will depend on the magnitude of the landslides (Cao et al. (2019), Saha et al. (2021)) . In the last two decades, efforts in assessing landslides have focused on studying susceptible zones, understanding the mechanisms that govern landslides (Tien Bui et al. (2017), Pourghasemi and Rahmati (2018)). This has made it possible to extract valuable knowledge from the analysis of geomorphological, tectonic, geological, climatic, and anthropomorphic characteristics (Bui et al. (2020), Gorsevski et al. (2006), Flentje et al. (2007), Hervás and Bobrowsky (2009)).

In Chile, landslides are one of the most important geological hazards, together with earthquakes, volcanic activity and floods. The geological, geomorphological, tectonic and climatic conditions of the country, characterized by the presence of the Andes Mountain Range on its eastern margin and the Coastal Mountain Range on its western margin, make it highly susceptible to the generation of mass movements, such as landslides and rockfalls, flows and falls (Lara and Sepúlveda, 2010). Within 52 declared events in Chile, there is a total of 1010 fatal victims caused by landslides, corresponding to 882 deaths and 128 people missing between 1928 and 2017 (90 years) (Marín et al., 2018). The Atacama Region is characterized by a geomorphology that renders it susceptible to landslide events. According to documented records, this region has witnessed the highest number of fatalities resulting from flow-type events in the country, with a total of 132 people (Marín et al., 2018). In the area of study, the Salado River basin, situated in the Chañaral Province of the region, a landslide event occurred in 1940 that resulted in the destruction of houses and interruption of roads, causing substantial disruption to the city of Chañaral. In 1972, another landslide affected Chañaral and towns upstream, which were flooded and 700 people were affected in Chañaral and 400 people in El Salado. Another flood in 1983 affected the city due to a rise in the river level caused by an increase in rainfall (González, 2018). In Chañaral, there have been at least 15 major landslide events in the last 150 years (Vargas Easton et al., 2018). With regard to the reports of deadly events and economic damage provoked by landslides, the identification of areas prone to these types of events and the determination of their risk level are the most critical actions in the assessment of the hazard (Abedini et al., 2019a).

In the last two decades, extensive research has been carried out on landslide methods using innovative technologies and tools to promote this field, such as in crisis management in mountain areas or near them (Dahal et al., 2008). Therefore, in order to obtain a reliable and accurate vulnerability map, it is

necessary to test and evaluate several quantitative methods for more effective management of mountainous areas Abedini et al. (2019a). The use of Geographical Information Systems (GIS) in the elaboration of susceptibility maps constitutes an effective method to identify and delineate landslide-prone areas. This allows for the creation of a geospatial database of occurrences or an exhaustive inventory. By using GIS data repositories, the geospatial attributes of landslide-prone sites that can influence the potential stability of slopes, called landslide conditioning factors (LCFs), can be aggregated into a database.

Several methodologies and techniques have been developed for the hazard susceptiblity cartography around the world. Literature has classified them into (Abedini et al. (2019b), Merghadi et al. (2020)):

1. Models founded on physical conditions
2. Models founded on expert knowledge (Shirzadi et al. (2012), Zhang et al. (2016)).
3. Multivariate statistical methods. The examples are the statistical index (SI) Regmi et al. (2014), the frequency ratio (FR) (Pham et al. (2015), Shirzadi et al. (2012)), and the logistic regression (Shirzadi et al. (2012), Tsangaratos and Ilia (2016), Chen et al. (2019), Mousavi et al. (2011))

4. Machine learning models, such as decision trees (DT) (Khosravi et al., 2018), random forest (RF) (Hong et al., 2016), artificial neural networks (ANN) (Pham et al. (2018), Shirzadi et al. (2012)) and some hybrid methods which include optimization algorithms (Abedini et al. (2019a), Ahmadlou et al. (2019), Tien Bui et al. (2017), Chen and Guestrin (2016)).

.

Each of the methods described has unique strengths and limitations (Khosravi et al., 2018). Physical models, for example, require extensive field analysis and are currently considered unbeatable in terms of prediction accuracy, making them suitable for local scale mapping. In order to work effectively, these models demand a complete knowledge of the landslide systems, obtained through meticulous observation and monitoring of the surface and the subsurface; this is essential in order to issue timely warnings of further slope collapse (Piciullo et al., 2018). However, when applied on a larger scale, the need for a large amount of substantial data to obtain reliable results becomes inconvenient due to the considerable financial and computational resources required. Therefore, the use of this technique for the segmentation of larger regions is not feasible. This has led to the proliferation of statistical and knowledge-based models for more than four decades (Guzzetti et al., 2012). The knowledge-based models operate on the premise of building a framework with limited information, which is then parameterised by a system of weights assigned to factors according to expert judgement. Statistical models, on the other hand, have benefited from recent advances in GIS. This has paved the way or the beginning and the successful application of a set of tools and cuantitave methodologies for the modelling of landslides, improving in this way the understanding of the associated patterns and the causative agents (Dou et al., 2019). At the moment, the thin line that separates the statistical models from machine learning is a subject of controversy (Ij (2018), Merghadi et al. (2020)). The synergy and differences between statistical methods and machine learning are not clearly explained in academic works, mainly because the approach for geoscientists is primarily to generate

and refine accurate results in landslide susceptibility mapping (LSM) rather than algorithmic categorisation. In essence, machine learning is characterised by its ability to extract knowledge from data without relying on rule-based functions, whereas statistical modelling aims to establish relationships between data variables through algebraic expressions. Although the two fields were once considered mutually exclusive, they have recently converged (Merghadi et al., 2020). A notable example is the adoption of the logistic regression (LR) algorithm, originally from statistics, to solve classification problems. Now machine learning has adopted LR and has become one of the most widely used algorithms. However, machine learning is more concerned with optimisation and efficiency, in contrast to the inferential approach of statistical models.

Machine learning methods have been used in engineering and science problems for more than two decades. This is the reason why the use of these techniques in the area of geosciences and remote sensing is quite new and limited. Machine learning focuses on the automatic extraction of information from data through computational and statistical methods. The areas of applicability are very diverse, and involve different topics such as rock mass characterization, ocean products, vegetation indices, etc. At present, data analysis methods play a central role in geosciences and remote sensing. While collecting large volumes of data is essential in the field, and the analysis of this information becomes a major challenge (Lary et al., 2016). Various machine learning techniques, including random forest (RF), support vector machine (SVM), and artificial neural network (ANN), have proven to be effective in dealing with nonlinear data across different scales in areas such as identification, prediction, mitigation, and modeling. Studies like Sekkeravani et al. (2022), Miao et al. (2018), Saha et al. (2022), Conforti et al. (2014) and others have demonstrated the success of these methods. Unlike traditional statistical models, which aim to infer relationships between variables, machine learning models autonomously identify logical criteria from input data to make highly accurate predictions (Miao et al., 2023). The main advantages of the machine learning are the accuracy (Machine Learning can learn complex, nonlinear patterns of data), flexibility (These models can be adapted to different types of data and probems), speed (ML can process large amounts of data quickly and efficiently) and generalization (Once the model is trained, it can be applied to new data to produce reliable predictions).

There are several research gaps in the modeling of landslide susceptibility using machine learning algorithms. Some of them are the

- Integration of multiple factors: Most studies have focused on assessing a limited and repetitive set of factors. The integration of new factors needs to be explored to improve model accuracy and ensure a more complete assessment of susceptibility.

- Development of interpretable models: Although machine learning models can achieve high accuracy in identifying landslide susceptibility, some of these models can be difficult to interpret and explain. There is a need to develop interpretable models that allow decision makers to understand how the data is being used and how susceptibility identification is being performed.

- The lack of simplification of the models in terms of the use of factors: Many models rely on multiple data sources, such as area-specific maps (geology, soils, roads, etc.), which hinder subsequent reproducibility and detract from some dynamic dimension of the model, which if only satellite parameters and/or digital elevation were used, could be updated as these images become available.

In this work, we seek to fill one of the existing research gaps in the Central Andes, in the sense that there are few studies that characterise the susceptibility to landslides, which can help to understand the development of this phenomenon, and thus apply the models generated in similar areas that have not yet been studied. For its part, the research question that this study seeks to address is how different Machine Learning techniques can be applied and compared to assess and predict susceptibility to landslides in a region of the Andes where no studies of this type have been carried out before.

The objectives of the work is to build a susceptibility model of the Chañaral province to identify the areas most exposed to landslide risk by using machine learning algorithms (SVM, RF, XGBoost and LR), and by comparing their performance; to build an inventory of landslides in the study area through historical records and the analysis of satellite images and to determine the most relevant factors in susceptibility assessment by using indices based on information theory